BRIDGING MUSIC INFORMATICS WITH MUSIC COGNITION

EDITED BY: Naresh N. Vempala, Frank A. Russo and Geraint A. Wiggins
PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88945-571-3
DOI 10.3389/978-2-88945-571-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

BRIDGING MUSIC INFORMATICS WITH MUSIC COGNITION

Topic Editors:
Naresh N. Vempala, Nuralogix Corporation and Ryerson University, Canada
Frank A. Russo, Ryerson University, Canada
Geraint A. Wiggins, Vrije Universiteit Brussel, Belgium

Citation: Vempala, N. N., Russo , F. A., Wiggins G, A., eds. (2018). Bridging Music Informatics With Music Cognition Frontiers Media. doi: 10.3389/978-2-88945-571-3

Table of Contents

SECTION 6
CORPUS ANALYSIS METHODS

SECTION 7
LISTENER BEHAVIOR

frontiers in Psychology

EDITORIAL
published: 08 May 2018
doi: 10.3389/fpsyg.2018.00633

Editorial: Bridging Music Informatics With Music Cognition

Naresh N. Vempala[1,2*] *and Frank A. Russo*[1]

[1] Psychology, Ryerson University, Toronto, ON, Canada, [2] Nuralogix Corporation, Toronto, ON, Canada

Keywords: music cognition, music informatics, music emotion, computational modeling, musical preference, music representation, music segmentation

Editorial on the Research Topic

Bridging Music Informatics With Music Cognition

Over 30 authors contributed 15 articles toward this research topic. Collectively this body of work represents a bridge between music informatics and music cognition, covering a broad range of research topics.

We can categorize these fifteen articles into one of the following groups or a combination of them, since the groups are not mutually exclusive:

(1) Research addressing problems or needs fundamental to one domain but borrowing methods, approaches, and/or insights from the other domain.
(2) Research addressing problems or needs common to both domains and borrowing methods and insights from either of the two domains.
(3) Research addressing problems or needs of one domain with strong implications for the other domain.

Eleven articles (i.e., 73.3%) attempt to elucidate underlying mental processes related to music. These articles may be thought of as predominantly aligned with music cognition (Baker and Müllensiefen; Barone et al.; Casey; Foubert et al.; Kim; McAdams et al.; McFee et al.; Siedenburg and Müllensiefen; Stober; van der Weij et al.; Vempala and Russo). Two articles (i.e., 13.3%) (Kaneshiro et al.; Thoret et al.) explore issues that fall mainly within the space of music informatics, while the two remaining articles (i.e., 13.3%) (Janssen et al.; Krumhansl) explore areas with research motivations relevant to both music cognition and music informatics. This cursory analysis might suggest that only limited interactions between these domains exist. With the majority of interactions biased toward music cognition, one might argue that this fragile new bridge is at risk of collapse!

However, a closer examination of the articles reveals a richer and balanced network of interactions. Of the eleven articles that are predominantly aligned with music cognition, no less than six (Barone et al.; Casey; Foubert et al.; McAdams et al.; Vempala and Russo; Siedenburg and Müllensiefen) use feature extraction methods hailing from music informatics. In other words, the dependence of these studies on music informatics should not be understated. Additionally, most of

OPEN ACCESS

Edited and reviewed by:
Bernhard Hommel,
Leiden University, Netherlands

***Correspondence:**
Naresh N. Vempala
nvempala@gmail.com

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *21 March 2018*
Accepted: *16 April 2018*
Published: *08 May 2018*

Citation:
Vempala NN and Russo FA (2018) Editorial: Bridging Music Informatics With Music Cognition Front. Psychol. 9:633. doi: 10.3389/fpsyg.2018.00633

these eleven articles have moderate to strong implications for music informatics. Likewise, the two articles that fall predominantly within music informatics, have implications for music cognition.

Since all the articles present research in more than one key area within music informatics and music cognition, they may be thought of as forming dynamic clusters that may be characterized differently depending on one's vantage point. The key areas driving these clusters include but are not limited to: statistical and computational modeling, machine learning, music and emotion, musical preference and engagement, rhythm and meter perception, musical timbre and instrument identification, music similarity, music representation, structural segmentation, implied harmony, music therapy, and big data analysis.

Baker and Müllensiefen, Kim, McAdams et al., van der Weij et al., Vempala and Russo, use computational modeling as a means to explain or interpret behaviors associated with music cognition. van der Weij et al. use a probabilistic model of meter expectation to explain the effects of enculturation. But their model is generative and borrows techniques from machine learning, thus bridging into music informatics. Both McAdams et al. and Vempala and Russo explore music and emotion. While McAdams et al. examine perceived emotion based on the acoustic properties of timbre, Vempala and Russo explore higher-level emotion judgments through a classic cognitive modeling framework using machine learning methods. Baker and Müllensiefen look at how similarity in compositional structure affects salience and recognition, specifically through the use of Wagner's leitmotives. Among all the computational modeling studies, Kim's gradient frequency neural network for estimating implied harmony, is the only biologically inspired low-level computational model consisting of tonotopically tuned nonlinear oscillators.

Both Stober and Casey present findings on music representation as assessed by neural activity—a topic that intersects music cognition, music information retrieval, and cognitive neuroscience. Stober explores music imagery information retrieval through EEG recordings whereas Casey examines neural representation of music in naturalistic listening conditions through fMRI. Both studies strongly depend on machine learning and deep learning methods. Stober's work also highlights the need for sharing open datasets. Open science is a practice common to music informatics and one that is fast gaining ground in music cognition. This approach promotes collaborative research endeavors and encourages replicability of research findings.

Several studies in this topic address the importance of timbre in music. While Siedenburg and Müllensiefen focus on music similarity judgments, Thoret et al. look at timbre and the modulation power spectrum as feature sources for musical instrument identification. Thoret et al.'s work is similar to, McAdams et al. since both inspect the role of timbre in music perception. However, given the importance of automatic source recognition in music informatics, it can be argued that Thoret et al.'s work on instrument identification is more closely aligned with music informatics than music cognition.

Barone et al., Kaneshiro et al., and Janssen et al. emphasize the role of corpus analysis methods in music informatics and music cognition. Janssen et al. uses a folk music corpus to study the relationship between musical memory and melodic variation with pattern matching—research that is more traditionally aligned with music cognition but has clear implications for music informatics. Barone et al. and Kaneshiro et al. focus on the analysis of big data - an area that has become especially relevant since the advent of cloud storage and high performance computing resources. Barone et al. examine statistical regularities in music download patterns of listeners. Specifically, they look at genre and emotion preference using acoustic features. Their work serves as yet another example of research problems fundamental to music cognition using methods borrowed from music informatics.

Kaneshiro et al. also explore musical behavior of listeners at scale. They study the types of musical events within a piece of music that lead to enhanced engagement of the listener. Despite addressing issues related to perception and preference in music cognition, their work adheres more to music informatics because of its application areas comprising music discovery, multimedia search, and musical engagement.

McFee et al.'s work focuses on the analysis of musical structure, and its role in hierarchical music segmentation by annotators. They present ways to overcome limitations during the occurrence of inter-annotator disagreements because of ambiguous musical structure. Segmentation algorithms are an active area of music informatics while perception of musical structure is also integral to music cognition. As such, this research falls well within the scope of both music informatics and music cognition.

Foubert et al.'s article stands out as the only article with application in music therapy. Their research is based on the hypothesis that abnormal timing deviations during musical improvisation can be used as predictors of interpersonal relationship instability—a characteristic of borderline personality disorder. A statistical model motivated from music cognition, with rhythm and tempo-based pattern matching features borrowed from music informatics, is used to diagnose patients with borderline personality disorder.

Krumhansl's article presents the results of an extensive survey on the contexts in which people heard popular music in their lifetimes, and how they developed their preferences for music. The survey shows several interesting results about the progression of music listening across the life span of different participants. The results also provide more insights and context about different effects such as generational effects, song specific age effect, decade effect, influence of emotion on memory and preference, among others. This study has relevance for music informatics in particular, and for the music industry more generally.

Given the breadth of research occurring at the intersection of music informatics and music cognition, these 15 articles represent a small sampling. Nonetheless, through their range and diversity of topics, these articles give us a sense of the nature and scope of research at this intersection. Hence, we can safely conclude that, far from risk of collapse, the bridge between music

informatics and music cognition is built on solid foundations. The diversity of interactions explored in this topic suggests that this bridge is sustainable and that it will continue to support fruitful activity for decades to come.

AUTHOR CONTRIBUTIONS

NV was responsible for writing. FR was responsible for writing.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 03 August 2017
doi: 10.3389/fpsyg.2017.01337

Evaluating Hierarchical Structure in Music Annotations

Brian McFee[1,2]*, *Oriol Nieto*[3], *Morwaread M. Farbood*[2] *and Juan Pablo Bello*[2]

[1] *Center for Data Science, New York University, New York, NY, United States,* [2] *Music and Audio Research Laboratory, Department of Music and Performing Arts Professions, New York University, New York, NY, United States,* [3] *Pandora, Inc., Oakland, CA, United States*

Music exhibits structure at multiple scales, ranging from motifs to large-scale functional components. When inferring the structure of a piece, different listeners may attend to different temporal scales, which can result in disagreements when they describe the same piece. In the field of music informatics research (MIR), it is common to use corpora annotated with structural boundaries at different levels. By quantifying disagreements between multiple annotators, previous research has yielded several insights relevant to the study of music cognition. First, annotators tend to agree when structural boundaries are ambiguous. Second, this ambiguity seems to depend on musical features, time scale, and genre. Furthermore, it is possible to tune current annotation evaluation metrics to better align with these perceptual differences. However, previous work has not directly analyzed the effects of hierarchical structure because the existing methods for comparing structural annotations are designed for "flat" descriptions, and do not readily generalize to hierarchical annotations. In this paper, we extend and generalize previous work on the evaluation of hierarchical descriptions of musical structure. We derive an evaluation metric which can compare hierarchical annotations holistically across multiple levels. sing this metric, we investigate inter-annotator agreement on the multilevel annotations of two different music corpora, investigate the influence of acoustic properties on hierarchical annotations, and evaluate existing hierarchical segmentation algorithms against the distribution of inter-annotator agreement.

Keywords: music structure, hierarchy, evaluation, inter-annotator agreement

OPEN ACCESS

Edited by:
Naresh N. Vempala, Ryerson University, Canada

Reviewed by:
Dipanjan Roy, Allahabad University, India
Thomas Grill, Austrian Research Institute for Artificial Intelligence, Austria
Matthew Davies, Institute for Systems and Computer Engineering of Porto, Portugal

***Correspondence:**
Brian McFee
brian.mcfee@nyu.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *01 November 2016*
Accepted: *20 July 2017*
Published: *03 August 2017*

Citation:
McFee B, Nieto O, Farbood MM and Bello JP (2017) Evaluating Hierarchical Structure in Music Annotations. Front. Psychol. 8:1337. doi: 10.3389/fpsyg.2017.01337

1. INTRODUCTION

Music is a highly structured information medium, containing sounds organized both synchronously and sequentially according to attributes such as pitch, rhythm, and timbre. This organization of sound gives rise to various musical notions of harmony, melody, style, and form. These complex structures include multiple, inter-dependent levels of information that are hierarchically organized: from individual notes and chords at the lowest levels, to measures, motives and phrases at intermediate levels, to sectional parts at the top of the hierarchy (Lerdahl and Jackendoff, 1983). This rich and intricate pattern of structures is one of the distinguishing characteristics of music when compared to other auditory phenomena, such as speech and environmental sound.

The perception of structure is fundamental to how listeners experience and interpret music. Form-bearing cues such as melody, harmony, timbre, and texture (McAdams, 1989) can be interpreted in the context of both short and long-term memory. Hierarchies are considered a

fundamental aspect of structure perception, as musical structures are best retained by listeners when they form hierarchical patterns (Deutsch and Feroe, 1981). Lerdahl (1988) goes so far as to advocate that hierarchical structure is absolutely essential for listener appreciation of music since it would be impossible to make associations between nonadjacent segments without it. Hierarchical structure is also experienced by listeners over a wide range of timescales on the order of seconds to minutes in length (Farbood et al., 2015). Although interpretation of hierarchical structure is certainly influenced by acculturation and style familiarity (Barwick, 1989; Clayton, 1997; Drake, 1998; Drake and El Heni, 2003; Bharucha et al., 2006; Nan et al., 2006), there are aspects of it that are universal. For example, listeners group together some elements of music based on Gestalt theory (Deutsch, 1999; Trehub and Hannon, 2006), and infants have been shown to differentiate between correctly and incorrectly segmented Mozart sonatas (Krumhansl and Jusczyk, 1990).[1]

The importance of hierarchical structure in music is further highlighted by research showing how perception of structure is an essential aspect of musical performance (Cook, 2003). Examination of timing variations in performances has shown that the lengthening of phrase endings corresponds to the hierarchical depth of the ending (Todd, 1985; Shaffer and Todd, 1987). Performers also differ in their interpretations much like listeners (or annotators) differ in how they perceive structure. A combination of converging factors can result in a clear structural boundary, while lack of alignment can lead to an ambiguous boundary. In ambiguous cases, listeners and performers may focus on different cues to segment the music. This ambiguity has not been the focus of empirical work, if only because it is (by definition) hard to generalize.

Unsurprisingly, structure analysis has been an important area of focus for music informatics research (MIR), dealing with tasks such as motif-finding, summarization and audio thumbnailing, and more commonly, segmentation into high-level sections (see Paulus et al., 2010 for a review). Applications vary widely, from the analysis of a variety of musical styles such as jazz (Balke et al., 2016) and opera (Weiß et al., 2016), to algorithmic composition (Herremans and Chew, 2016; Roy et al., 2016) and the creation of mash-ups and remixes (Davies et al., 2014).

This line of work, however, is often limited by two significant shortcomings. First, most existing approaches fail to account for hierarchical organization in music, and characterize structure simply as a sequence of non-overlapping segments. Barring a few exceptions (McFee and Ellis, 2014a,b; McFee et al., 2015a; Grill and Schlüter, 2015), this flat temporal partitioning approach is the dominant paradigm for both the design and evaluation of automated methods. Second, and more fundamentally, automated methods are typically trained and evaluated using a single "ground-truth" annotation for each recording, which relies on the unrealistic assumption that there is a single valid interpretation to the structure of a given recording or piece. However, it is well known that perception of musical structure is ambiguous, and that annotators often disagree in their interpretations. For example, Nieto (2015) and Nieto et al. (2014) provide quantitative evidence of inter-annotator disagreement, differentiating between content with high and low ambiguity, and showing listener preference for over- rather than under-segmentation. The work of Bruderer (2008) shows that annotators tend to agree when quantifying the degree of ambiguity of music segment boundaries, while in Smith et al. (2014) disagreements depend on musical attributes, genre, and (notably) time-scale. Differences in time-scale are particularly problematic when hierarchical structures are not considered, as mentioned above. This issue can potentially result in a lack of differentiation between *superficial* disagreements, arising from different but compatible analyses of a piece, from *fundamental* discrepancies in interpretation, e.g., due to attention to different acoustic cues, prior experience, cultural influences on the listener, etc.

The main contribution of this article is a novel method for measuring agreement between hierarchical music segmentations, which we denote as the *L-measure*. The proposed approach can be used to compare hierarchies of different depths, including flat segmentations, as well as hierarchies that are not aligned in depth, i.e., segments are assigned to the same hierarchical level but correspond to different time-scales. By being invariant to superficial disagreements of scale, this technique can be used to identify true divergence of interpretation, and thus help in isolating the factors that contribute to such differences without being confounded by depth alignment errors.

The L-measure applies equally to annotated and automatically estimated hierarchical structures, and is therefore helpful to both music cognition researchers studying inter-subject agreement and to music informatics researchers seeking to train and benchmark their algorithms. To this end, we also describe three experimental studies that make use of the proposed method. The first experiment compares the L-measure against a number of standard flat metrics for the task of quantifying inter-annotator agreement, and seeks to highlight the properties of this technique and the shortcomings of existing approaches. The second experiment uses the L-measure to identify fundamental disagreements and then seeks to explain some of those differences in terms of the annotators focus on specific acoustic attributes. The third experiment evaluates the performance of hierarchical segmentation algorithms using the L-measure and advances a novel methodology for MIR evaluation that steps away from the "ground-truth" paradigm and embraces the possibility of multiple valid interpretations.

2. CORPORA

In our experiments, we use publicly available sets of hierarchical structural annotations produced by at least two music experts per track. To the best of our knowledge, the only published data sets that meet these criteria are SALAMI (Smith et al., 2011) and SPAM (Nieto and Bello, 2016).

[1]In the context of the present article, these two elements (cultural and universal) are not differentiated because the listeners who provide hierarchical analyses all had prior experience with Western music.

2.1. SALAMI

The publicly available portion of the *Structural Annotations for Large Amounts of Music Information* (SALAMI) set contains two hierarchical annotations for 1,359 tracks, 884 of which have annotations from two distinct annotators and are included in this study. These manual annotations were produced by a total of 10 different music experts across the entire set, and contain three levels of segmentations per track: *fine*, *coarse*, and *function*. The *fine* level typically corresponds to short phrases (described by lower-case letters), while the *coarse* section represents larger sections (described by upper-case letters). The *function* level applies semantic labels to large sections, e.g., "verse" or "chorus" (Smith et al., 2011). The boundaries of the function level often coincide with those of the coarse level, but for simplicity and consistency with SPAM (described below), we do not use the function level. The SALAMI dataset includes music from a variety of styles, including jazz, blues, classical, western pop and rock, and non-western ("world") music. We manually edited 171 of the annotations to correct formatting errors and enforce consistency with the annotation guide.[2] The corrected data is available online.[3]

2.2. SPAM

The *Structural Poly Annotations of Music* is a collection of hierarchical annotations for 50 tracks of music, each annotated by five experts. Annotations contain *coarse* and *fine* levels of segmentation, following the same guidelines used in SALAMI. The music in the SPAM collection includes examples from the same styles as SALAMI. The tracks were automatically sampled from a larger collection based on the degree of segment boundary agreement among a set of estimations produced by different algorithms (Nieto and Bello, 2016). Forty-five of these tracks are particularly challenging for current automatic segmentation algorithms, while the other five are more straightforward in terms of boundary detection. In the current work we treat all tracks equally and use all 10 pairs of comparisons between different annotators per track. The SPAM collection includes some of the same audio examples as the SALAMI collection described above, but the annotators are distinct, so annotation data is shared between the two collections.

3. METHODS FOR COMPARING ANNOTATIONS

The primary technical contribution of this work is a new way of comparing structural annotations of music that span multiple levels of analysis. In this section, we formalize the problem statement and describe the design of the experiments in which we test the method.

3.1. Comparing Flat Segmentations

Formally, a *segmentation* of a musical recording is defined by a temporal partitioning of the recording into a sequence of labeled time intervals, which are denoted as *segments*. For a recording of duration T samples, a segmentation can be encoded as mapping of samples $t \in [T] = \{1, 2, \ldots, T\}$ to some set of segment labels $Y = \{y_1, y_2, \ldots, y_k\}$, which we will generally denote as a function $S : [T] \rightarrow Y$.[4] For example, Y may consist of functional labels, such as *intro* and *verse*, or section identifiers such as A and B. A *segment boundary* is any time instant at the boundary between two segments. Usually this corresponds to a change of label $S(t) \neq S(t-1)$ (for $t > 1$), though boundaries between similarly labeled segments can also occur, e.g., when a piece has an AA form, or a verse repeats twice in succession.

When comparing two segmentations—denoted as the *reference* S^{R} and *estimate* S^{E}—a variety of metrics have been proposed, measuring either the agreement of segment boundaries, or agreement between segment labels. Two segmentations need not share the same label set Y, since different annotators may not use labels consistently, so evaluation criteria need to be invariant with respect to the choice of segment labels, and instead focus on the patterns of label agreement shared between annotations. Of the label agreement metrics, the two most commonly used are *pairwise classification* (Levy and Sandler, 2008) and *normalized conditional entropy* (Lukashevich, 2008).

3.1.1. Pairwise Classification

The pairwise classification metrics are derived by computing the set A of pairs of similarly labeled distinct time instants (u, v) within a segmentation:

$$A(S) := \left\{(u, v) \;\middle|\; S(u) = S(v)\right\}. \tag{1}$$

Pairwise precision (P-Rrecision) and recall (P-Recall) scores are then derived by comparing $A\left(S^{\mathrm{R}}\right)$ to $A\left(S^{\mathrm{E}}\right)$:

$$\text{P-Precision}\left(S^{\mathrm{R}}, S^{\mathrm{E}}\right) := \frac{\left|A\left(S^{\mathrm{R}}\right) \cap A\left(S^{\mathrm{E}}\right)\right|}{\left|A\left(S^{\mathrm{E}}\right)\right|} \tag{2}$$

$$\text{P-Recall}\left(S^{\mathrm{R}}, S^{\mathrm{E}}\right) := \frac{\left|A\left(S^{\mathrm{R}}\right) \cap A\left(S^{\mathrm{E}}\right)\right|}{\left|A\left(S^{\mathrm{R}}\right)\right|}. \tag{3}$$

The precision score measures the correctness of the predicted label agreements, while the recall score measures how many of the reference label agreements were found in the estimate. Because these scores are defined in terms of exact label agreement between time instants, they are sensitive to matching the exact level of specificity in the analysis encoded by the two annotations in question. If S^{E} is at a higher (coarser) or lower (finer) level of specificity than S^{R}, the pairwise scores can be small, even if the segmentations are mutually consistent. Examples of this phenomenon are provided later in Section 4.

3.1.2. Normalized Conditional Entropy

The normalized conditional entropy (NCE) metrics take a different approach to measuring similarity between annotations.

[2]The SALAMI annotation guide is available at http://music.mcgill.ca/~jordan/salami/SALAMI-Annotator-Guide.pdf.

[3]https://github.com/DDMAL/salami-data-public/pull/15

[4]Although segmentations are typically produced by annotators without reference to a fixed time grid, it is standard to evaluate segmentations after re-sampling segment labels at a standard resolution of 10 Hz (Raffel et al., 2014), which we adopt for the remainder of this article.

Given the two flat segmentations S^{R} and S^{E}, a joint probability distribution $\mathbf{P}\left[y^{\mathrm{R}}, y^{\mathrm{E}}\right]$ is estimated as the frequency of time instants t that receive label y^{R} in the reference S^{R} and y^{E} in the estimate S^{E}:

$$\mathbf{P}\left[y^{\mathrm{R}}, y^{\mathrm{E}}\right] \propto \left|\left\{t \middle| S^{\mathrm{R}}(t) = y^{\mathrm{R}} \wedge S^{\mathrm{E}}(t) = y^{\mathrm{E}}\right\}\right| \tag{4}$$

From the joint distribution $\mathbf{P}$, the conditional entropy is computed between the marginal distributions $\mathbf{P}^{\mathrm{R}}$ and $\mathbf{P}^{\mathrm{E}}$:

$$\mathbb{H}\left(\mathbf{P}^{\mathrm{E}} \mid \mathbf{P}^{\mathrm{R}}\right) = \sum_{y^{\mathrm{R}}, y^{\mathrm{E}}} \mathbf{P}\left[y^{\mathrm{R}}, y^{\mathrm{E}}\right] \log \frac{\mathbf{P}^{\mathrm{R}}\left[y^{\mathrm{R}}\right]}{\mathbf{P}\left[y^{\mathrm{R}}, y^{\mathrm{E}}\right]} \tag{5}$$

The conditional entropy therefore measures how much information the reference distribution $\mathbf{P}^{\mathrm{R}}$ conveys about the estimate distribution $\mathbf{P}^{\mathrm{E}}$: if this value is small, then the segmentations are similar, and if it is large, they are dissimilar.

The conditional entropy is then normalized by $\log\left|Y^{\mathrm{E}}\right|$: the maximum possible entropy for a distribution over labels Y^{E}.[5] The normalized entropy is subtracted from 1 to produce the so-called *over-segmentation score* NCE_o, and reversing the roles of the reference and estimate yields the *under-segmentation score* NCE_u:

$$\mathrm{NCE}_o := 1 - \frac{\mathbb{H}\left(\mathbf{P}^{\mathrm{E}} \mid \mathbf{P}^{\mathrm{R}}\right)}{\log\left|Y^{\mathrm{E}}\right|} \tag{6}$$

$$\mathrm{NCE}_u := 1 - \frac{\mathbb{H}\left(\mathbf{P}^{\mathrm{R}} \mid \mathbf{P}^{\mathrm{E}}\right)}{\log\left|Y^{\mathrm{R}}\right|}. \tag{7}$$

The naming of these metrics derives from their application in evaluating automatic segmentation algorithms. If the estimate has large conditional entropy given the reference, then it is said to be *over-segmented* since it is difficult to predict the estimated segment label from the reference: this leads to a small NCE_o. Similar reasoning applies to NCE_u: if $\mathbb{H}\left(\mathbf{P}^{\mathrm{R}} \middle| \mathbf{P}^{\mathrm{E}}\right)$ is large, then it is difficult to predict the reference from the estimate, so the estimate is thought to be *under-segmented* (and hence a small NCE_u score). If both NCE_o and NCE_u are large, then the estimate is neither over- nor under-segmented with respect to the reference.

3.1.3. Comparing Annotations

When comparing two annotations in which there is no privileged "reference" status for either—such as the case with segmentations produced by two different annotators of equal status—the notions of precision and recall, or over- and under-segmentation can be dubious since neither annotation is assumed to be "correct" or *ground truth*. Arbitrarily deciding that one annotation was the reference and the other was the estimate would produce precision and recall scores, but reversing the roles of the annotations would exchange the roles of precision and recall, since P-Precision(S^1, S^2) = P-Recall(S^2, S^1).

[5]It has been recently noted that maximum-entropy normalization can artificially inflate scores in practice because the marginal distribution $\mathbf{P}^{\mathrm{E}}$ is often far from uniform. See https://github.com/craffel/mir_eval/issues/226 for details. For the remainder of this article, we focus comparisons on the pairwise classification metrics, but include NCE scores for completeness.

A common solution to this ambiguity is to combine precision and recall scores into a single summary number. This is most often done by taking the harmonic mean of precision P and recall R, to produce the $F1$-score or F-measure:

$$F := 2\frac{P \cdot R}{P + R}. \tag{8}$$

For the remainder of this article, we summarize the agreement between two annotations by the F-measure, using precision and recall for pairwise classification, and over- and under-segmentation for NCE metrics.

3.2. Hierarchical Segmentation

A *hierarchical segmentation* is a sequence of segmentations

$$H = (S_0, S_1, S_2, \ldots, S_m), \tag{9}$$

where the ordering typically encodes a coarse-to-fine analysis of the recording. Each S_i in a hierarchy is denoted as a *level*. We assume that the first level S_0 always consists of a single segment which spans the entire track duration.[6]

Most often, when presented with two hierarchical segmentations H^{R} and H^{E}, practitioners assume that the hierarchies span the same set of levels, and compare the hierarchies level-by-level: S_1^{R} to S_1^{E}, $S_2^{\mathrm{R}}, S_2^{\mathrm{E}}$, *etc.*, or between all pairs of levels (Smith et al., 2011). This results in a set of independently calculated scores for the set of levels, rather than a score that summarizes the agreement between the two hierarchies. Moreover, this approach does not readily extend to hierarchies of differing depths, and is not robust to depth alignment errors, where one annotator's S_1 may correspond to the other's S_2.

To the best of our knowledge, no previous work has addressed the problem of holistically comparing two labeled hierarchical segmentations. Our previous work (McFee et al., 2015a) addressed the unlabeled, boundary-detection problem, which can be recovered as a special case of the more general formulation derived in the present work (where each segment receives a unique label).

3.2.1. Hierarchical Label Agreement

Given a hierarchical segmentation H as defined in Equation (9) and time instants u, v, define the *meet* of u and v under H as

$$M(u, v \mid H) := \max k \text{ such that } S_k(u) = S_k(v), \tag{10}$$

that is, $M(u, v \mid H)$ is the deepest level of H where u and v receive the same label. The meet induces a partial ordering over pairs of time instants: large values of $M(u, v \mid H)$ indicate a high degree of similarity, and small values indicate low similarity.

To compare two hierarchical segmentations H^{R} and H^{E}, we examine triples of distinct time instants t, u, v in terms of the pairwise meets $M\left(t, u \middle| H^{\mathrm{R}}\right)$ and $M\left(t, v \middle| H^{\mathrm{R}}\right)$. We define the reference comparison set for a hierarchy H as

$$A(H) := \left\{(t, u, v) \middle| M\left(t, u \middle| H\right) > M\left(t, v \middle| H\right)\right\}, \tag{11}$$

[6]If S_0 is not provided, it can be trivially synthesized. Including S_0 in the hierarchy is useful for ensuring that the metrics derived in Section 3.2.1 are well-formed.

that is, the set of triples where (t, u) agree at a deeper level than the pair (t, v).

Level-independent precision and recall scores—*L-Precision* and *L-Recall*—can be defined, just as in the pairwise classification method of Section 3.1.1, by comparing the size of the intersection to the reference comparison set:

$$\text{L-Precision}\left(H^{\text{R}}, H^{\text{E}}\right) := \frac{\left|A(H^{\text{R}}) \cap A(H^{\text{E}})\right|}{\left|A(H^{\text{E}})\right|} \tag{12}$$

$$\text{L-Recall}\left(H^{\text{R}}, H^{\text{E}}\right) := \frac{\left|A(H^{\text{R}}) \cap A(H^{\text{E}})\right|}{\left|A(H^{\text{R}})\right|}. \tag{13}$$

These scores capture the rank-ordering of pairwise similarity between time instants, and can be interpreted as a relaxation of the pairwise classification metrics. We define the *L-Measure* as the harmonic mean of L-Precision and L-Recall.

Rather than asking if an annotation describes two instants (u, v) as the *same* or *different*, the scores defined here ask whether (t, u) as *more similar* or *less similar* to each-other than the pair (t, v), and whether that ordering is respected in both annotations. An example of this process is illustrated in **Figure 1**. Consequently, the proposed scores are robust to depth alignment errors between annotations, and readily support comparison between hierarchies of differing depth.

4. EXPERIMENT 1: L-MEASURES AND FLAT METRICS

Our first experiment investigates how the L-measure described above quantifies inter-annotator agreement for hierarchical music segmentation as compared to metrics designed for flat segmentations.[7]

4.1. Methods

The data sets described in Section 2 consist of musical recordings, each of which has at least two hierarchical annotations, which are each comprised of flat *upper* (high-level) and *lower* (low-level) segmentations. For each pair of annotations, we compare the L-measure to existing segmentation metrics (pairwise classification and normalized conditional entropy) at both levels of the hierarchy.

From this set of comparisons, we hope to identify examples illustrating the following behaviors: pairs where the flat metrics are small because the two annotations exist at different levels of analysis; and pairs where the flat metrics are large at one level, but small at the other, indicating hierarchical disagreement. In the calculation of all evaluation metrics, segment labels are sampled at a rate of 10 Hz, which is the standard practice for segmentation evaluation (Raffel et al., 2014).

4.2. Results and Discussion

Figure 2 illustrates the behavior on SALAMI of the L-measure compared to the flat segmentation metrics (right column), as well as all other pairs of comparisons between metrics. Overlaid in red on each plot is the best-fit robust (Huber's T) linear regression line, with shaded regions indicating the 95% confidence intervals as estimated by bootstrap sampling ($n = 500$ trials). This figure demonstrates a general trend of positive correlation between the L-measure and flat segmentation metrics at both levels, indicating that the L-measure integrates information across the entire hierarchy. Additionally, this plot exhibits a high degree of correlation between the pairwise classification and NCE metrics when confined to a single level. For the remainder of this section, we will focus on comparing L-measure to the pairwise classification metrics, which are more similar in implementation to L-measure.

To get a better sense of how the L-measure captures agreement over the full hierarchy, **Figure 3** compares the L-measure to the maximum and minimum agreements across levels of the hierarchy: that is, $L(H^{\text{R}}, H^{\text{E}})$ compared to $\max\left(F(S_1^{\text{R}}, S_1^{\text{E}}), F(S_2^{\text{R}}, S_2^{\text{E}})\right)$. The resulting plots are broken into quadrants I–IV along the median values of each metric, indicated in red. To simplify the presentation, we only compared the L-measure to the pairwise F-measure scores, though the results using normalized conditional entropy scores are qualitatively similar. Of particular interest in these plots are the points where the maximum is small (disagreement at both levels) or the minimum is large (agreement at both levels), and how the L-measure scores these points.

Quantitatively, of the points below the median of maximum F-measure (quadrants II and III of **Figure 3**, left), 81% lie below the median L-measure (quadrant III). Conversely, the points above the median of minimum F-measure (quadrants I and IV of **Figure 3**, right) have 75% above the median L-measure (quadrant I). These two quadrants (I and III) correspond to subsets of examples where the L-measure broadly agrees with the pairwise F-measure scores, indicating that there is little additional discriminative information encoded in the hierarchy beyond what is captured by level-wise comparisons. The remaining points correspond to inversions of score from what would be expected by level-by-level comparison: quadrant II in the left plot (9.5% of points), and IV in the right plot (12.6% of points).

Figure 4 illustrates example annotations drawn from each quadrant of the left plot of **Figure 3** (across-layer maximum vs. L-measure). The two plots in the left column, corresponding to quadrants II and III, illustrate examples where the flat metrics disagree at both levels. The top-left plot (track 347) achieves a large L-measure because the first annotator's upper-level matches well to the second annotator's lower level, but not to the second annotator's upper-level. However, the two hierarchies are generally consistent with one another, and the L-measure identifies this consistency. The top-right plot (track 555) achieves large pairwise agreement at the upper level (aside from *E/E'*, these annotations are equivalent up to a permutation of the labels), but small pairwise agreement at the lower level, because the annotators disagree about whether the lower-level segment labels repeat in the second half of the song. Just as in the previous example (347), these two hierarchies are mutually consistent, and the L-measure produces a high score for this pair. The bottom-left plot (track 436) appears to consist of genuinely incompatible

[7]Our implementations for the experiments included in this paper are available at https://github.com/bmcfee/segment_hierarchy_labels.

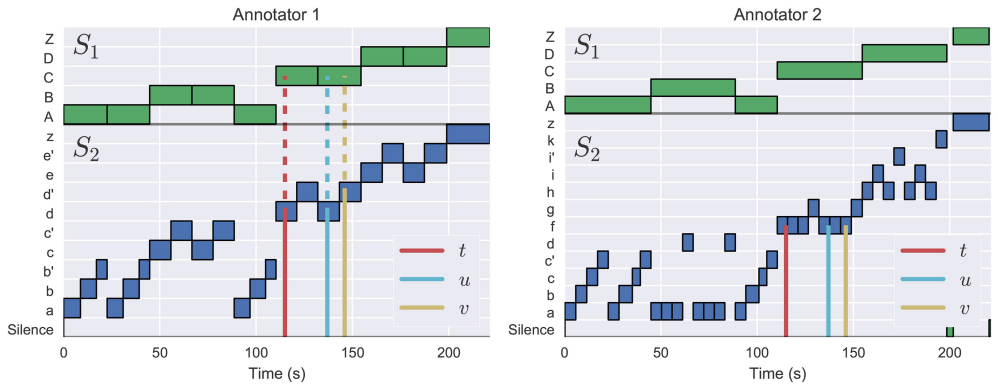


FIGURE 1 | The L-measure is computed by identifying triples of time instants (t, u, v) where (t, u) meet at a deeper level of the hierarchy (indicated by solid lines) than (t, v) (dashed lines), as illustrated in the left plot (Annotator 1). In this example, the left annotation has $M(t, u) = 2$ (both belong to lower-level segments labeled as *d*), and $M(t, v) = 1$ (both belong to upper-level segments labeled as *C*). The right annotation has $M(t, u) = M(t, v) = 2$: all three instants belong to segment label *f*, as indicated by the solid lines. This triple is therefore counted as evidence of disagreement between the two hierarchies.

hierarchies, resulting in small scores across all metrics. The bottom-right plot (track 616) illustrates agreement in the upper level, but significant disagreement in the lower level, which is taken as evidence of hierarchical disagreement and produces a small L-measure (0.30).

Similarly, **Figure 5** illustrates examples drawn from each quadrant of the right plot in **Figure 3** (across-layer minimum vs. L-measure). Here, the right column is of interest, since it lists annotations where the flat metrics agree at both levels (quadrants I and IV). The top-right plot (track 829) contains virtually identical hierarchies, and produces high scores under all metrics. The bottom-right plot (track 1342) consists of two essentially flat hierarchies where each lower-level contains the same label structure as the corresponding upper level. The large flat metrics here ($F = 0.80$) are easily understood since the majority of pairs of instants are labeled similarly in both annotations, excepting those (u, v) for which u is in section *C/c* for the second annotation and v is not, which are in the minority. The small L-measure (0.39) for this example is a consequence of the lack of label diversity in the first annotation, as compared to the second. By the definition in Equation (11), the L-measure only compares triples (t, u, v) where the labels for u and v differ, and in the second annotation, most of these triples contain an example from the *C/c* sections. Since the second annotation provides no information to disambiguate whether *C* is more similar to *A* or *Z*, the L-measure assigns a small score when compared to the first annotation.

A similar phenomenon can be observed in the bottom-left plot (track 768), in which the first annotator used a single label to describe the entire track in each level. In this case, nearly all of the comparison triples derived from the second annotation are not found in the first, resulting in an L-measure of 0.06. It is worth noting that the conditional entropy measures would behave similarly to the L-measure here, since the first annotation has almost no label entropy in either level.

To summarize, the L-measure broadly agrees with the level-by-level comparisons on the SALAMI dataset without requiring assumptions about equivalent level structure or performing comparisons between all pairs of levels. In the minority of cases (22%) where the L-measure substantially disagrees with the level-by-level comparison, the disagreements between metrics are often explained by the flat segmentations not accounting for hierarchical structure in the annotations. The exception to this are annotations with low label diversity across multiple levels, where the L-measure can assign a small score due to insufficiently many contrasting triples to form the evaluation (**Figure 5**, bottom-right).

5. EXPERIMENT 2: ACOUSTIC ATTRIBUTES

In the second experiment, we investigate annotator disagreement with respect to acoustic attributes. Two annotations that produce a small L-measure may be due to annotators responding to different perceptual or structural cues in the music.

5.1. Methods

To attempt to quantify attribute-based disagreement, we extracted four acoustic features from each recording, designed to capture aspects relating to tempo, rhythm, harmony, and timbre. Our hypothesis was that if hierarchical annotations receive small L-measure, and the annotators are indeed cued by different acoustic properties, then this effect should be evident when comparing annotations in a representation derived from acoustic features. All audio was down-sampled and mixed to 22,050 Hz mono prior to feature extraction, and all analysis was performed with librosa 0.5 dev (McFee et al., 2015b). A visualization of the features described in this section is provided in **Figure 6**.

5.1.1. Tempo Features

The tempo features consist of the short-time auto-correlation of the onset strength envelope of the recording. This feature loosely captures the timing structure of note onsets centered around each

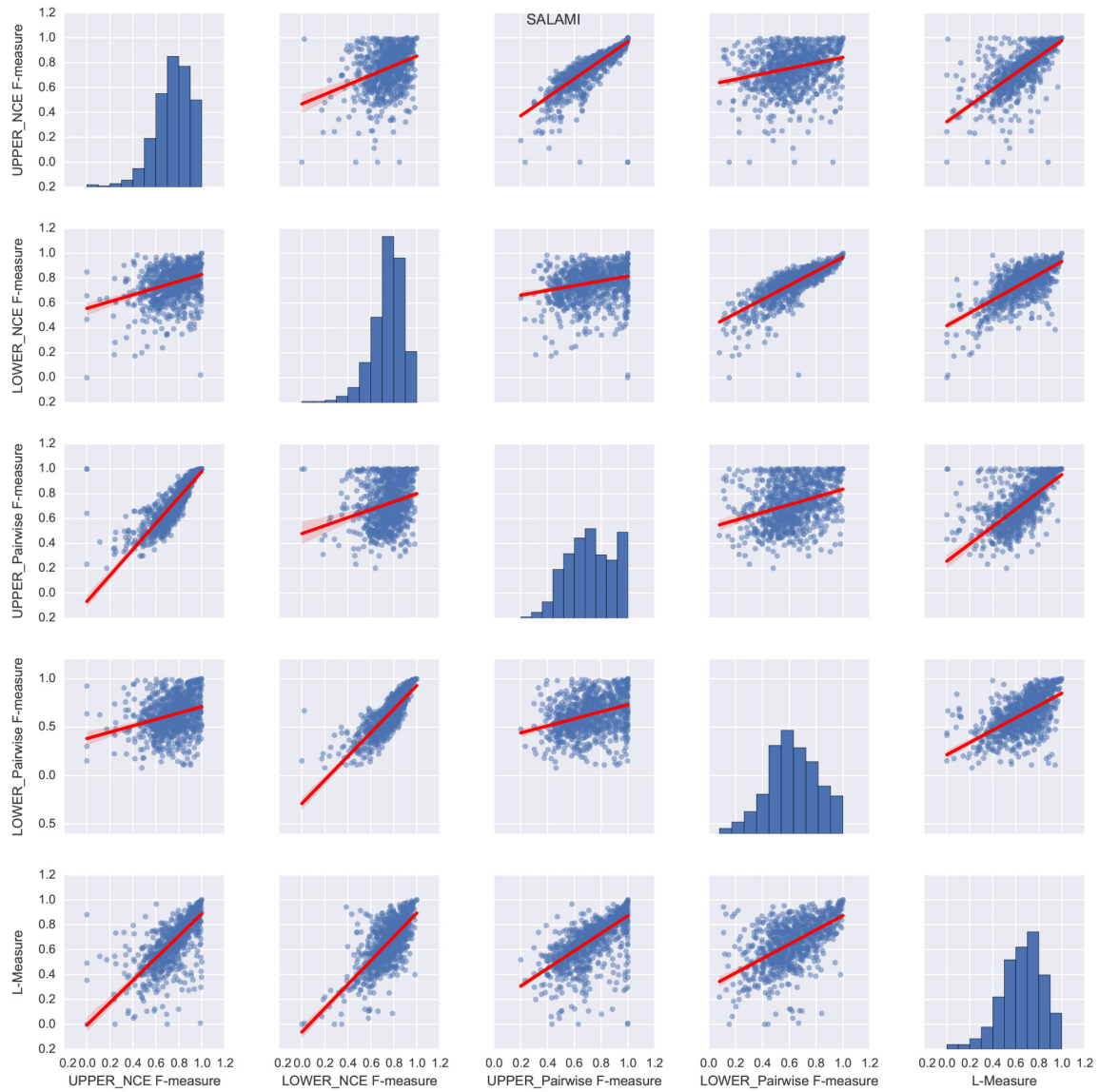


FIGURE 2 | Relations between the different segment labeling metrics on the SALAMI dataset. Each subplot (i, j) corresponds to a pair of distinct metrics for $i \neq j$, while the main diagonal illustrates the histogram of scores for the ith metric. Each point within a subplot corresponds to a pair of annotations of the same recording. The best-fit linear regression line between each pair of metrics is overlaid in red, with shaded regions indicating the 95% confidence intervals.

time point in the recording. The location of peaks in the onset strength auto-correlation can be used to infer the tempo at a given time.

The onset strength is computed by the spectral flux of a log-power Mel spectrogram of 128 bins sampled at a frame rate of $\sim$ 43 Hz (hop size of 512 samples), and spanning the frequency range up to 11,025 Hz. The short-time auto-correlation is computed over centered windows of 384 frames ($\sim$ 8.9 s) using a Hann window, resulting in a feature matrix $X_t \in \mathbb{R}_+^{384 \times T}$ (for T frames). The value at $X_\tau[i, j]$ is large if an onset envelope peak at frame j is likely to co-occur with another peak at frame $j + i$. Each column was normalized by its peak amplitude.

5.1.2. Rhythm Features

The rhythm features were computed by applying the scale (Mellin) transform to the tempo features derived above (Cohen, 1993; De Sena and Rocchesso, 2007). The scale transform magnitude has been used in prior work to produce an approximately tempo-invariant representation of rhythmic information (Holzapfel and Stylianou, 2011), so that similar rhythmic patterns played at different speeds result in similar feature representations.

At a high level, the scale transform works by re-sampling the onset auto-correlation—i.e., each column of X_τ defined above—on a logarithmic lag scale from a minimum lag $t_0 > 0$ to the maximum lag, which in our case is the auto-correlation window

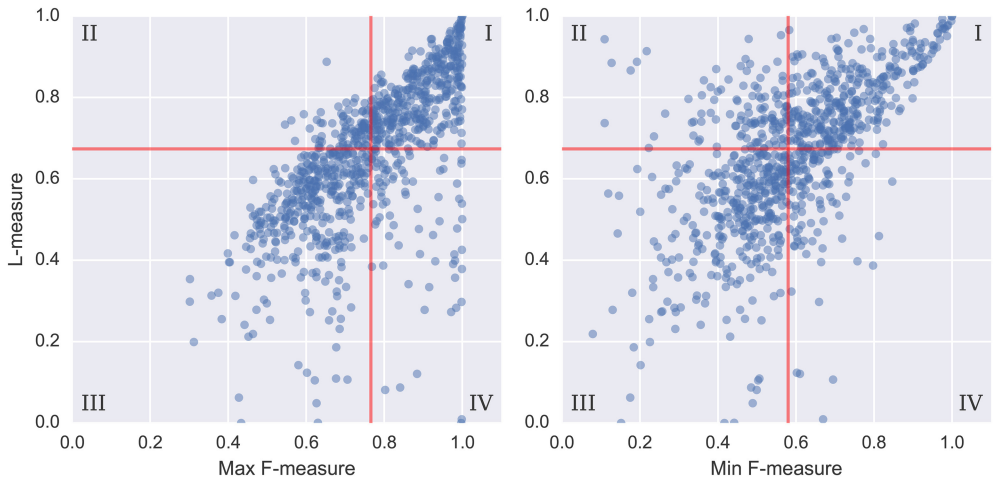


FIGURE 3 | For each pair of annotations in the SALAMI dataset, we compare the L-measure to the maximum and minimum agreement between the upper and lower levels. Agreement is measured by pairwise frame classification metrics. Red lines indicate the median values for each metric. A small maximum F-measure (quadrants II and III in the left plot) indicates disagreement at both levels; a large minimum F-measure (quadrants I and IV in the right plot) indicates agreement at both levels.

length (384 frames). This transforms multiplicative scaling in time to an additive shift in logarithmic lag. The Fourier transform of this re-sampled signal then encodes additive shift as complex phase. Discarding the phase information, while retaining the magnitude, produces a tempo-invariant rhythm descriptor.

The scale transform has two parameters which must be set: the minimum lag t_0 (in fractional frames), and the number of scale bins n (analogous to FFT bins), which we set to $t_0 = 0.5$ and $n = 64$. Because the input (onset autocorrelation) is real-valued, its scale transform is conjugate-symmetric, so we discard the negative scale bins to produce a representation of dimension $\lfloor n/2 \rfloor + 1$. The log-power of the scale transform magnitude was computed to produce the rhythm features $X_\rho \in \mathbb{R}^{33\times T}$.

5.1.3. Chroma Features

The harmony features were computed by extracting pitch class (*chroma*) features at the same time resolution as the tempo and rhythm features. Specifically, we applied the constant-Q transform magnitude using 36 bins per octave spanning the range $(C1, C8)$, summed energy within pitch classes, and normalized each frame by peak amplitude. This resulted in a chromagram $X_\chi \in \mathbb{R}_+^{12\times T}$.

5.1.4. Timbre Features

Finally, timbre features were computed by extracting the first 20 Mel frequency cepstral coefficients (MFCCs) using a log-power Mel spectrogram of 128 bins, and the same frame rate as the previous features. This resulted in the MFCC feature matrix $X_\mu \in \mathbb{R}^{20\times T}$.

5.1.5. Comparing Audio to Annotations

To compare audio features to hierarchical annotations, we converted the audio features described above to self-similarity matrices, described below. However, because the features are sampled at a high frame rate, the resulting $T \times T$ self-similarity matrices would require a large amount of memory to process ($\sim$ 3 GB for a four-minute song). We therefore down-sampled the feature matrices to a frame rate of 4 Hz by linear interpolation prior to computing the self-similarity matrices below. The tempo and rhythm features are relatively stable across large extents of time (each frame spans 8.9s), but the chroma and MFCC features are confined to much smaller local regions defined by their window sizes. To improve the stability of similarity for the chroma and MFCC features, each frame was extended by time-delay embedding (Kantz and Schreiber, 2004): concatenating the features of the previous two frames (after down-sampling). This provides a small amount of local context for each observation, and is a commonly used technique in music structure analysis algorithms (Serra et al., 2012).

We then computed self-similarity matrices for each feature with a Gaussian kernel:

$$G[u, v] := \mathrm{e}^{-\frac{1}{\sigma}\|X[u]-X[v]\|^2} \tag{14}$$

where $X[t]$ denotes the feature vector at frame t, and the bandwidth σ is estimated as

$$\sigma := \mathrm{mean}_u \ \mathrm{median}_v \|X[u] - X[v]\|^2. \tag{15}$$

Similarly, for each annotation, we computed the meet matrix M by Equation (10) (also at a frame rate of 4 Hz). **Figures 9**, **10** illustrate examples of the feature-based self-similarity matrices, as well as the meet matrices for two annotations each.

To compare M to each of the feature-based self-similarity matrices $G_\tau, G_\rho, G_\chi, G_\mu$, we first standardized each matrix by subtracting its mean value and normalizing to have unit Frobenius norm:

$$\widehat{D} := \frac{D - \mathrm{mean}_{u,v} D[u, v]}{\left\| D - \mathrm{mean}_{u,v} D[u, v] \right\|_{\mathrm{F}}}. \tag{16}$$

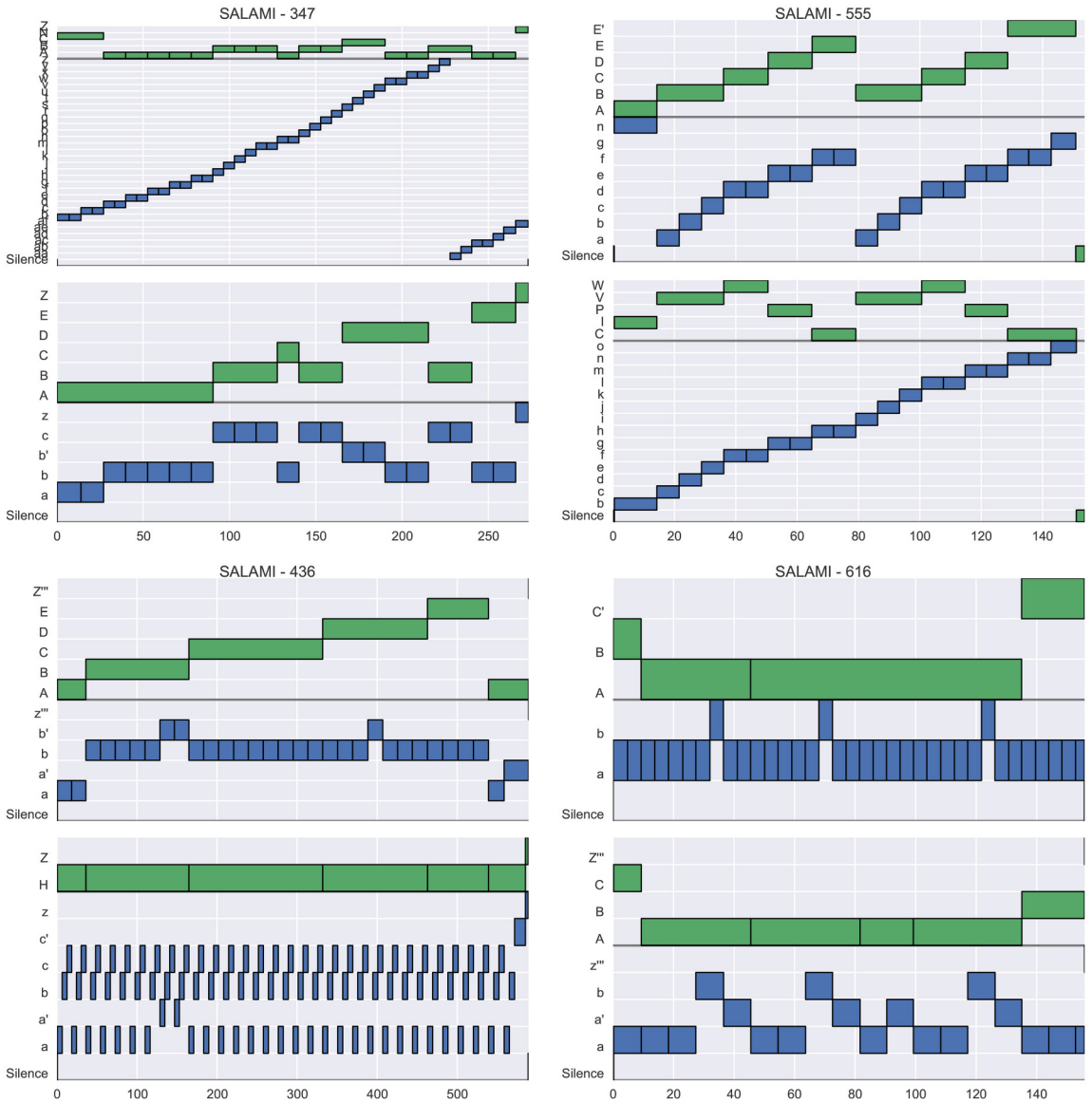


FIGURE 4 | Four example tracks from SALAMI, one drawn from each quadrant of **Figure 3 (Left)**, which compares L-measure to the maximum of upper- and lower-level pairwise F-measure between tracks. For each track, two hierarchical annotations are displayed (top and bottom), and within each hierarchy, the upper level is marked in green and the lower in blue. **(Upper right)** Track 555 ($L = 0.94$, upper $F = 0.92$, lower $F = 0.69$) has high agreement at the upper level, and small agreement at the lower level. **(Upper left)** Track 347 ($L = 0.89$, upper $F = 0.65$, lower $F = 0.19$) has little within-level agreement between annotations, but the upper level of the top annotation is nearly identical to the lower level of the bottom annotation, and the L-measure identifies this consistency. **(Bottom left)** Track 436 ($L = 0.24$, upper $F = 0.35$, lower $F = 0.44$) has little agreement at any level, and receives small scores in all metrics. **(Bottom right)** Track 616 ($L = 0.30$, upper $F = 0.998$, lower $F = 0.66$) has high agreement within the upper level, but disagreement in the lower levels.

The inner product between normalized self-similarity matrices

$$\left\langle \widehat{M}, \widehat{G} \right\rangle_{\mathrm{F}} := \sum_{u,v} \widehat{M}[u,v]\widehat{G}[u,v] \tag{17}$$

can be interpreted as a cross-correlation between the vectorized forms of M and G, and due to normalization, takes a value in $[-1, 1]$. Collecting these inner products against each G matrix results in a four-dimensional vector of feature-based similarity to the annotation M:

$$z(M) := \left(\left\langle \widehat{M}, \widehat{G}_i \right\rangle_{\mathrm{F}} \right)_{i \in \{\tau, \rho, \chi, \mu\}} \tag{18}$$

To compare two annotations $H^{\mathrm{R}}, H^{\mathrm{E}}$ with meet matrices $M^{\mathrm{R}}, M^{\mathrm{E}}$, we could compute the Euclidean distance between the corresponding z-vectors. However, correlated features (such as tempo and rhythm) could artificially inflate the distance calculation. We therefore define a whitening transform W^{-1},

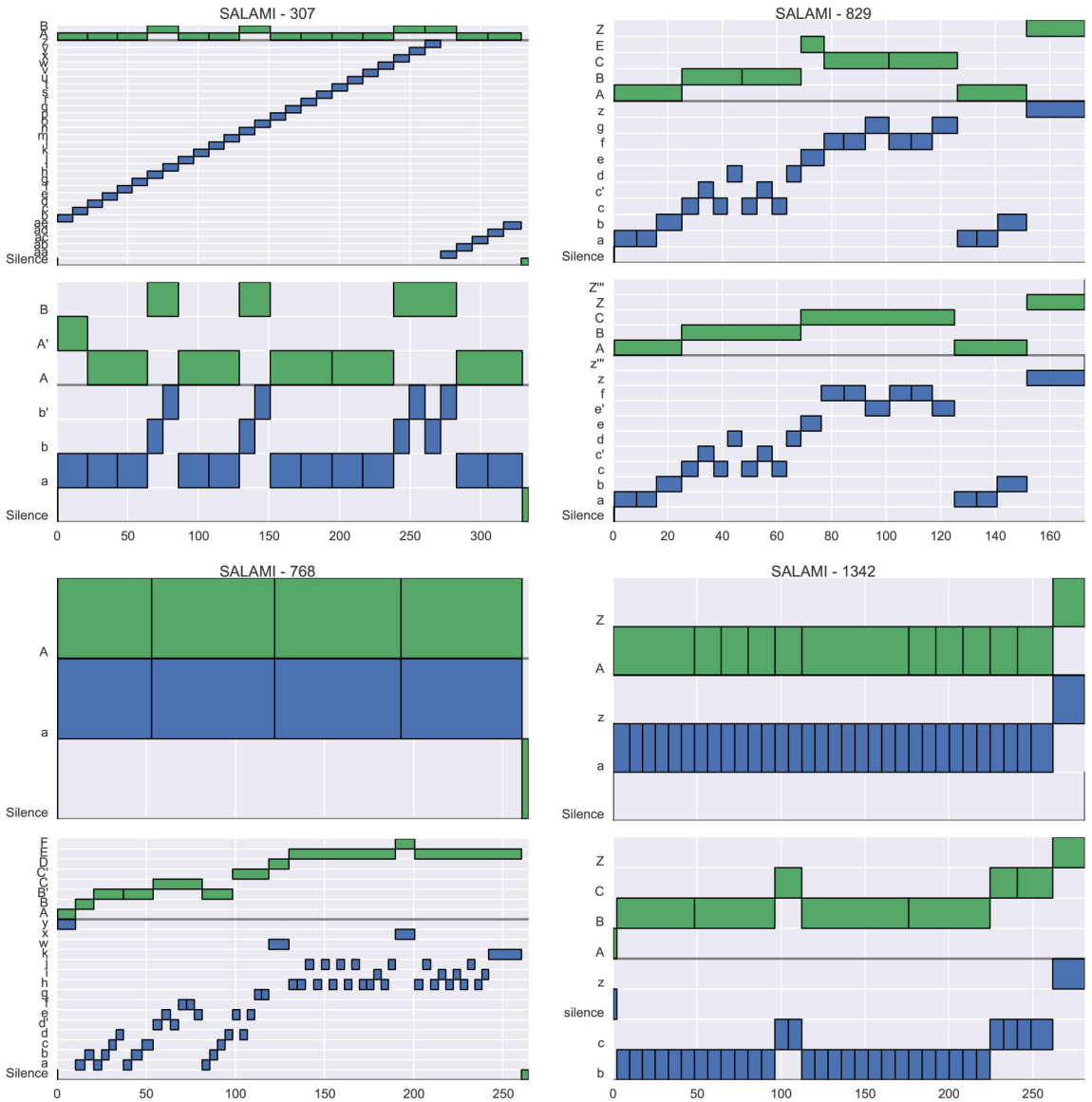


FIGURE 5 | Four example tracks from SALAMI, one drawn from each quadrant of **Figure 3 (Right)**, which compares L-measure to the minimum of upper- and lower-level pairwise F-measure between tracks. **(Upper right)** Track 829 ($L = 0.94$, upper $F = 0.93$, lower $F = 0.96$) has high agreement at the both levels, and consequently a large L-measure. **(Upper left)** Track 307 ($L = 0.94$, upper $F = 0.92$, lower $F = 0.11$) has high agreement in the upper level, but the first annotator did not detect the same repetition structure as the second in the lower level. **(Bottom left)** Track 768 ($L = 0.06$, upper $F = 0.43$, lower $F = 0.18$) has little agreement at any level because the first annotator produced only single-label annotations. **(Bottom right)** Track 1342 ($L = 0.39$, upper $F = 0.80$, lower $F = 0.80$) has high pairwise agreement at both levels, but receives a small L-measure because the first annotator did not identify the distinct *C/c* sections indicated by the second annotator.

where

$$W[i,j] := \left\langle \widehat{G}_i, \widehat{G}_j \right\rangle_{\mathrm{F}}. \tag{19}$$

This provides a track-dependent, orthogonal basis for comparing meet matrices M^{R} and M^{E}. The distance between annotations is then defined by

$$\delta\left(H^{\mathrm{R}}, H^{\mathrm{E}}\right) := \sqrt{\left(z\left(M^{\mathrm{R}}\right) - z\left(M^{\mathrm{E}}\right)\right)^{\mathrm{T}} W^{-1} \left(z\left(M^{\mathrm{R}}\right) - z\left(M^{\mathrm{E}}\right)\right)}. \tag{20}$$

By introducing the whitening transformation, we reduce the influence of correlations between acoustic features on the resulting annotation distance δ. A large distance δ indicates that the hierarchies correlate with different subsets of features, so we expect an inverse relationship between δ and the L-measure between the annotations.

5.2. Results and Discussion

The results of the acoustic feature correlation experiment are displayed in **Figure 7**. As expected, the δ score is inversely

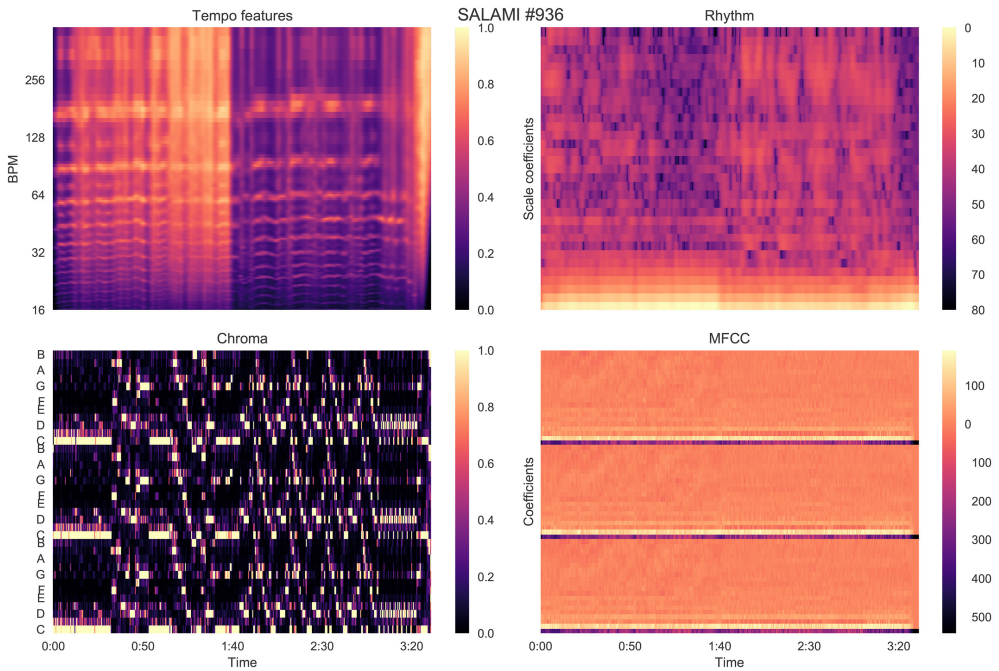


FIGURE 6 | Features extracted from an example track in the SALAMI dataset, as described in Section 5.

related to the L-measure ($r = -0.61$ on the SALAMI data set, $r = -0.32$ on SPAM). Because the SPAM dataset was explicitly constructed from difficult examples, it produces smaller L-measures on average than the SALAMI dataset. However, the SPAM annotators did not appear to produce low label-diversity annotations that generate small L-measures, so the overall distribution is more concentrated. The δ distribution is similar across both datasets, which explains the apparently large discrepancy in correlation coefficients.

The estimated mean feature correlations are displayed in **Figure 8**. Because the SPAM dataset provides all combinations of the five annotators with the fifty tracks, it is more amenable to statistical analysis of annotator behavior than the SALAMI dataset. Using the SPAM dataset, we investigated the relationship between feature types and annotators. A two-way, repeated-measures ANOVA was performed with annotator and feature type as fixed effects and tracks as a random effect (all results Greenhouse-Geisser corrected). The main effects of annotator and feature type were both significant: $F_{(2.92,\ 142.85)} = 3.44$, $p = 0.02$, $\eta^2 = 0.068$, $\eta_p^2 = 0.066$ for annotator and $F_{(2.52,\ 123.37)} = 28.33$, $p = 1.49 \times 10^{-12}$, $\eta^2 = 0.159$, $\eta_p^2 = 0.366$ for feature type. The interaction effect was also significant, $F_{(8.26,\ 404.97)} = 3.00$, $p = 2.46 \times 10^{-3}$, $\eta^2 = 5.17 \times 10^{-3}$, $\eta_p^2 = 0.058$. There was a large effect size for feature type and very small effect sizes for annotator and interaction.

Tukey's test for multiple comparisons revealed a significant difference between Annotators 3 and 4 ($|z| = 2.88$, $p = 0.032$) and a slight difference between 2 and 4 ($|z| = 2.52$, $p = 0.086$). **Figure 8** (right) indicates that most of this difference is likely attributable to the tempo feature, which annotator 4 correlates with considerably less than the other annotators. These results demonstrate that a small set of annotators are likely to produce significantly different interpretations of musical structure, even when they are following a common set of guidelines.

Figure 9 illustrates the self-similarity matrices for SALAMI track 410: *Erik Truffaz–Betty*, a jazz recording featuring trumpet, piano, bass, and drums. The two annotations for this track produce a small L-measure of 0.25, and a large δ score of 0.67. In this example, the two annotators appear to be expressing different opinions about the organization of the piece, as illustrated in the right-most column of **Figure 9**. Annotator 1 first separates the extended final fermata from the rest of the recording in the upper level, and then segments into repeated 4-bar progressions in the lower level. Annotator 2 groups by instrumentation or texture in the upper level, separating the piano and trumpet solos (center blocks) from the head section, and then grouping by repeated 8-bar segments. The first annotation correlates well with all of the feature-based similarity matrices, which exhibit low contrast for the majority of the piece. The second annotation is generally uncorrelated with the feature similarities, leading to the large δ score between the two. Note that this does not imply that one annotator was more "accurate" than the other, but it does suggest that the differences in the annotations can be attributed, at least in part, to perceptual characteristics of the music in question. In this case, Annotator 2 accounted for both instrumentation and harmony, while Annotator 1 accounted only for harmony.

Figure 10 illustrates a second example, SALAMI track 936: *Astor Piazzola – Tango Aspasionado*, which produces L-measure of 0.46 and a relatively large $\delta = 0.45$. The two annotators in

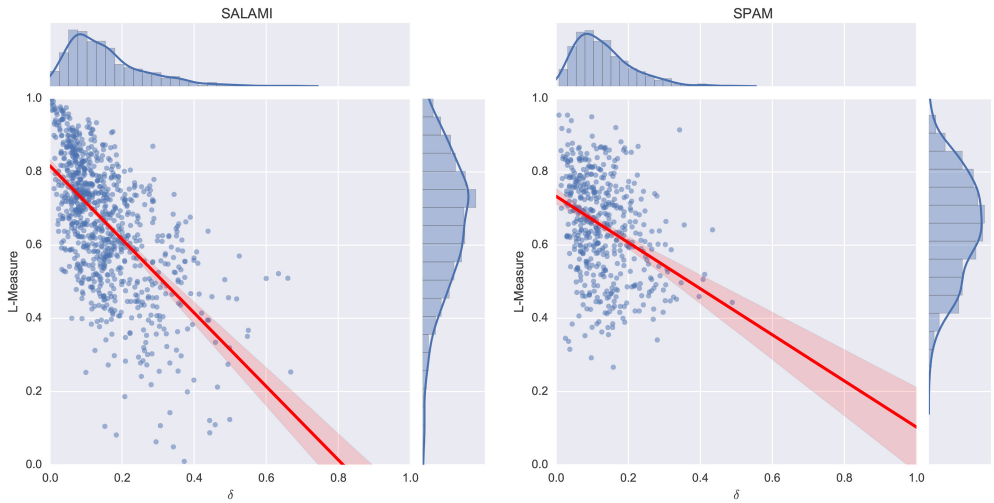


FIGURE 7 | Feature correlation compared to L-measures on the SALAMI **(Left)** and SPAM **(Right)** datasets.

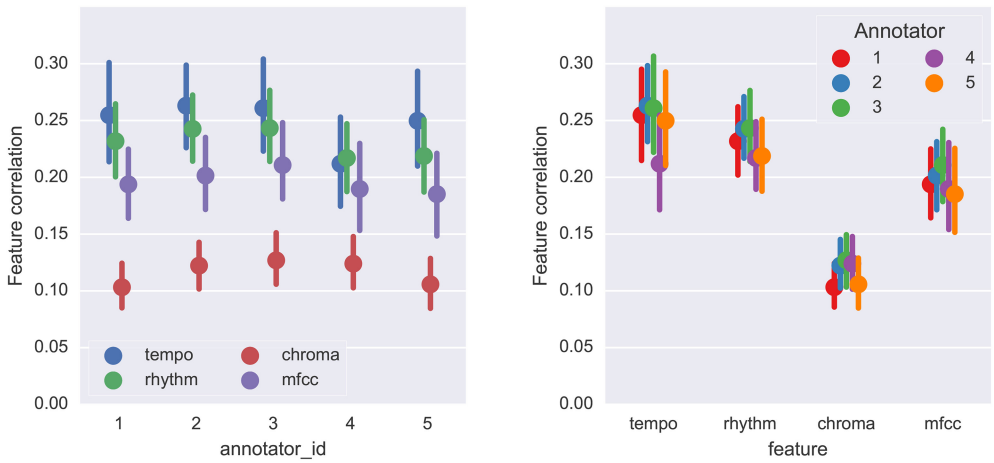


FIGURE 8 | The mean feature correlation for each feature type and annotator on the SPAM dataset. Error bars indicate the 95% confidence intervals estimated by bootstrap sampling ($n = 1,000$). **Left:** results are grouped by annotator ID; **Right:** results are grouped by feature type.

this example have again identified substantially different large-scale structures, with the first annotation correlating highly with tempo (0.57) and rhythmic (0.40) similarity as compared to the second annotator (0.16 and 0.12, respectively). The second annotator identified repeating melodic and harmonic themes that persist across changes in instrumentation and rhythm. This persistence explains the comparatively low correlation scores for the tempo and rhythm features. The two annotators appear to disagree on the relative importance of rhythmic and instrumental characteristics, compared to melodic and harmonic features, in determining the structure of the piece.

In both of these examples, and as a general trend illustrated in **Figure 8**, annotations that relied on solely on harmony produced lower correlation scores than those which align with timbre and rhythm descriptors. This is likely a consequence of the dynamic structure of harmony and chroma representations, which evolve rapidly compared to the more locally stationary descriptors of timbre, rhythm, and tempo. Chroma self-similarity matrices (**Figures 9**, **10**, bottom-left) tend to exhibit diagonal patterns rather than solid blocks of self-similar time intervals, which are easier to match against the annotation-based meet matrices (right column). It may be possible to engineer locally stable harmony representations that would be more amenable to this kind of correlation analysis, but doing so without supposing a pre-existing segmentation model is a non-trivial undertaking and beyond the scope of the present experiment.

6. EXPERIMENT 3: HIERARCHICAL ALGORITHMS

This last experiment focuses on using the L-measure to compare hierarchical results estimated by automatic approaches with those annotated by music experts. Assuming that the L-measure

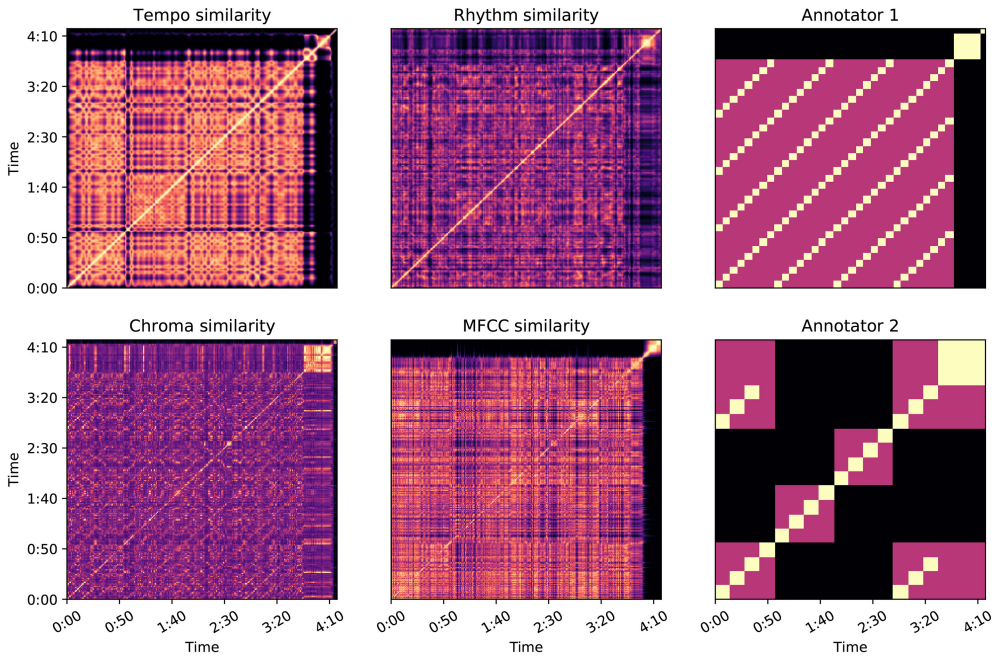


FIGURE 9 | Feature correlation for SALAMI track #410: *Erik Truffaz–Betty*, which achieves $\delta = 0.67$, L-measure = 0.25. The two annotations encode different hierarchical repetition structures, depicted in the meet matrices in the right-most column. Annotator 1's hierarchy is more highly correlated with the feature-based similarities: $z = (0.62, 0.42, 0.26, 0.48)$ for tempo, rhythm, chroma, and MFCC, compared to $z = (0.03, 0.07, 0.07, 0.04)$ for Annotator 2.

between human annotations defines the upper limit in terms of performance for the automated hierarchical segmentation task, we explore how the L-measure behaves when assessing this type of algorithms. We are particularly interested in better understanding how much room there is for improvement when designing new approaches to this task.

6.1. Methods

To the best of our knowledge, only two automatic methods that estimate hierarchical segmentations have been published with open source implementations: Laplacian structural decomposition (McFee and Ellis, 2014a), and Ordinal Linear Discriminant Analysis (McFee and Ellis, 2014b). The Laplacian method generates hierarchies of depth 10, where each layer i consists of $i + 1$ unique segment labels McFee and Ellis (2014a). For each layer index, this method first partitions the recording into a set of discontinuous clusters (segment labels), and then estimates segment boundaries according to changes in cluster membership between successive time instants. Consequently, each layer can have arbitrarily many segments, but the number of unique segment labels is always fixed.

The OLDA method, as described by McFee and Ellis (2014b), operates by agglomerative clustering of time instants into segments, resulting in a binary tree with time instants at the leaves, and the entire recording at the root. Each layer i of this tree has $i + 1$ contiguous segments, and the tree is automatically pruned based on the statistics of segment lengths and the overall track duration. This results in a hierarchy of variable depth, typically between 15 and 30 levels, where each level can be seen as splitting one segment from the previous level into two. Because OLDA only estimates segment boundaries, segment labels were estimated at each level by using the 2D-Fourier Magnitude Coefficients method (Nieto and Bello, 2014), which yields state-of-the-art results in terms of automatic flat segment label prediction. The 2D-FMC method is set to identify a maximum of 7 unique labels per level of segmentation, as this number was previously found to produce the best results in The Beatles[8] and SALAMI datasets. These sets are the most popular in the task of structural segmentation, and it is a standard practice to tune the parameters according to them (Kaiser and Sikora, 2010; Nieto and Jehan, 2013; Nieto and Bello, 2014).

The standard approach to measuring the performance of automatic algorithms is to compare the average scores derived from a sample of tracks, each of which has one "ground truth" annotation. However, as demonstrated in the previous sections, there is still significant disagreement between annotators when it comes to hierarchical segmentation, so selecting a single annotation to use as a point of reference would bias the results of the evaluation. Instead, we compared the output of each algorithm to all annotations for a given track, with results presented in terms of the full empirical distribution over scores rather than the mean score. We quantify the

[8]http://isophonics.net/content/reference-annotations-beatles

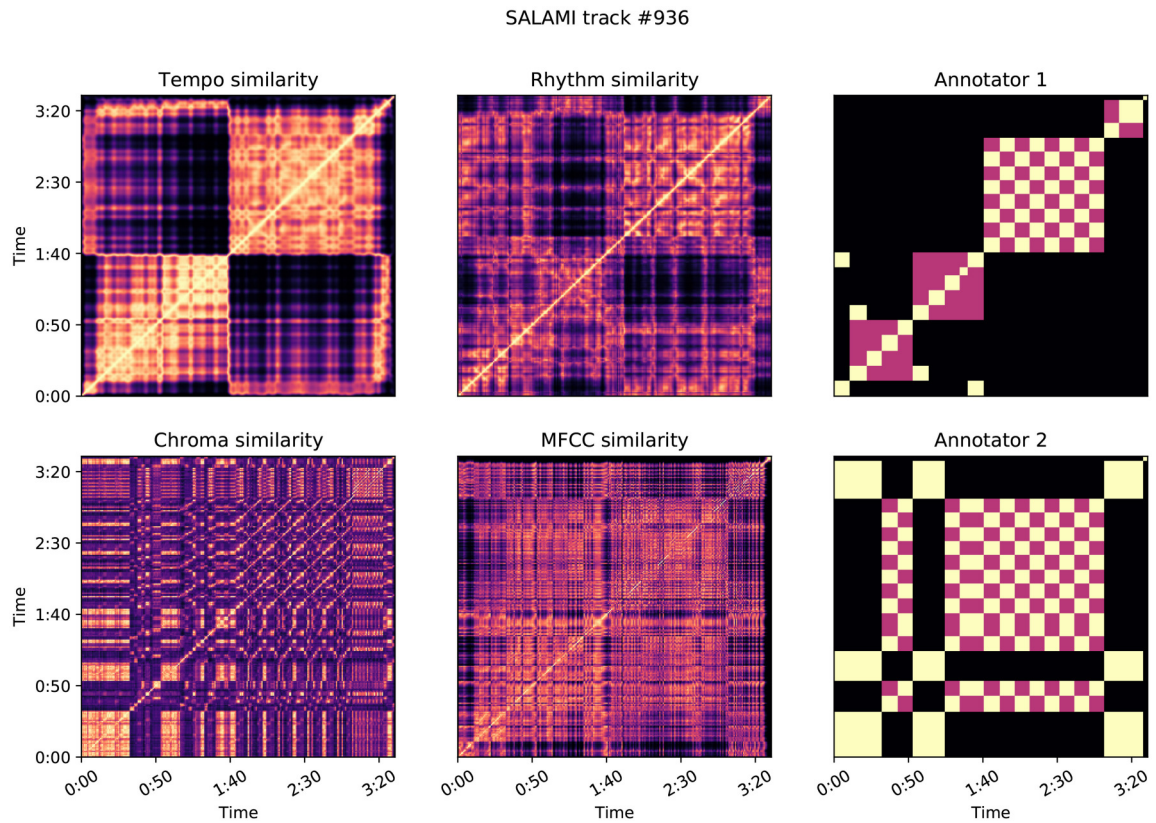


FIGURE 10 | Feature correlation for SALAMI track #936: *Astor Piazzola–Tango Aspasionado*, which achieves $\delta = 0.45$, L-measure $= 0.46$. Annotator 1 is highly correlated with the features: $z = (0.57, 0.40, 0.11, 0.25)$ for tempo, rhythm, chroma, and MFCC, compared to $z = (0.16, 0.12, 0.13, 0.25)$ for Annotator 2.

difference in distributions by the two-sample Kolmogorov-Smirnov statistic, which measures the maximum difference between the empirical cumulative distributions: a small value (near 0) indicates high similarity, a large value (near 1) indicates low similarity. For this experiment, the set of human annotations had a privileged interpretation (compared to the automatic methods), so we reported L-precision, L-recall, and L-measure separately.

Both algorithms (OLDA and Laplacian) were run on both datasets (SALAMI and SPAM) using the open-source implementations found in the Music Structure Analysis Framework, version 0.1.2-dev (Nieto and Bello, 2016). All algorithm parameters were left at their default values.

6.2. Results and Discussion

The results of the automatic hierarchical segmentation algorithm experiment are displayed in **Figure 11**. Both algorithms achieve larger average L-recall (center column) than L-precision (left column), which suggests that the automated methods, which produce much deeper hierarchies than the reference annotations, have identified more detailed structures than were encoded by the human annotators. Notably, the Laplacian method achieved a recall distribution quite close to that of the human annotators. This indicates that the L-measure is robust to differences in hierarchical depth: structures encoded in the depth-2 human annotations can also be found in the depth-10 automatic annotations.

The right column shows the total L-measure distribution (combining precision and recall). In both datasets, the Laplacian method was significantly more similar to the inter-annotator distribution than the OLDA-2DFMC method was, despite the mode at the bottom of the L-measure scale visible in **Figure 11** (right). The region of low performance can be attributed to an apparent weakness of the method on longer recordings (e.g., *SALAMI-478* at 525 s, or *SALAMI-108* at 432 s) where it tends to over-emphasize short discontinuities and otherwise label the remainder of the track as belonging primarily to one component. This behavior can also be seen in the SALAMI distribution, though such examples make up a smaller portion of the corpus, and therefore exert less influence on the resulting distribution.

The results of this experiment demonstrate a rather large gap between the distribution of inter-annotator agreement and algorithm-annotator agreement. In the examples presented here, and especially the Laplacian method, much of this gap can be attributed to low precision. Low precision may arise naturally from comparisons between deep and shallow hierarchies. Because the reference annotations in both SALAMI and SPAM have fixed depth, this effect is not observable in the inter-annotator comparison distribution. This effect suggests a trade-off between precision and recall as a function of hierarchy

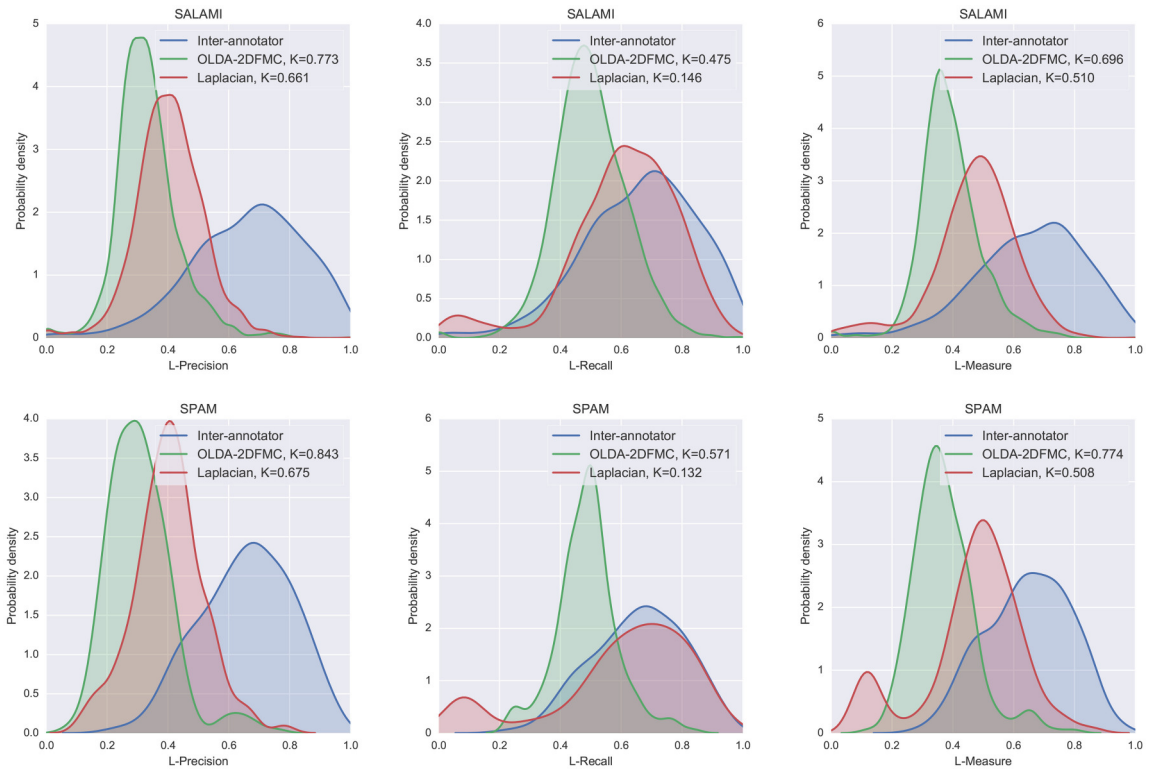


FIGURE 11 | The distribution L-measure scores for inter-annotator agreement, OLDA-2DFMC, and Laplacian on the SALAMI **(Top row)** and SPAM **(Bottom row)** datasets. The left, middle, and right columns compare algorithm L-precision, L-recall, and L-measure to inter-annotator scores. For each algorithm, the two-sample Kolmogorov-Smirnov test statistic K is computed against the inter-annotator distribution (smaller K is better).

depth. If a practitioner was interested in bounding hierarchy depth to optimize this trade-off, the L-measure would provide a means to do so.

7. GENERAL DISCUSSION

From the perspective of music informatics research, the hierarchical evaluation technique described here opens up new possibilities for algorithm development. Most existing automatic segmentation methods, in one way or another, seek to optimize the existing metrics for flat boundary detection and segment label agreement. Boundary detection is often modeled as a binary classification problem (boundary/not-boundary), and labeling is often modeled as a clustering problem. The L-measure suggests instead to treat both problems from the perspective of similarity ranking, and could therefore be used to define an objective function for a machine-learning approach to hierarchical segmentation.

As demonstrated in Section 4, the L-measure can reduce bias in the evaluation due to superficial differences between two hierarchical segmentations, which better exposes meaningful structural discrepancies. Still, there appears to be a considerable amount of inter-annotator disagreement in commonly used corpora. Disagreement is a pervasive problem in music informatics research, where practitioners typically evaluate an algorithm by comparing its output to a single "ground truth" annotation for each track in the corpus. The evaluation described in Section 6 represents a potentially viable alternative method of evaluation, which seeks not to measure "agreement" against human annotators, but rather to match the distribution of agreement *between* human annotators. This approach could be easily adapted to other tasks involving high degrees of inter-annotator disagreement, such as chord recognition or automatic tagging.

While the L-measure resolves some problems with evaluating segmentations across different levels, it still shares some limitations with previous label-based evaluation metrics. Notably, none of the existing methods can distinguish between adjacent repetitions of the same segment label (*aa*) from a single segment spanning the same time interval (*A*). This results in an evaluation which is blind to boundaries between similarly labeled segments, and therefore discards important cues indicating repetition. Similarly, variation segments—e.g., (*A*, *A*') in SALAMI notation—are always treated as distinct, and equally distinct as any other pair of dissimilar segments (*A*,*B*). While the L-measure itself does not present a solution to these problems, its ability to support hierarchies of arbitrary depth could facilitate solutions in the future. Specifically, one could augment an existing segmentation with additional lower layers that distinguish among each instance of a label, so that *a, a* decomposes into *a1, a2*, without losing the information that both segments ultimately receive the same label. Similarly,

variations could be resolved by introducing a layer above which unifies *A, A'* both as of type *A*. Because this approach requires significant manipulation of annotations, we leave it as future work to investigate its effects.

The work described here also offers both insight and a potential tool for researchers in the field of music cognition. The results from Experiment 1 reveal that flat segmentation metrics are confounded by superficial differences between otherwise consistent hierarchical annotations, while the L-measure is robust to these differences. The L-measure can therefore provide a window into the individual differences inherent in the perception of musical structure. Furthermore, the L-measure can provide a quantitative metric for directly comparing hierarchical analyses of musical form in experimental work. It can serve as a means to objectively assess response similarity between subjects on tasks that require analysis of metrical, grouping, and prolongational hierarchies.

The results of Experiment 2 present evidence for distinct modes of listening predicated on different acoustical features of the music. Comparing differences in feature correlations can help identify potential causal factors contributing to listener interpretation of musical form. The feature analysis offers objective evidence in support of qualitative observations for how and why listeners interpret musical structure differently, particularly in cases of significant disagreement.

AUTHOR CONTRIBUTIONS

All authors contributed to the research conceptually, including the experimental design and data interpretation. All authors also contributed to writing and editing the paper. Additional individual contributions are as follows: BM contributed to data preparation, software implementation, and conducted experiments; ON contributed to data preparation and conducted experiments; MF contributed to part of the statistical analysis.

FUNDING

BM acknowledges support from the Moore-Sloan Data Science Environment at New York University. JB acknowledges support from the NYU Global Seed Grant for collaborative research, and the NYUAD Research Enhancement Fund.

ACKNOWLEDGMENTS

The authors thank Jordan Smith for helpful discussions about the SALAMI dataset. We also thank Schloss Dagstuhl for hosting the Computational Music Structure Analysis seminar, which provided the motivation for much of the contents of this article. We are thankful to the reviewers for their constructive feedback.

REFERENCES

Balke, S., Arifi-Müller, V., Lamprecht, L., and Müller, M. (2016). "Retrieving audio recordings using musical themes," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Shanghai).

Barwick, L. (1989). Creative (ir) regularities: the intermeshing of text and melody in performance of central australian song. *Aus. Aboriginal Stud.* 1, 12–28.

Bharucha, J. J., Curtis, M., and Paroo, K. (2006). Varieties of musical experience. *Cognition* 100, 131–172. doi: 10.1016/j.cognition.2005.11.008

Bruderer, M. J. (2008). *Perception and Modeling of Segment Boundaries in Popular Music*. PhD thesis, Doctoral dissertation, JF Schouten School for User-System Interaction Research, Technische Universiteit Eindhoven.

Clayton, M. (1997). Le mètre et le tāl dans la musique de l'inde du nord. *Cahiers Musiques Traditionnelles* 10, 169–189. doi: 10.2307/40240271

Cohen, L. (1993). The scale representation. *IEEE Trans. Signal Process.* 41, 3275–3292. doi: 10.1109/78.258073

Cook, N. (2003). "Music as performance," in *The Cultural Study of Music A Critical Introduction*, eds M. Clayton, T. Herbert, and R. Middleton (New York, NY: Routledge), 204–214.

Davies, M. E. P., Hamel, P., Yoshii, K., and Goto, M. (2014). Automashupper: automatic creation of multi-song music mashups. *IEEE/ACM Trans. Audio Speech Lang. Process.* 22, 1726–1737. doi: 10.1109/TASLP.2014.2347135

De Sena, A., and Rocchesso, D. (2007). A fast mellin and scale transform. *EURASIP J. Appl. Signal Process* 2007, 75–84.

Deutsch, D. (ed.). (1999). "Grouping mechanisms in music," in *The Psychology of Music, 2nd Edn*. (New York, NY: Academic Press), 299–348. doi: 10.1016/B978-012213564-4/50010-X

Deutsch, D., and Feroe, J. (1981). The internal representation of pitch sequences in tonal music. *Psychol. Rev.* 88, 503–522. doi: 10.1037/0033-295X.88.6.503

Drake, C. (1998). Psychological processes involved in the temporal organization of complex auditory sequences: universal and acquired processes. *Music Percept. Interdisc. J.* 16, 11–26. doi: 10.2307/40285774

Drake, C., and El Heni, J. B. (2003). Synchronizing with music: intercultural differences. *Anna. N.Y. Acad. Sci.* 999, 429–437. doi: 10.1196/annals.1284.053

Farbood, M. M., Heeger, D. J., Marcus, G., Hasson, U., and Lerner, Y. (2015). The neural processing of hierarchical structure in music and speech at different timescales. *Front. Neurosci.* 9:157. doi: 10.3389/fnins.2015.00157

Grill, T., and Schlüter, J. (2015). "Music boundary detection using neural networks on combined features and two-level annotations," in *Proceedings of the 16th International Society for Music Information Retrieval Conference* (Málaga: Citeseer).

Herremans, D., and Chew, E. (2016). *Music Generation with Structural Constraints: An Operations Research Approach*. Louvain-La-Neuve.

Holzapfel, A., and Stylianou, Y. (2011). Scale transform in rhythmic similarity of music. *IEEE Trans. Audio Speech Lang. Process.* 19, 176–185. doi: 10.1109/TASL.2010.2045782

Kaiser, F., and Sikora, T. (2010). "Music structure discovery in popular music using non-negative matrix Factorization," in *Proceedings of the 11th International Society of Music Information Retrieval* (Utrecht), 429–434.

Kantz, H., and Schreiber, T. (2004). *Nonlinear Time Series Analysis*, Vol. 7. Cambridge, UK: Cambridge University Press.

Krumhansl, C. L., and Jusczyk, P. W. (1990). Infants' perception of phrase structure in music. *Psychol. Sci.* 1, 70–73. doi: 10.1111/j.1467-9280.1990.tb00070.x

Lerdahl, F. (1988). Tonal pitch space. *Music Percept.* 5, 315–349. doi: 10.2307/40285402

Lerdahl, F., and Jackendoff, R. (1983). An overview of hierarchical structure in music. *Music Percept. Interdisc. J.* 1, 229–252. doi: 10.2307/40285257

Levy, M., and Sandler, M. (2008). Structural segmentation of musical audio by constrained clustering. *IEEE Trans. Audio Speech Lang. Process.* 16, 318–326. doi: 10.1109/TASL.2007.910781

Lukashevich, H. (2008). "Towards Quantitative Measures of Evaluating Song Segmentation," in *Proceedings of the 10th International Society of Music Information Retrieval* (Philadelphia, PA), 375–380.

McAdams, S. (1989). Psychological constraints on form-bearing dimensions in music. *Contemp. Music Rev.* 4, 181–198. doi: 10.1080/07494468900640281

McFee, B., and Ellis, D. P. W. (2014a). "Analyzing song structure with spectral clustering," in *Proceedings of the 15th International Society for Music Information Retrieval Conference* (Taipei), 405–410.

McFee, B., and Ellis, D. P. W. (2014b). "Learning to Segment Songs With Ordinal Linear Discriminant Analysis," in *Proceeings of the 39th IEEE International Conference on Acoustics Speech and Signal Processing* (Florence), 5197–5201.

McFee, B., Nieto, O., and Bello, J. (2015a). "Hierarchical evaluation of segment boundary detection," in *16th International Society for Music Information Retrieval Conference* (ISMIR) (Malaga).

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., et al. (2015b). "Librosa: audio and music signal analysis in pyhon," in *Proceeding of the 14th Python in Science Conference* (Austin, TX), 18–25.

Nan, Y., Knösche, T. R., and Friederici, A. D. (2006). The perception of musical phrase structure: a cross-cultural ERP study. *Brain Res.* 1094, 179–191. doi: 10.1016/j.brainres.2006.03.115

Nieto, O. (2015). *Discovering Structure in Music: Automatic Approaches and Perceptual Evaluations*. Ph.d dissertation, New York University.

Nieto, O., and Bello, J. P. (2014). "Music segment similarity using 2D-Fourier magnitude coefficients," in *Proceedings of the 39th IEEE International Conference on Acoustics Speech and Signal Processing* (Florence), 664–668.

Nieto, O., and Bello, J. P. (2016). "Systematic exploration of computational music structure research," in *Proceedings of ISMIR* (New York, NY).

Nieto, O., Farbood, M. M., Jehan, T., and Bello, J. P. (2014). "Perceptual analysis of the f-measure for evaluating section boundaries in music," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)* (Taipei), 265–270.

Nieto, O., and Jehan, T. (2013). "Convex non-negative matrix factorization for automatic music structure identification," in *Proceedings of the 38th IEEE International Conference on Acoustics Speech and Signal Processing* (Vancouver, BC), 236–240.

Paulus, J., Müller, M., and Klapuri, A. (2010). "State of the art report: audio-based music structure analysis," in *ISMIR* (Utrecht), 625–636.

Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., et al. (2014). "mir_eval: a transparent implementation of common mir metrics," in *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR* (Taipei: Citeseer).

Roy, P., Perez, G., Rgin, J.-C., Papadopoulos, A., Pachet, F., and Marchini, M. (2016). "Enforcing structure on temporal sequences: the allen constraint," in *Proceedings of the 22nd International Conference on Principles and Practice of Constraint Programming - CP* (Toulouse: Springer).

Serra, J., Müller, M., Grosche, P., and Arcos, J. L. (2012). "Unsupervised detection of music boundaries by time series structure features," in *Twenty-Sixth AAAI Conference on Artificial Intelligence* (Toronto, ON).

Shaffer, L. H., and Todd, N. (1987). "The interpretive component in musical performance," in *Action and Perception in Rhythm and Music*, ed A. Gabrielsson (Stockholm: Royal Swedish Academy of Music), 139–152.

Smith, J. B., Burgoyne, J. A., Fujinaga, I., De Roure, D., and Downie, J. S. (2011). "Design and creation of a large-scale database of structural annotations," in *ISMIR*, Vol. 11 (Miami, FL), 555–560.

Smith, J. B., Chuan, C.-H., and Chew, E. (2014). Audio properties of perceived boundaries in music. *IEEE Trans. Multimedia* 16, 1219–1228. doi: 10.1109/TMM.2014.2310706

Todd, N. (1985). A model of expressive timing in tonal music. *Music Percept.* 3, 33–57. doi: 10.2307/40285321

Trehub, S. E., and Hannon, E. E. (2006). Infant music perception: domain-general or domain-specific mechanisms? *Cognition* 100, 73–99. doi: 10.1016/j.cognition.2005.11.006

Weiß, C., Arifi-Müller, V., Prätzlich, T., Kleinertz, R., and Müller, M. (2016). "Analyzing measure annotations for western classical music recordings," in *Proceedings of the 17th International Society for Music Information Retrieval Conference* (New York, NY).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

HYPOTHESIS AND THEORY
published: 22 May 2017
doi: 10.3389/fpsyg.2017.00824

A Probabilistic Model of Meter Perception: Simulating Enculturation

Bastiaan van der Weij[1]*, *Marcus T. Pearce*[2] *and Henkjan Honing*[1]

[1] *Music Cognition Group, Amsterdam Brain and Cognition, Institute for Logic, Language, and Computation, University of Amsterdam, Amsterdam, Netherlands,* [2] *Music Cognition Lab, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom*

Enculturation is known to shape the perception of meter in music but this is not explicitly accounted for by current cognitive models of meter perception. We hypothesize that the induction of meter is a result of predictive coding: interpreting onsets in a rhythm relative to a periodic meter facilitates prediction of future onsets. Such prediction, we hypothesize, is based on previous exposure to rhythms. As such, predictive coding provides a possible explanation for the way meter perception is shaped by the cultural environment. Based on this hypothesis, we present a probabilistic model of meter perception that uses statistical properties of the relation between rhythm and meter to infer meter from quantized rhythms. We show that our model can successfully predict annotated time signatures from quantized rhythmic patterns derived from folk melodies. Furthermore, we show that by inferring meter, our model improves prediction of the onsets of future events compared to a similar probabilistic model that does not infer meter. Finally, as a proof of concept, we demonstrate how our model can be used in a simulation of enculturation. From the results of this simulation, we derive a class of rhythms that are likely to be interpreted differently by enculturated listeners with different histories of exposure to rhythms.

Keywords: rhythm, cognition, meter perception, predictive coding, enculturation, computational modeling

OPEN ACCESS

Edited by:
Naresh N. Vempala,
Ryerson University, Canada

Reviewed by:
Xavier Serra,
Pompeu Fabra University, Spain
Maarten Grachten,
Austrian Research Institute for Artificial Intelligence, Austria

***Correspondence:**
Bastiaan van der Weij
b.j.vanderweij@uva.nl

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *25 October 2016*
Accepted: *05 May 2017*
Published: *22 May 2017*

Citation:
van der Weij B, Pearce MT and Honing H (2017) A Probabilistic Model of Meter Perception: Simulating Enculturation. Front. Psychol. 8:824. doi: 10.3389/fpsyg.2017.00824

1. INTRODUCTION

In a variety of settings, perception appears to be tuned to statistical properties of the environment. It has for example been found that certain properties of neuron receptive fields in early visual processing (Olshausen and Field, 1996) and early auditory processing (Smith and Lewicki, 2006) emerge from information theoretically efficient learning algorithms trained respectively on natural images or sounds. Such tuning, it has been suggested, happens both on an evolutionary time-scale through gradual adaptation, and on an ontogenetic time scale, through brain plasticity (Clark, 2013).

The perception of meter in music appears to be shaped by cultural differences in musical conventions. Exposure to rhythmically different music has been shown to influence perception from an early age (Hannon and Trehub, 2005a,b), but such shaping possibly continues into adulthood (Creel, 2011, 2012). In the current paper, we hypothesize that considering meter perception from the perspective of *predictive coding* (Rao and Ballard, 1999; Friston, 2005; Clark, 2013) can help to understand how meter perception is shaped by one's environment.

Rhythm is an important component of music traditions all over the world (Savage et al., 2015). When listening to rhythms, onsets in the rhythm are perceived relative to a periodic and hierarchically organized framework of beats (Honing, 2013). This mental framework, called meter,

is induced in the mind of the listener by the rhythm. The relation between rhythm and meter is complex. For a meter to be perceived, not every beat in the meter needs to coincide with onsets in the rhythm. In many cases, listeners can, through conscious effort, alter their metrical interpretation of a rhythm. At the same time, not every meter is equally easy to hear in every rhythm. Meter, once induced, tends to show a certain resistance to change. Therefore, meter perception is a fundamentally incremental process (Longuet-Higgins and Steedman, 1971): the same rhythmic passage can sound different depending on the meter induced by the rhythm preceding the passage (Honing, 2013).

The organizing structure of meter is commonly described as a hierarchy of pulses, yielding a periodic pattern of metrical accents varying in salience at different points in time. Metrical accent, or metrical salience, is commonly treated as a proxy for temporal expectation, or the probability of an event onset at a particular pulse (Palmer and Krumhansl, 1990). By investigating a corpus of Western classical music, Palmer and Krumhansl (1990) found that the distribution of onsets over different positions relative to the meter reflected theoretical descriptions of metrical hierarchy (Lerdahl and Jackendoff, 1983). Using a goodness-of-fit paradigm, Palmer and Krumhansl (1990) found that temporal *expectations* of North-American listeners also reflect metrical hierarchy, although musicians showed evidence of deeper hierarchical differentiation than non-musicians. Based on these findings, Palmer and Krumhansl (1990) suggested that composers communicate meter to listeners through the distribution of onsets at different metrical positions. Listeners, in turn, acquire their knowledge about meter through the distribution of onsets over metrical positions in the music they are exposed to.

More recent work has addressed the question of whether hierarchical organization of onset distributions is a general property of rhythmic organization or whether it is specific to Western classical music and related styles. Holzapfel (2015), for instance, found that in traditional Turkish makam music, the distribution of onsets is modulated by the specific *usul*—a type of rhythmic mode, corresponding in some ways to meter—underlying a piece. Furthermore, the distribution of onsets within one usul in Turkish makam music does not always exhibit hierarchical organization. London et al. (2016) found that peaks in onset distributions in a corpus of Malian drumming recordings are not periodically spaced. London et al. (2016) conclude that in makam music and Malian drumming, distributions of onsets do reflect metrical structure, but this structure is not always isochronous or strictly hierarchical.

London et al. (2016) point out that their and Holzapfel (2015) results question a basic assumption made by many computational models, as well as empirical studies, namely that metrical accent is equivalent to the likelihood of an onset. A more likely alternative is that metrical expectations are derived from extensive exposure to a musical idiom, by which, beyond distributions of onsets and style-specific, stereotypical rhythmic patterns associated with certain meters are learned.

Consistent with this suggestion, an increasing number of empirical studies show that rhythm perception is affected by enculturation (cf. Morrison and Demorest, 2009). For example, Bulgarian or Macedonian adults are better in detecting metrical violations in meters with a non-isochronous *tactus* level—the level of beat that listeners are most likely to tap along with—(e.g., 5/8 or 7/8) than North-American listeners (Hannon and Trehub, 2005a). This effect appears to be specific to complex meters to which the listeners have been exposed (Hannon et al., 2012).

There have also been a number of observations in the ethnomusicological literature suggesting that individuals from different cultures perceive rhythms differently. For example, during field work in the Bolivian Andes, while studying Easter songs from Northern Potosí, (Stobart and Cross, 2000) realized that while they had assumed many of the tunes where indisputably anacrustic (i.e., a rhythm starting on an off-beat), the local populations appeared to perceive them as beginning on a downbeat. Another example is provided by rhythms from West-African Sub-Saharan musical cultures, which are characterized by a great deal of metrical ambiguity (Locke, 1982). In particular, many of these rhythms can be interpreted as having a binary or ternary pulse. While individuals from West-African cultures appear to perceive both pulses with equivalent ease, it can take great effort for Western listeners to hear the ternary pulse in some of these rhythms.

The idea that perception, in general, is shaped by statistical properties of the environment is not new (e.g., Barlow, 1961). However, it recently has been developed into a framework which has been argued to bear the promise of providing an overarching theory of perception (Clark, 2013). Under the name of predictive coding (Rao and Ballard, 1999), this framework firmly grounds perception in prediction, based largely on previous sensory experience. In fact, the theory proposes that the brain's primary occupation is to explain sensory input using hierarchical generative models gleaned from previous experience (Clark, 2013). Such models are realized in a hierarchical organization of layers. The lowest layer in the hierarchy represents sensations received directly from the senses. Through feed-forward connections, information travels upward in the hierarchy. Meanwhile, layers higher up in the hierarchy attempt to predict information, propagated by layers below. These predictions are cast to lower layers through feedback connections. Successful prediction cancels out the upward propagation of information. As a result, only *prediction error*, information that higher layers failed to predict, propagates upwards in the hierarchy. Based on prediction error, layers gradually adapt their processing characteristics in a way that minimizes prediction error with respect to layers lower in the hierarchy. By this process of adaptation, the hierarchy of layers is gradually shaped into a *generative* model of sensations, where layers higher up in the hierarchy track causes in the external world that underlie the received sensations (Friston, 2005). From an information-theoretic point of view, the resulting coding scheme is highly efficient: the more accurate the top-down predictions, the less bottom-up information is left to be processed.

We propose a predictive coding account of meter perception that involves statistical learning of musical rhythms and generation of probabilistic expectations for event timings. Meters

are modeled as distinct causes underlying the musical surface. Inferring the underlying meter from rhythm allows the rhythm to be related to rhythms previously heard in that meter, which may help prediction performance. Enculturation is modeled by estimating the parameters of the generative model on a corpus of quantized rhythms annotated with meter. Since the model learns the statistical properties of rhythms through exposure and performs metrical inference based on these, it has the potential to simulate enculturation effects in meter perception.

The paper is organized in six sections. In the remaining part of the current section, Section 1.1 develops an account of meter perception based on predictive coding, while Section 1.2 discusses relevant work in computational modeling of music perception. Section 2 presents the probabilistic model of meter perception in detail, concluding with a set of behaviors we expect the model to exhibit. Section 3 presents the methods used in a series of simulations designed to test these behaviors, while Section 4 presents the results of the simulations. Section 5 discusses the results in the context of the existing literature and includes implications for future research.

1.1. Meter Perception as Predictive Coding

The dynamic interaction of top-down and bottom-up processing postulated by predictive coding is reminiscent of dynamic interaction of bottom-up meter-induction and top-down influence exerted by the induced meter, as pointed out by Vuust and Witek (2014).

The hypothesis we explore in this paper is that predictive coding can explain how meter perception is influenced by enculturation. To explore the consequences of this idea, we present a probabilistic model of meter perception, based on an empirical Bayes scheme. Empirical Bayes schemes describe how generative systems, such as the generative models posited by predictive coding, are updated by experience (Friston, 2005). We model meters as virtual causes underlying the rhythmic surface: a meter imposes constraints the likelihood of rhythms. A listener commanding an appropriate generative model reflecting this relationship (i.e., how rhythms are generated from meters), can, when presented only with a rhythmic surface, infer the underlying meter. This process of inferring underlying causes (meters) of experienced sensations (rhythms) involves inverting the generative model of those sensations (which are the end-product of the generative process). We hypothesize that interpreting the rhythm in the context of an inferred meter will reduce the discrepancy between predicted and experienced sensations. In other words, inferring meter makes the rhythm more predictable.

The generative model includes prior expectations, obtained from previous experience, about which metrical categories are likely to occur in general. For example, meters with non-isochronous pulses ("complex" meters) are relatively uncommon in Western-European music, but much more common in music from the Balkans and Eastern Mediterranean region. Listeners from these regions may be more likely to interpret a rhythm in a meter with non-isochronous pulses than listeners from Western Europe. These kind of prior biases might underlie the findings of Hannon and Trehub (2005a) mentioned in the previous section.

Metrical categories favored by prior biases entail expectations regarding the surface structure of rhythms. As bottom-up evidence from the rhythm begins to flow in, these (top-down) expectations are either confirmed or violated. Prediction error results from a violation of the top-down expectations by the incoming evidence. To reduce prediction error, the listener revises their metrical interpretation of the rhythm, which in turn alters the flow of top-down predictions. A predictive coding perspective of meter perception thus posits a dynamic interplay between bottom-up evidence and top-down expectations.

Crucially, both prior biases toward certain meters and the dependencies between meter and the rhythmic surface—which rhythms can be generated by a certain meter—are the result of previous exposure. The generative model in the mind of the listener underlying these representations is carved out by previous experience in predictive processing of rhythmic signals. Since the statistical properties of rhythms vary between styles (e.g., Holzapfel, 2015; London et al., 2016), the processing biases of listeners with significant differences in their exposure to musical styles are likely to vary as well.

1.2. Related Work

Our approach in some respects resembles other recent probabilistic models, in particular a generative model presented by Temperley (2007). Temperley (2007, ch. 2) models meter perception as probabilistic inference on a generative model whose parameters are estimated using a training corpus. Meter is represented as a multi-leveled hierarchical framework, which the model generates level by level. The probability of onsets depends only on the metrical status of the corresponding onset time. Temperley (2009) generalizes this model to polyphonic musical structure, and introduces a metrical model that conditions onset probability on whether onsets occur on surrounding metrically stronger beats. This approach introduces some sensitivity to rhythmic context into the model. In later work, Temperley (2010) evaluates this model, the *hierarchical position model*, and compares its performance to other metrical models with varying degrees of complexity. One model, called the first-order metrical position model, was found to perform slightly better than the hierarchical position model, but this increase in performance comes at the cost of a higher number of parameters. Temperley concludes that the hierarchical position model provides the best trade-off between model-complexity and performance.

In a different approach, Holzapfel (2015) employs Bayesian model selection to investigate the relation between usul (a type of rhythmic mode, similar in some ways to meter) and rhythm in Turkish makam music. The representation of metrical structure does not assume hierarchically organization, allowing for arbitrary onset distributions to be learned. Like the models compared by Temperley (2010), this model is not presented explicitly as a meter-finding model, but is used to investigate the statistical properties of a corpus of rhythms.

The approach presented here diverges from these models in that it employs a general purpose probabilistic model of *sequential* temporal expectation based on statistical learning (Pearce, 2005) combined with an integrated process of metrical inference such that expectations are generated given an inferred

meter. The sequential model is a variable-order metrical position model. Taking into account preceding context widens the range of statistical properties of rhythmic organization that can be learned by the model. In particular, the model is capable of representing not only the frequency of onsets at various metrical positions, but also the probability of onsets at metrical positions conditioned on the preceding rhythmic sequence. The vastly increased number of parameters of this model introduces a risk of *over-fitting*; models with many parameters may start to fit to noise in their training data, which harms generalization performance. However, we employ sophisticated smoothing techniques that avoid over-fitting (Pearce and Wiggins, 2004). Furthermore, we to some extent safe-guard against over-fitting by evaluating our model using cross-validation.

2. THE PROBABILISTIC MODEL

In this section and the sections that follow, we use the words metrical category and metrical interpretation in a specific sense. *Metrical categories*, denoted by m, represent different metrical frameworks in which rhythms can be interpreted. Metrical categories correspond directly to time signatures taken from scores. Each metrical category has an associated *period*, denoted by T_m. The period is encoded as a discrete number representing the duration of one bar of m in basic quantized units of time (see Section 2.1). The *phase* parameter, ϕ, encodes how a metrical category aligns with the rhythmic surface. More precisely, ϕ encodes the time-interval between the downbeat of the first bar and the time point marked by zero in the encoding of the rhythmic pattern. Together, a metrical category and phase form a *metrical interpretation*.

The approach described below deals not with real audio signals. Instead, the musical surface is represented as a sequence of events. Each event corresponds to a note, as it might be found in a musical score. The nth event in a sequence is denoted by e_n. A sequence of events, starting at event n and ending at event m is denoted by $\mathbf{e}_n^m$. Section 2.1 provides more details the representation of rhythmic patterns.

Predictive coding postulates internal generative models reflecting the causal structure of the external world. In analogy to this, we model meter perception as the inversion of a generative model of rhythms. Enculturation through exposure to rhythms is modeled by deriving the parameters of the generative model from a corpus of rhythms annotated with metrical interpretation. During listening, the metrical category underlying a given rhythm is generally not known to the listener. Instead, it has to be inferred from rhythmic surface, which is assumed to result from the generative model. The likelihood of a metrical interpretation given an observed rhythm (i.e., a sequence of events) can be inferred from the generative model through the application of Bayes' formula, as shown in Equation (1).

$$\overbrace{p(m,\phi|\mathbf{e}_0^n)}^{\text{posterior}} = \frac{\overbrace{p(\mathbf{e}_0^n|m,\phi)}^{\text{likelihood}}\overbrace{p(m,\phi)}^{\text{prior}}}{\underbrace{p(\mathbf{e}_0^n)}_{\text{evidence}}}. \tag{1}$$

Two factors play a role in calculating the likelihood of a metrical category: The *a priori* likelihood of the metrical category itself, operationalized here as the metrical category's conventionality. In Equation (1), this distribution is labeled *prior*. The other factor is the likelihood of the rhythmic pattern given a certain metrical structure. In Equation (1), this function is labeled *likelihood*. The distribution over metrical interpretations inferred from the observed events is called the *posterior* distribution. The factor labeled *evidence* in Equation (1) is a constant with respect to metrical interpretation. It ensures that the distribution sums to unity.

The proposed generative model is illustrated in **Figure 1**. To generate a rhythm, a metrical category is first generated from a distribution, $p(m)$, reflecting the prior likelihood of metrical categories. Next, a phase is sampled from a uniform distribution over a range of discrete phases allowed in m. From a model associated with the selected metrical category, events are then generated in an incremental fashion. As can be seen in **Figure 1**, the likelihood of an event is conditioned on underlying metrical category and preceding events.

Equation (1) can be expanded into the incremental and recursive equation shown in Equation (2). This equation expresses the posterior distribution given all events as proportional to the product of the likelihood of the last event, e_n and the posterior given all but the last event, $\mathbf{e}_0^{n-1}$. Inferring the posterior incrementally after each event by refining the posterior that resulted from the previous events can be interpreted intuitively as the listener integrating the (bottom-up) information provided by each event into their (top-down) beliefs about the underlying metrical category. Note that the evidence normalization constant has been omitted for clarity.

$$\overbrace{p(m,\phi|\mathbf{e}_0^n)}^{\text{per-event posterior}} \propto \begin{cases} \overbrace{p(e_n|m,\phi,\mathbf{e}_0^{n-1})}^{\text{per-event likelihood}}\overbrace{p(m,\phi|\mathbf{e}_0^{n-1})}^{\text{updated prior}} & \text{if } n>0, \\ p(e_n|m,\phi)p(m,\phi) & \text{else.} \end{cases} \tag{2}$$

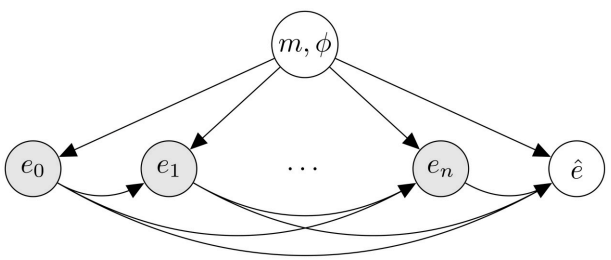


FIGURE 1 | Conditional dependency relations assumed by the model between its probabilistic variables visualized as a graphical model (Bishop, 2007, ch. 8). Shaded nodes represent observed variables, unshaded nodes represent unobserved, *hidden* variables. Each node in the graph is associated with a discrete probability distribution. If one or more arrows terminate at a node, its associated probability distribution is conditioned on the node(s) that the arrows originate from. Nodes labeled e_n represent musical events indexed by n. The hidden variable at the top represents a metrical interpretation. The hidden variable labeled $\hat{e}$ represents a predicted subsequent event.

To infer the posterior distribution over metrical interpretations, Equation (2) is evaluated for a set of possible metrical interpretations. This set is constrained to include only metrical categories that occur in the model's training data. The number of different phases considered per metrical category depends on the period of the category, T_m.

To evaluate Equation (2), two probability distributions need to be approximated: the prior distribution over metrical interpretations, $p(m, \phi)$, and the likelihood function $p(\mathbf{e}_0^n | m, \phi)$. We discuss both in the following paragraphs.

First, we consider estimating the prior, which uses supervised learning from a corpus of rhythms labeled with metrical category. The parameters of the distribution defining the *a priori* likelihood of metrical categories (not phases), $p(m)$, are set to their maximum likelihood estimate, namely the relative frequency of occurrence of a metrical category in the empirical training data.

$$p(m) = \frac{N_m}{N}, \tag{3}$$

where N_m is the number of times m was observed in the training data and N is the total number of training examples (rhythms) in the training data.

The prior distribution over metrical interpretations (i.e., the joint distribution over phase and metrical category) is defined as follows:

$$p(m, \phi) = \frac{p(m)}{\sum^{m'} T_{m'} p(m')}. \tag{4}$$

Each metrical interpretation is assigned a probability proportional to the probability of its category. This definition entails a reweighing of metrical categories to compensate for the duration of their periods; it prevents meters with long periods (many possible phases) from being at a disadvantage due to the uniform spreading out of their probability over a large number of phases.

Second, we consider estimating the likelihood. Calculating the likelihood of an observed rhythm given a hypothesized metrical interpretation involves two steps: First, the rhythm under consideration is *interpreted* in a hypothesized metrical interpretation specified by m and ϕ. Interpretation is operationalized in the present model as converting the events in the rhythm into a sequence of symbols encoding the position of each event relative to the beginning of the bar in which it occurs under the currently considered metrical interpretation. The details of this conversion are discussed in Section 2.3. Second, the likelihood of the resulting sequence of symbols is estimated using an unsupervised probabilistic model trained on metrically interpreted rhythms in the training corpus annotated with the same metrical category, m. The likelihood that a rhythm is generated by given metrical interpretation thus becomes the likelihood of the sequence of symbols resulting from metrically interpreting the event onset times in the rhythm. The likelihood of the metrically interpreted rhythm, in turn, is determined on the basis of a corpus of rhythms belonging to the same metrical category.

Equation (2) decomposes the likelihood function into the product of the per-event likelihoods, i.e., the likelihood of each (metrically interpreted) event given the sequence of (metrically interpreted) preceding events. In the present work, IDyOM (Pearce, 2005) is used to approximate the per-event likelihood function.

IDyOM is a flexible modeling framework based on variable-order Markov modeling combined with a multiple-viewpoint system for music prediction (Conklin and Witten, 1995). It was designed for modeling dynamically changing auditory expectations, based on long-term and short-term statistical learning, which evolve as a piece of music unfolds. Empirical research has demonstrated that IDyOM accurately simulates listeners' predictive processing of melody in many perceptual tasks involving pitch expectation (Pearce, 2005; Pearce et al., 2010; Omigie et al., 2012, 2013), uncertainty (Hansen and Pearce, 2014), segmentation (Pearce et al., 2010) and emotional response (Egermann et al., 2013; Gingras et al., 2015).

Section 2.3 describes how our model is implemented on top of IDyOM. While the present model does not make use of the full range of modeling opportunities that the multiple-viewpoint approach has to offer, presenting the model as an extension of IDyOM highlights the continuity between the two probabilistic modeling approaches.

Aspects of multiple viewpoint systems and IDyOM relevant to the present model are introduced in Section 2.1 and Section 2.2. Our treatment of this topic is far from complete; for a complete overview, we refer the reader to Conklin and Witten (1995) and Pearce (2005).

2.1. Representation of Rhythmic Patterns

Multiple viewpoint systems represent the musical surface as a sequence of multi-dimensional datapoints encoding basic attributes of musical events, such as pitch, onset time and duration. These basic attributes of events are accessed through *viewpoints*. A viewpoint maps sequences of events, rather than individual events, to an element of its corresponding *type*, τ. The set of all possible elements of a type τ is called the *alphabet* of τ and is denoted by $[\tau]$. A viewpoint function may be undefined for some sequences of events. The inter-onset-interval viewpoint, for example, is undefined for the sequence $\mathbf{e}_0^0$, which consists of only a single event, e_0. Hence, a viewpoint is defined by a partial function that maps sequences of events to elements of a type

$$\Psi_\tau : \zeta^* \rightharpoonup [\tau],$$

where the symbol ζ^* denotes the set of all possible sequences of events.

A distinction between two types of viewpoints is made. A *basic viewpoint* simply returns one of the basic attributes of the last event in the sequence to which it was applied (i.e., a projection function). The alphabet of a basic viewpoint is determined by the set of values of its corresponding attribute observed in the training corpus (see Section 2.2). A *derived viewpoint* derives more abstract attributes from one or more basic attributes of one or more basic events. Its alphabet can be derived from the alphabets of the basic viewpoints that the viewpoint is derived of. The inter-onset-interval viewpoint and metrical viewpoints introduced in Section 2.3 are examples of derived viewpoints. For

derived viewpoints, multiple different sequences of events may map to the same element.

The function Φ_τ returns the sequence of viewpoint elements of type τ obtained by applying the viewpoint function Ψ_τ incrementally to to all prefixes of the sequence in order of increasing length:

$$\Phi_\tau(\mathbf{e}_0^n) = \begin{cases} \Phi_\tau(\mathbf{e}_0^{n-1})\Psi_\tau(\mathbf{e}_0^n) & \text{if } \Psi_\tau(\mathbf{e}_0^n) \neq \perp, \\ \Phi_\tau(\mathbf{e}_0^{n-1}) & \text{else,} \end{cases}$$

where $\perp$ is a symbol indicating that the viewpoint is undefined for the given sequence of events.

The model introduced here makes use of a single basic viewpoint, namely `on`, returning the onset attribute of the last event in a sequence, and a set of derived metrical viewpoints. The alphabet of onset, `[on]`, contains natural numbers that encode the temporal position of a note as an integer-multiple of basic quantized units. To obtain a finite, meaningful alphabet for `on`, the onset alphabet is constructed online by adding the set of inter-onset intervals encountered in the training data to the onset of the previous event.

2.2. Predicting Musical Events

Predicting sequences of musical events in IDyOM requires specifying a set of viewpoints, $\tau_0, \tau_1, \cdots, \tau_n$, on which to base predictions. A predictive model is associated with each of these viewpoints. Each predictive model is trained on the set of symbol sequences obtained by applying the associated viewpoint function Φ_τ to all event sequences in the training corpus. To approximate the predictive distribution for a future event, $p(\hat{e}|\mathbf{e}_0^n)$, given a sequence of preceding events $\mathbf{e}_0^n$, the function Φ_τ is applied, once for each of the specified viewpoints, to $\mathbf{e}_0^n$ to obtain a set of sequences of viewpoint elements.

The per-viewpoint predictions, $p_\tau(\Psi_\tau(\hat{e})|\Phi_\tau(\mathbf{e}_0^n))$ are then combined into a single event prediction, using a mechanism that involves a weighted geometric mean. Some subtleties are involved in converting the predictive distributions to a single domain so that they can be combined (Pearce, 2005, ch. 7). These need not concern us, as the model proposed here only uses a single viewpoint to predict a single attribute of the event representation (although it could be extended in the future to use multiple viewpoints).

IDyOM thus reduces the challenge of estimating $p(\hat{e}|\mathbf{e}_0^n)$ to the parallel prediction of symbol sequences by estimating $p_\tau(\Psi_\tau(\hat{e})|\Phi_\tau(\mathbf{e}_0^n))$ for each viewpoint $\tau_0, \tau_1, \cdots, \tau_n$. The (domain-general) method employed by IDyOM for predicting symbol sequences is based on a data-compression scheme called prediction by partial matching (PPM), introduced by Cleary and Witten (1984). Pearce and Wiggins (2004) provide an overview of various modifications and improvements to the original PPM scheme that have been proposed over the years, and compare their performance using an information-theoretic performance measure (see Section 2.4). IDyOM implements multiple prediction schemes and furthermore allows predictions to be based on two separate models: a long-term model trained on a corpus of training data and a short-term model trained, online, on only the current sequence of events. In our simulations, we use only a long-term model (see Pearce et al., 2005), employing a PPM* scheme using method C (Moffat, 1990) for calculating escape probabilities and adapted to use interpolated smoothing—the configuration Pearce and Wiggins, 2004 found to yield the best results for a long-term model. A parameter called model order-bound parameter limits the amount of previous events taken into account in the predicting the next event, $\hat{e}$: An order-bound of b means that it is assumed that $p(\hat{e}|\mathbf{e}_0^n) \approx p(\hat{e}|\mathbf{e}_{n-b}^n)$. While Pearce and Wiggins (2004) found that an unbounded model order worked best, the present paper presents results for varying model order-bounds of up to four.

2.3. Metrical Viewpoints, Metrical Models, and Metrical Inference

The per-event likelihood function in Equation (2) is a predictive distribution that, based on events observed so far and a hypothesized metrical interpretation, specified by m and ϕ, predicts the next event. This relies on interpreting the sequence of events in the given metrical interpretation and estimating the likelihood of the resulting sequence of symbols given a predictive model of such sequences in the provided metrical category. Interpretation of a rhythm in a specific metrical interpretation is achieved in IDyOM through the introduction of a set of *metrical viewpoints*. Metrical viewpoints transform a sequence of absolute onset times into a sequence of symbols that depend on the metrical interpretation implemented by the viewpoint.

The general form of a metrical viewpoint $\tau_{m,\phi}$ is

$$\Psi_{\tau_{m,\phi}}(\mathbf{e}_0^n) = f(m, \phi, \mathbf{e}_0^n),$$

where f is a function that implements the metrical interpretation given a phase and metrical category.

The present model uses a simple metrical interpretation function that returns the *metrical position* of an onset. This function makes few assumptions about the structural organization of meter, and can accommodate complex, non-isochronous meters. The metrical position of an onset is defined as its position relative to the period and phase of an interpretation. The general definition of the resulting metrical position viewpoint, `mp`, is given below

$$\Psi_{\mathtt{mp}_{m,\phi}}(\mathbf{e}_0^n) = \left(\Psi_{\mathtt{on}}(\mathbf{e}_0^n) - \phi\right) \mod T_m,$$

where the viewpoint `on` is a basic viewpoint that returns the onset of the last event in a sequence of events.

One metrical viewpoint is created for each metrical interpretation considered by the model by instantiating m and ϕ to a specific value.

The alphabet of the `mp` viewpoint is given by

$$[\mathtt{mp}_{m,\phi}] = \{0, 1, \cdots, T_m - 1\}.$$

Using metrical viewpoints, metrical inference can be implemented on top of the standard IDyOM machinery, with one important caveat: the predictive model of a metrical viewpoint, $\tau_{m,\phi}$ is trained only on those sequences in the training

data that have been annotated with metrical category m. Hence, the predictability of a metrically interpreted rhythm depends only on rhythms previously observed in the corresponding metrical category.

One further subtlety needs to be addressed to complete the model. Note that the per-viewpoint predictive distributions mentioned in Section 2.2 are defined over a viewpoint's alphabet $[\tau]$. In order to predict the onset of the next event this alphabet needs to be mapped back to the alphabet of the onset viewpoint, [on]. However, any metrical position in [mp] theoretically corresponds to an infinite number of periodically spaced onset times. To be able to generate predictions for *specific* onset times, and for metrical inference to work correctly, it is necessary that the alphabet of a metrical viewpoint maps to unique onset times. This can be achieved by *linking* the metrical position viewpoint to another metrical viewpoint, which encodes the distance in bars between the last event and the predicted event.

The equation below defines the bar distance viewpoint, bd in terms of an intermediate metrical viewpoint, bn (bar number), which calculates the number of bars elapsed between time zero and the onset of the last event.

$$\Psi_{\mathtt{bd}_{m,\phi}}(\mathbf{e}_0^n) = \Psi_{\mathtt{bn}_{m,\phi}}(\mathbf{e}_0^n) - \Psi_{\mathtt{bn}_{m,\phi}}(\mathbf{e}_0^{n-1}),$$

where metrical viewpoint bn is defined as

$$\Psi_{\mathtt{bn}_{m,\phi}}(\mathbf{e}_0^n) = \text{integer}\left(\frac{\left(\Psi_{\mathtt{on}}\mathbf{e}_0^n - \phi\right)}{T_m}\right).$$

A linked viewpoint is a special case of a derived viewpoint composed of a number of constituent viewpoints. The elements of linked viewpoints are tuples containing the values of the constituent viewpoints. A linked viewpoint composed of $\tau_1, \cdots, \tau_n$ is denoted by $\tau_1 \otimes \cdots \otimes \tau_n$, its alphabet is given by the Cartesian product of the constituent viewpoints' alphabets: $[\tau_1] \times \cdots \times [\tau_n]$.

The linked metrical viewpoint used in our simulations is denoted by $\mathtt{mp} \otimes \mathtt{bd}$, and encodes metrical position and distance in bars between the penultimate and last event. Elements in the alphabet of this viewpoint have a one-to-one correspondence to elements in [on].

To summarize: metrical viewpoints and separate predictive models per metrical category enable using IDyOM to estimate the per-event likelihood function in Equation (2). In this model, the likelihood of a metrical interpretation m depends on the predictability of the sequence of symbols that results from interpreting the rhythm in that metrical interpretation. This predictability in turn depends on the set of rhythms previously observed in m.

2.4. Expectation and Information Content

We have focussed our discussion so far on the issue of inferring a posterior distribution over metrical interpretations. In order to calculate prediction error, it is necessary to derive the predictive distribution over future note onsets given a preceding rhythmic context and an inferred meter.

To estimate prediction error, we look at the amount of information communicated by each observation. Although it is sometimes referred to as cross-entropy (e.g., Manning and Schütze, 1999, ch. 2), we call this quantity the *information content* (MacKay, 2003) of an event. Information content is defined as the negative logarithm of the likelihood of observing the next event given the predictive distribution conditioned on the sequence of events observed so far:

$$h(\hat{e}|\mathbf{e}_0^n) = -\log_2 p(\hat{e}|\mathbf{e}_0^n). \tag{5}$$

In an information-theoretic sense, this quantity is equivalent to prediction error. An unlikely (unexpected) event results in a high prediction error, signaled by high information content. Conversely, a likely event results in a low prediction error, signaled by low information content.

The predictive distribution corresponds to the probability distribution associated with the hidden variable labeled $\hat{e}$ in the graphical model in **Figure 1**. This distribution is obtained from the generative model by marginalizing out meter and phase from the posterior distribution inferred from the preceding events:

$$p(\hat{e}|\mathbf{e}_0^n) = \sum^{m}\sum^{\phi} p(\hat{e}|m,\phi,\mathbf{e}_0^n)p(m,\phi|\mathbf{e}_0^n), \tag{6}$$

where the summation over meters sums over all metrical categories considered by the model, $m \in M$, and the summation over phases sums over all possible phase of category m, $\phi \in \{0, 1, \cdots, T_m - 1\}$.

Equation (6) shows that the prediction of the onset of the next event is subject to top-down influence from the distribution over metrical interpretations inferred from bottom-up information from the events observed so far.

2.5. Hypotheses

We expect an accurate computational model of human meter perception to show certain patterns of behavior. First, we expect it to be able to infer meters that agree with the time signatures in notated scores (Longuet-Higgins and Lee, 1982; Temperley, 2004). Second, we argued that the metrical knowledge, acquired by listeners through exposure to a musical idiom, is characterized not only by the distribution of onsets over metrical positions, but also by the probabilistic properties of how rhythms in particular meters sequentially unfold. Thus, we expect that a model that can learn such properties will lead to increased performance in finding time signatures notated in scores compared to a similar model that does not learn these properties. Third, we argued above that categorizing rhythms into metrical categories can plausibly be regarded as a strategy to reduce prediction error for those rhythms. Therefore, we expect that our model will show better performance in predicting the timing of musical events than a comparable model that is agnostic of meter. Fourth, we expect that our model will simulate enculturation by showing sensitivity to the statistical properties of the rhythms it was trained on. A model trained on rhythms with similar statistical properties as the rhythms it is evaluated on will perform better than a model that was trained on rhythms with different statistical properties. If the statistical properties of rhythms originating

from two cultures with different cultural practices regarding rhythm are sufficiently different, we expect that a model trained on rhythms from the same culture as the rhythms it is evaluated on will outperform a model trained on rhythms from a culture with different rhythmic practices. We evaluate these expectations in Sections 3 and 4.

3. METHODS

3.1. Resolution of Onset Time and Phase

For reasons of computational efficiency, the resolution the phase parameter of metrical interpretations is restricted to sixteenth notes. This means that, for example, in the 3/4 category twelve different phases are possible (since the duration of one 3/4 bar is twelve sixteenth notes). Since all onset times in rhythms used in this study encode distance from the beginning of the first bar in the annotated meter, the correct phase of a rhythm can be represented under any phase resolution. The representation of rhythms in a phase of zero does not influence the evaluation: as far as the model is concerned, all phases are initially equally likely since the prior distribution over phase is uniform. The presence of 32th notes and 16th-note triplets in the training data requires that onset times are represented as integer multiples of symbolic units corresponding to 96th notes.

3.2. Training Data

Except for one artificially constructed test set, the datasets used in our simulations are all derived from the Essen folksong collection (Schaffrath and Huron, 1995). The Essen folksong collection is a corpus consisting of monophonic transcriptions of folksongs, originating from various geographical regions across the globe. The majority of the folksongs in this dataset originate from regions in Germany and China. We use a version of the Essen folksong collection encoded in humdrum format, which we obtained from http://kernscores.stanford.edu.

Folksongs without an annotated time signature, or with multiple time signatures are filtered out. The simulations described below use different subsets of this filtered version of the Essen folksong collection.

3.3. Classification Performance and the Influence of Preceding Context

The first expectation formulated in Section 2.5 concerns the model's ability to infer meters that agree with time signatures notated in scores. To evaluate this, classification performance is measured using ten-fold cross validation on a dataset of German folksongs. In a cross validation scheme, a model is trained and evaluated ten times on different partitions of the dataset into a training set and a test set. Reported classification scores are based on the average classification score over all ten partitions.

The second expectation we formulated is that models exploiting sequential probabilistic properties will perform better in this task than a similar model that does not exploit such properties. To evaluate this, we measure classification performance of five different models configured with order-bounds ranging from zero to four using cross validation. The order-bound parameter (see Section 2.2) allows us to vary the degree to which the model can learn sequential probabilistic properties of rhythms, interpolating between a model that can only learn distributions of onsets over metrical positions (order-bound zero) and a model that predicts the subsequent metrical position based on the metrical positions of the last four events (order-bound four).

The result of performing inference on the generative model—inferring meter from a rhythm—is not a single classification, but a posterior probability distribution over metrical interpretations. To determine in which meter the model interprets a rhythm, an additional inferential step is required. All classification scores reported in this paper are based on the interpretations with the highest posterior probability after observing the entire rhythm. An interpretation is considered correct if its phase and category agree with the annotated time signature.

For these simulations, we used rhythms extracted from 4,966 German folksongs in the Essen folksong collection. This set is constructed by selecting all melodies with an "ARE" record (area of origin; Huron, 1999) indicating a region of Germany from the Essen folksong collection, subject to the constraints described in Section 3.2. **Figure 2** shows the distribution of meters in the resulting dataset. The most frequently appearing time signatures in this set are 4/4, 2/4, 3/4, and 6/8.

3.4. Does Metrical Inference Reduce Prediction Error?

The third expectation we formulated is that a model using inferred meter to predict the onsets of musical events will outperform comparable models that do not use metrical inference. To assess whether metrical inference increases predictive performance we compare the model an IDyOM model that predicts event onset time without inferring meter. Prediction performance is measured by looking at average information content (see Section 2.4), which represents the discrepancy between predicted and observed events.

This IDyOM model is configured to use a single viewpoint, encoding inter-onset intervals between subsequent events, to predict onset time. Inter-onset interval is defined as the difference

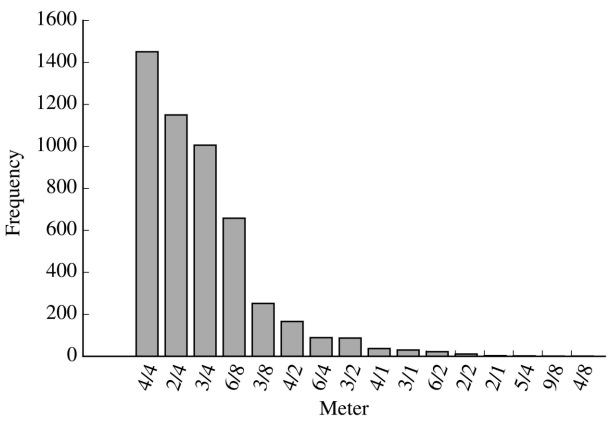


FIGURE 2 | Histogram showing the of the distribution of meters in the dataset of 4,966 German folksongs from the Essen folksong collection.

between the onset time of the final and penultimate event. Both models are trained and evaluated on the same dataset using cross-validation, and the input of both models consists only of onset times encoded in the event representation.

The results are reported, as before, for order-bounds varying from zero to four. The values represent average information content over cross validation folds.

3.5. Simulating Enculturation

The fourth expectation concerning the model's behavior we formulated is that it should show sensitivity to the statistical properties of its training data. To investigate this, two types of statistical aspects of training data that affect the model's behavior in different ways are distinguished. The first aspect is the distribution of metrical categories in the training rhythms. This distribution is directly reflected in the prior distribution, encoding a priori likelihood of different metrical categories. The effect of the prior distribution on the model's behavior can be seen as *inferential biases*. The second aspect concerns the sequential structure of the training rhythms themselves. This aspect includes the distribution of onsets over different metrical positions, but also the typical unfolding of rhythms interpreted in a specific meter and the presence of stereotypical rhythmic patterns.

These two aspects of training data may influence the encountered prediction error on novel rhythms as well as the metrical category in which rhythms are interpreted. To investigate the effect of inferential biases, we focus on consequences of inferential biases for metrical interpretation. In the investigation of the statistical properties of rhythms themselves we focus on the effects of training data on prediction error.

The simulations described below are all conducted using an order-bound of four, since the cross validation results indicate that, out of the considered order-bounds, four works best (see Section 4).

3.5.1. Inferential Biases

A high prevalence of certain metrical categories in the music to which a listener has been exposed to previously may lead to inferential biases: a tendency to interpret rhythms in the pervasive category. In probabilistic terms, this is a sensible behavior: in the presence of uncertainty, it is optimal to tend toward categories with a high *a priori* likelihood of occurring. Such likelihoods are represented in the prior distribution over metrical categories. Inferential biases are top-down in the sense that they are independent of the particular rhythm encountered by the model. Once the model begins to process a rhythm, the prior distribution is updated by bottom-up evidence from the rhythm. Inferential biases can alternatively be understood as changing the initial state of meter induction. Meters favored by the prior distribution require less evidence from rhythmic events to gain a high posterior likelihood. In cases where a rhythm is ambiguous (i.e., provides evidence for two or more metrical categories), inferential biases toward either category can be decisive in the model's interpretation.

To investigate the effect of inferential biases, we train two models on a subset of the German folksongs described in Section 3.3 containing 658 2/4 (a simple duple meter), 658 3/4 (a simple triple meter) and 658 6/8 (a compound duple meter) training examples. We bias the prior distribution of one model to favor 3/4 interpretations while the other model is biased to favor 6/8 interpretations.

In this simulation the prior distribution is not estimated empirically using the relative frequency of metrical categories in the training data. Instead, the parameters of the prior distribution are manually set to the values shown in **Table 1**. The rationale behind this choice is that if we would manipulate the prior distribution by altering the number of training rhythms in a metrical category, the number of training examples from which the model predictive model of that category is learned would be affected, which introduces performance differences that cannot be attributed solely to the prior distribution.

The consequences of the biased prior distribution are investigated using an artificially constructed test set. To construct this set, first, a set of rhythmic patterns is constructed by generating all possible patterns within the following constraints: the total duration of a pattern is exactly twelve sixteenth notes, none of the patterns begin with a rest and the minimum inter-onset interval is a sixteenth note. The resulting set consists of 2^{11} rhythmic patterns: each pattern begins with an onset and each sixteenth-note time point between the second and twelfth sixteenth-note can contain an event onset. Because twelve sixteenth notes is exactly the duration of one 3/4 or 6/8 bar, this set contains all rhythms with a minimum interval of a sixteenth note that fit in one bar of a 3/4 or 6/8 meter. To construct the final test set, each of these patterns is repeated four times. The repetition allows the model more time to converge on a single interpretation.

Both models are used to infer meter for each rhythm in the test set. Note that while three different categories, 2/4, 3/4, and 6/8, are considered, the quadruple repetition of patterns with a duration of twelve sixteenth notes may favor 3/4 and 6/8 interpretations. Since this potential bias is a property of the test set on which both models are evaluated, it does not cause problems for the evaluation of the effect of inferential biases.

We expect that inferential biases will increase the number of rhythms interpreted in the category corresponding to the bias. Due to the juxtaposition of 3/4 and 6/8 inferential biases, and the bar-level period-correspondence between these two meters, we expect to find the greatest degree of disagreement in interpretation of rhythms in the test set between the 3/4 and 6/8 categories: the 3/4 biased model will likely interpret rhythms classified by the 6/8 biased model as 6/8 in 3/4 and vice versa.

TABLE 1 | Prior probabilities of metrical categories used for simulating inferential biases.

Category	Prior probability	
	3/4 biased	6/8 biased
2/4	4/9	4/9
3/4	4/9	1/9
6/8	1/9	4/9

It seems plausible that 3/4 and 6/8 inferential biases will lead to some disagreement about the 2/4 category. An inferential bias may lead a model to interpret rhythms classified by the other model as 2/4 in the category corresponding to its bias. At the tactus level, 2/4 and 3/4 exhibit structural similarities: by convention, 2/4 and 3/4 both imply simple meters, where beats are subdivided into two smaller units. The 6/8 time signature, on the other hand, implies a compound meter. These (music-theoretic) similarities between 2/4 and 3/4 may lead the 3/4 biased model to interpret more rhythms, interpreted in 2/4 by the 6/8 biased model, according to its bias than the 6/8 biased model will out of the rhythms interpreted in 2/4 by the 3/4 biased model. It is worth noting that 2/4 and 6/8 have a different structural similarity at the level above the tactus: they are both duple meters. However, the duration of beat in 2/4 and 6/8, in our quantized input representation, is different, preventing this similarity from playing a role in our model.

The set of rhythms interpreted differently by both models likely consists of rhythms that do not strongly imply one specific interpretation. We expect such rhythms to be either ambiguous, or metrically over- or under-determined (London, 2012, pp. 75–76). Because we define a classification as the interpretation with the maximum posterior probability, the model always produces an interpretation of a rhythm, even if evidence from the rhythm is weak or conflicting. Therefore, some of the rhythms about which the models disagree may be metrically vague, i.e., not strongly suggesting any interpretation.

3.5.2. Cultural Distance between Chinese and German Rhythms

In two simulations, we investigate how the model responds to being trained on folksongs originating from China or Germany. Music from these two areas might be different enough to lead to differences in rhythmic processing between enculturated individuals. By training the model on a dataset of Chinese and German folksongs, we can simulate how, according to the model, exposure to these stylistically different sets of rhythms affects perception.

To this end, we use two dataset sets, containing folksongs originating respectively from Germany and China. The German dataset is the same one that is used for the cross validation simulations described in Section 3.3. The dataset of Chinese folksongs is constructed in the same way as the German dataset, namely by selecting all folksongs from the Essen folksong collection whose "ARE" reference record (Huron, 1999) indicated a region in China and after first filtering out folksongs with zero or more than one annotated time signatures.

We run simulations in two separate conditions. In both conditions, two models are trained: one on a Chinese training set, and one on a German training set. Both of these models are subsequently evaluated on a separate Chinese and German test set consisting of rhythms that do not occur in the training data. In contrast to the simulation described above, we estimate the prior distribution in its normal way (see Equations 3 and 4).

The number of rhythms of each metrical category used in the test and train sets in the first and second condition are shown in **Table 2**.

In the first condition (see the columns under "identical" in **Table 2**), we control for the effect of the prior distribution and use identical distributions of metrical categories in the training data of both models. This allows us to attribute observed effects to differences in the statistical properties of rhythms, ruling out effects of differences in the number of training examples or the differences in prior distributions. Meters considered in the simulation need to be well represented in both datasets. In the German and Chinese dataset that we have available, this constraint leaves 2/4, 3/4, and 3/8 as suitable categories. Despite this reduction, the number of rhythms in meters other than 2/4 in the Chinese dataset remains rather small.

Due to the small number of rhythms in meters other than 2/4 in the Chinese dataset, it is not possible to use a uniform distribution of meters in the test sets for this condition. Instead, we only include rhythms in 2/4 in the German and Chinese test set.

In the second condition (see the columns under "empirical" in **Table 2**), we allow the prior distribution to influence results and use empirical distributions of metrical categories in the training data of both models. By empirical, we mean that the relative frequencies of meters in the test and training sets that we used are equal to those observed in the Essen folksong collection. Both training sets contained in total an equal number of training examples.

Rhythms in the test sets for this condition are distributed to the same proportions as in the corresponding training sets. The Chinese test set predominantly contains rhythms annotated in 2/4 while the German test set also contains substantial numbers of rhythms in 3/4 and 4/4.

We expect that, on the Chinese and the German test sets, the model trained and tested on culturally similar music will exhibit lower average information content and higher classification performance than the model trained on culturally different music. We expect to see this pattern of results both for the identical, as well as for the empirical distribution of meters in the training data.

4. RESULTS

4.1. Classification Performance and Preceding Context

Figure 3A shows the average number of correct interpretations found by our model at order-bounds ranging from zero to four. The averages are obtained by first averaging all per-event information contents (see Section 2.4) in the test set of one cross validation fold, and subsequently over all cross-validation folds. The standard deviations are calculated over the averages per cross validation fold. At order-bound zero, the model interprets rhythms in agreement with annotated time signatures in, on average 38%, of the cases. At order-bound one, classification performance increases sharply to, on average, 67% of the rhythms in agreement with the annotated time signature. Increasing order-bound further yields modest improvements. At order-bound four, the highest we tested, on average, 71% the rhythms were interpreted in agreement with the annotated time signature.

TABLE 2 | Number of rhythms in different metrical categories in the training and test sets used in the simulation of enculturation.

Distribution of meters		Identical				Empirical			
Country of origin		Germany		China		Germany		China	
Dataset		Training	Test	Training	Test	Training	Test	Training	Test
Meter	2/4	950	200	950	200	339	60	1,009	178
	4/4	132	0	132	0	427	75	90	16
	3/4	35	0	35	0	296	52	24	4
	3/8	19	0	19	0	74	13	13	2
	All	1,136	200	1,136	200	1,136	200	1,136	200

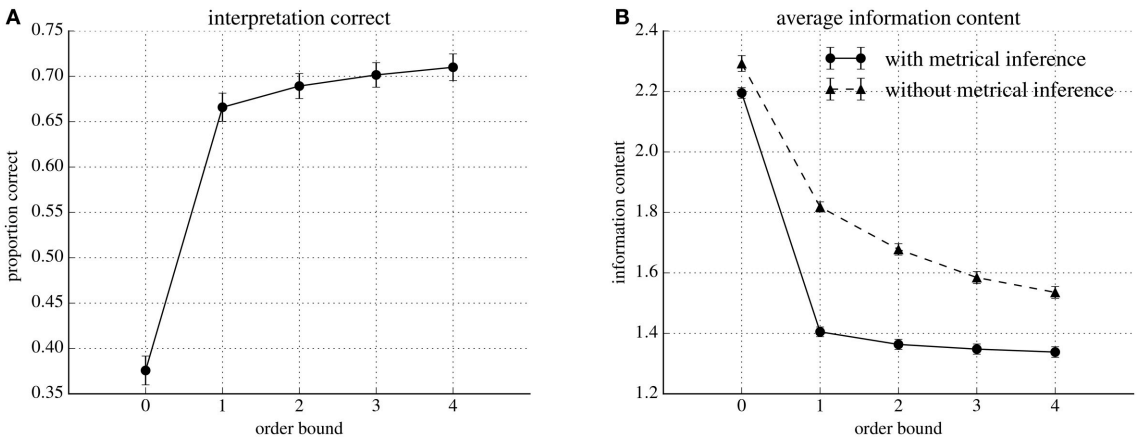


FIGURE 3 | Classification performance and average information content for five different models varying in order-bound, evaluated using ten-fold cross-validation. Markers represent values obtained by averaging over the ten folds. Error bars extend one standard deviation above and below the average values. **(A)** Proportions of correctly classified interpretations. **(B)** Average information contents for the model (with metrical inference) compared to IDyOM without metrical inference.

Variability in performance between different partitions of the data in a training and test set is low, as the small error bars in **Figure 3A** show.

4.2. Metrical Inference and Prediction Error

Figure 3B shows prediction performance in terms of average per-event information content of rhythms under IDyOM (without metrical inference) and our extended version of IDyOM (with metrical inference). Both models were tested at order-bounds ranging from zero to four.

The results shows that, in general, information content decreases as order-bound increases for both the IDyOM model (without metrical inference) and our model (with metrical inference). The results also show that for all tested order-bounds, the average information content is lower our model (with metrical inference): for example 2.19 compared to 2.29 for order-bound zero and 1.34 compared to 1.54 at order-bound four.

4.3. Simulating Enculturation

4.3.1. Inferential Biases

The results obtained from contrasting two models with manually manipulated prior distributions on an artificially generated test set are summarized in **Table 3**.

TABLE 3 | A contingency table showing the number of time signature classifications by a 3/4 biased model and a 6/8 biased model.

		3/4 biased			
		6/8	3/4	2/4	All
6/8 biased	6/8	471	83	40	594
	3/4	0	395	0	395
	2/4	0	54	1,005	1,059
	All	471	532	1,045	2,048

The results shows that both models interpret approximately half of all rhythms in 2/4. The rightmost column shows that the 6/8 biased model interprets more rhythms in 6/8 than in 3/4, while the bottom row shows that the 3/4 biased model interprets more rhythms in 3/4 than in 6/8.

The numbers on the diagonal show that both models agree on the vast majority of interpretations. Both models agree on the interpretation of rhythms that are classified as 3/4 or 6/8 *despite* inferential bias: None of the rhythms that the 3/4 biased model interprets as 6/8 are interpreted differently by the 6/8 biased model. Similarly, none of the rhythms that the 6/8 biased model interprets as 3/4 are classified differently by the 3/4 biased model.

The numbers off the diagonal show that the greatest degree of disagreement occurs between the 6/8 and 3/4 categories, but there is also substantial disagreement between 2/4 and 3/4 and 2/4 and 6/8.

There are two categories of rhythms sensitive to inferential biases: The first category consists of 83 rhythms that the 6/8 biased model interprets in 6/8 while the 3/4 biased model interprets them in 3/4. The second category consists of rhythms that one model interprets in 2/4 while the other model interprets them in the category its biased toward. The 6/8 biased model interprets 40 rhythms in 6/8 that the 3/4 biased model interprets in 2/4. Out of the rhythms classified by the 6/8 biased model as 2/4, the 3/4 biased model interprets slightly more rhythms in agreement with its bias (namely 54), than the 6/8 biased model does out of the rhythms classified by the 3/4 biased model as 2/4 (namely 40).

4.3.2. Cultural Distance between Chinese and German Rhythms

Table 4 shows average information content and classification performance obtained in the simulations of enculturation with German or Chinese folksongs. Results from two conditions are reported: one in which the German and Chinese training sets have an identical distribution of metrical categories and one in which they have empirical distributions of metrical categories.

In both conditions the results can be said to show effects of enculturation: The average information content for models evaluated on rhythms from the same country as the rhythms in their training data (culturally familiar) is lower than for models trained on rhythms from the other country (culturally unfamiliar). Classification performance shows a similar pattern: in most cases, classification performance is better for models evaluated on culturally familiar rhythms. However, in the identical prior condition, classification performance of the German model on the Chinese test set was slightly higher than of the Chinese model. Furthermore, in the identical prior condition, the average information content of the Chinese model is lower when evaluated on the German test set compared to the Chinese test set.

For both models and in both conditions, but most notably in the identical priors condition, information content of rhythms in the Chinese test set was slightly higher than that of rhythms in the German test set.

Figures 4A,B project the rhythms from both test tests onto a two-dimensional plane. The coordinates of each rhythm are determined by the average information content of events in the rhythm under the Chinese model (x-axis) and German model (y-axis). Under this projection, rhythms from the two cultures form clusters that are to some degree spatially separated. The degree of separation is stronger in the empirical prior condition (**Figure 4B**). For both conditions, average information content of events in a single test set is highly correlated between both models (see **Table 5**).

5. DISCUSSION

A predictive coding view of perception entails that perception depends on generative models in the mind of the perceiver that are tuned by statistical properties of the environment, through evolutionary adaptation and sensory experience, to predict

TABLE 4 | Average information content and classification performance of models trained and evaluated on test sets with rhythms from Germany and China.

	Training set	Test set			
		Identical priors		Empirical priors	
		German	Chinese	German	Chinese
Information content	German	1.21	1.63	1.34	1.72
	Chinese	1.32	1.49	1.70	1.49
Classification	German	0.84	0.80	0.73	0.72
	Chinese	0.59	0.77	0.47	0.75

Results are reported for two different conditions. One in which training sets contain identical distributions of metrical categories, and one in which training sets contain empirical distributions of metrical categories.

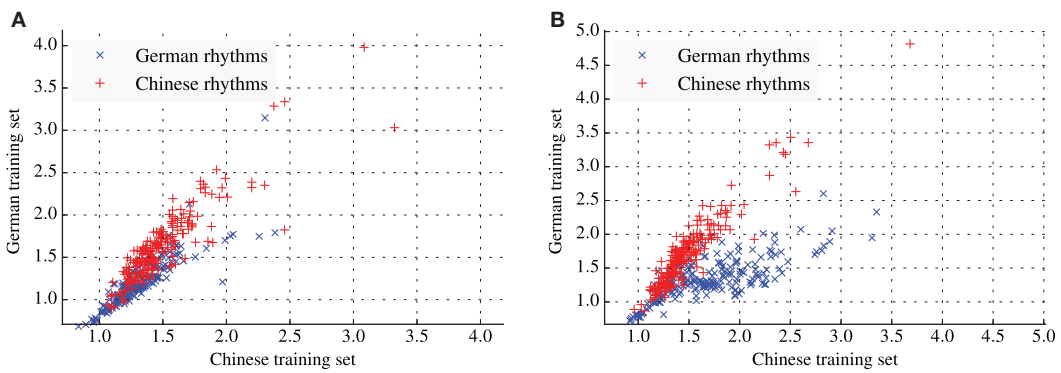


FIGURE 4 | Scatter plots of the average information content of rhythms for the Chinese and German models. (A) Results for the training and test sets with fixed distributions of meters. **(B)** Results for the training and test sets with empirical distributions of meters.

TABLE 5 | Pearson product-moment correlation coefficients between average information content per rhythm under the German and the Chinese model, showing the degree to which information-content assigned to the same rhythms by both models is related.

	German test set	Chinese test set
Fixed prior	0.74	0.94
Empirical prior	0.86	0.89

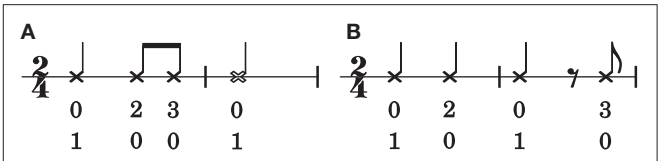


FIGURE 5 | Two rhythms that result in different orderings of the same set of mp ⊗ bd viewpoint elements. The number-pairs below the notes are the values of mp ⊗ bd. The top number represents the value of the bd (bar-distance) viewpoint, the bottom number represents the value of the mp (metrical position; expressed in multiples of an eighth note duration) viewpoint. The rhythm in **(A)** is structurally different from the rhythm in **(B)**, but under a zeroth-order mp ⊗ bd viewpoint, they are indistinguishable.

sensations. We hypothesized that effects of enculturation on the perception of meter can be understood in terms of predictive coding. To explore the consequences of this idea, we presented a probabilistic model of meter perception for which predictive coding served as the conceptual basis. The underlying hypothesis is that meter perception is the result of a strategy, based on statistical learning, probabilistic prediction and inference, for increasing predictive accuracy in processing of temporal events in music.

A set of expectations concerning the model's behavior was derived based on: the relevance of the model as a cognitive model of meter perception, theoretical proposals about the relation between rhythm and meter, the model's ability to reduce prediction error, and finally the model's potential to simulate enculturation. To investigate the degree to which the model meets these expectations, we ran a series of simulations. The results show that the model can infer metrical structure from rhythms, and that this ability improves when statistical properties of the succession of onsets in the metrical context are taken into account. A comparison with a similar model that does not use metrical inference demonstrates that metrical inference reduces prediction error in predicting the timing of musical events. Finally the results show hypothesized patterns of enculturation when models are trained on corpora varying, both naturally and artificially, in terms of distribution of meters and rhythmic properties.

The following sections discuss the simulation results in detail.

5.1. Meter Classification and Preceding Context

A model of meter perception can reasonably be expected to interpret a simple rhythm in a meter that agrees with the time signature that an educated listener would use when transcribing that rhythm. The used rhythms were taken from folksongs in the Essen folksong collection (Schaffrath and Huron, 1995). Despite its possible relevance to determining the time signature, melodic information was disregarded. This limitation notwithstanding, cross-validation results indicate that the model generally infers interpretations that agree both in category and phase with annotated time signatures. The best performing model configuration interprets rhythms in a time signature and phase that agrees with annotations in the Essen folksong collection in 71% of the cases. These classifications were selected by the model out of a large pool of alternatives. Summing the number of possible phases per considered metrical category (see Section 2) yields 320 possible metrical interpretations. Many of these categories occur very infrequently in the training data, resulting in a low *a priori* likelihood for these categories. If we limit interpretations to the four most frequently occurring metrical categories—4/4, 2/4, 3/4, and 6/8—the number of interpretation options reduces to 48.

By varying the model's order-bound (the amount of preceding events that inform the prediction of the next event, see Section 2.2), we investigated to what degree learning statistical properties of the succession of metrical positions in rhythms improved the model's performance.

Increasing the order-bound from zero to one yields the most significant improvement in classification performance. This finding is consistent with results obtained by Temperley (2010) in a comparison of six onset-prediction models. Some of these models were metrical, which means they made use of provided (rather than probabilistically inferred) metrical information. Temperley (2010) found that out of the compared models, the two metrical and context-sensitive models, namely the first-order metrical duration model and hierarchical position model, yielded the lowest cross-entropy (information content) score.

The performance increase between order bound zero and one is unsurprising. In a zeroth-order model, events in a rhythm are conditionally independent given a meter. If the meter is known, the probability of the next event only depends on its metrical status and is independent of preceding events[1]. In a zeroth-order model, a rhythm is a "bag of notes": the order in which notes occur is irrelevant to the final outcome. However, note-order bears consequences for the metrical interpretation of a rhythm, as illustrated in **Figure 5**. The rhythm in **Figure 5A** is structurally different from the rhythm in **Figure 5B**, yet under a zeroth-order model using mp ⊗ bd metrical viewpoints (see Section 2.3) these rhythms are indistinguishable.

The results show that classification and prediction performance, increases further when order-bound is increased to four. Since this improvement is relatively modest, it remains to be seen to what extent probabilistic information about the succession of multiple events facilitates metrical inference. Perhaps the effect of order-bound would be more pronounced

[1]The bd viewpoint used in our simulations indirectly introduces minor context dependency: if its value zero it means that the current note is the first note in the bar.

for music styles with more complex rhythms than the folksongs used here.

5.2. Metrical Inference Reduces Prediction Error

We proposed that meter perception may result from predictive coding: interpreting onsets in a rhythm as the result of a generative model with different periodic categories (meters), that are inferred from the pattern of onsets itself, may facilitate prediction of future onsets. Interpreting a rhythm in a metrical framework allows a listener to relate the observed events to patterns they observed previously. A computational probabilistic model that infers meter to predict the timing of events, such as the one presented here, should therefore encounter a lower prediction error in empirical rhythms compared to a similar model that does not infer meter.

To evaluate this, we compared prediction performance of the presented model to an IDyOM model that predicts the event onset times without using metrical inference. This comparison seems natural because the presented model implements metrical inference directly on top of IDyOM as explained in Section 2.

Simulations show that the meter inferring model reduces prediction error compared to IDyOM (without metrical inference) under all tested order-bounds. These results support the suggestion that inferring meter may improve temporal prediction of events in rhythms.

5.3. Simulating Enculturation

The goals of the simulations concerning enculturation were to investigate how our model's behavior is shaped by the statistical properties of rhythms in its training data, and to investigate the extent to which these statistical properties can be exploited to improve the prediction and metrical interpretation of stylistically similar rhythms. We first explored the consequences of inferential biases on an artificially constructed set of potentially ambiguous rhythms. Then, we studied the effect of statistical properties of sets of rhythms on metrical inference. The results show that when tested on Chinese rhythms, models trained on rhythms of Chinese folksongs show better prediction performance than models trained on German folksongs. The converse was true when the models were tested on German folksongs.

This simulation of enculturation should be seen as a proof-of-concept: Patterns of quantized onset times annotated with meter are a limited representation of the rich variety of musical and non-musical experiences that may shape listeners' perception of meter. In the musical domain, timbre, polyphony, expressive timing and dynamics are some examples of aspects not considered by our approach that all could plausibly form part of the experiences that shape meter perception. Nevertheless, it is possible that monophonic corpora of rhythms from different cultures can predict some enculturation effects. The methodology presented here is an illustration of how such predictions could be made.

5.3.1. Inferential Biases

Inferential biases were introduced into the model by directly manipulating the prior distribution, while avoiding differences in the amount of training examples per metrical category, which would influence the results.

We contrasted two models: one with a 6/8 inferential bias, another with a 3/4 inferential bias. The models were evaluated on an artificially constructed test set of rhythms with the potential for ambiguity between 3/4 and 6/8. These test rhythms were not annotated, as we intended find the set of rhythms for which inferential biases could swing the model's interpretation.

The results show that inferential biases affected the distribution of interpretations over metrical categories in ways that we expected: Each model interpreted more rhythms in the category corresponding to its bias than the other model. Both models agreed on the interpretation of the majority rhythms. These rhythms contained enough evidence toward a particular interpretation to override the model's inferential bias. As we expected on music theoretic grounds, the 3/4 biased model swung the interpretation of slightly more rhythms, interpreted in 2/4 by the 6/8 biased model, to a 3/4 interpretation than the 6/8 model did out of the set of rhythms interpreted in 2/4 by the 3/4 biased model.

Eight rhythms interpreted were interpreted in 3/4 without pick up by the 3/4 biased model and in 6/8 without pick up by the 6/8 biased model. These rhythms are shown, by way of example, in **Figure 6**, along with metrical grids contrasting a simple 3/4 interpretation with a compound 6/8 interpretation. That the interpretation of these rhythms could be influenced depending on inferential bias of the model suggests that they are ambiguous (e.g., **Figure 6vi**), and/or metrically underdetermined (e.g., **Figures 6i,iv**), or metrically vague, i.e., not strongly suggesting any interpretation (e.g., **Figure 6viii**).

5.3.2. Cultural Distance between Chinese and German Rhythms

In general agreement with the hypotheses presented in Section 2.5, the results in **Table 4** show that models evaluated on a test set with rhythms from the same country as the rhythms they were trained on exhibit a lower average per-event information content. The classification scores for models trained on culturally familiar rhythms were also higher compared to models trained on culturally unfamiliar rhythms, except on the Chinese test set in the identical prior scenario. It could be that rhythms in the Chinese portion of the Essen folksong collection (Schaffrath and Huron, 1995) were less consistently annotated, but further investigation is necessary to determine whether this is the case. The pattern of results suggests that the statistical properties of Chinese and German rhythms are different, and that these differences can be exploited to optimize prediction and metrical inference on rhythms from one of the countries.

In a recent study comparing recognition memory in North American listeners on Turkish classical music and Western art music, Demorest et al. (2016) found that rhythmic properties of music did not contribute to an enculturation effect on memory performance. At a first sight, these results seem surprising in the light of earlier studies that did find effects of enculturation

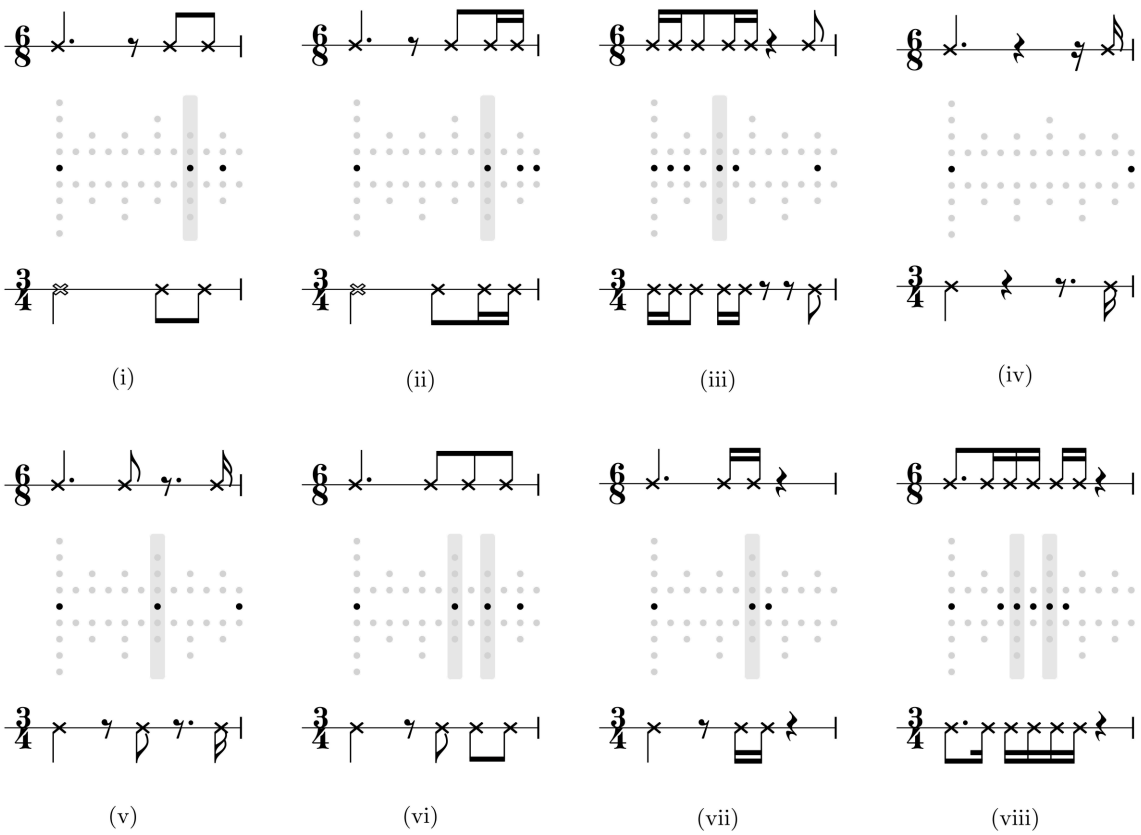


FIGURE 6 | The full set of rhythms interpreted as non-anacrustic 3/4 rhythm by a model with a 3/4 inferential bias and a non-anacrustic 6/8 rhythm by a model with a 6/8 inferential bias. Each rhythm is shown in dot notation between a compound 6/8 metrical grid (above) and a 3/4 simple metrical grid (below). The gray bars highlight onsets that fall on beats with different theoretical saliencies in both interpretations. Score transcriptions of the rhythms in 6/8 and 3/4 are shown above and below the grids.

related to rhythmic organization of music (Hannon and Trehub, 2005a,b; Hannon et al., 2012). However, it is possible that while rhythms are capable of eliciting effects of enculturation, such rhythms did not occur in the stimuli used by Demorest et al. (2016). Demorest et al. (2016) used a small set of stimuli that were not specifically selected to contain rhythms likely to elicit an effect of enculturation. The methodology applied in this simulation of enculturation is an example of how probabilistic models of rhythm perception can be employed to predict which rhythms are likely to elicit an effect of enculturation.

We hypothesized that fine-tuning of perception to the statistical properties of musical rhythms in one's environment in a way that leads to a reduction of prediction-error in rhythms typical of one's environment leads to differences in the processing of meter. This idea is closely related to the notion of *cultural distance*—the degree to which pitch relations in a musical excerpt resemble the pitch relations typical to music from one's own culture—introduced recently by Demorest and Morrison (2016). The *cultural distance hypothesis* (Demorest and Morrison, 2016) states that cultural distance is predictive of various culturally dependent responses such as preference, tension, expectation, and memory. This hypothesis is supported by a series of studies where cultural distance of stimulus material was found to affect memory performance (for an extensive overview, see Morrison and Demorest, 2009). Demorest and Morrison (2016) propose that cultural distance could be measured using probabilistic models of melodic expectancy, such as IDyOM, that learn the statistical properties of music from a particular culture. Music that is culturally distant from the music such a model is trained on should be predicted less effectively than culturally familiar music. As such, in the context of a cross-cultural study, average information content—the degree to which observed events deviate from one's expectations—can be seen as an operational definition of cultural distance.

The model presented here can supplement predictions about *melodic* cultural distance as provided by existing probabilistic models, with predictions about *rhythmic* cultural distance. Cultural distance, as predicted by our probabilistic model, can then be read directly from **Figures 4A,B**. If the probabilistic aspects of rhythm learned by the presented model correspond to those implicitly learned by human listeners, then, according to the cultural distance hypothesis, rhythms in the top-right part of **Figures 4A,B** should be more difficult to remember for German listeners while rhythms in the bottom-right part of **Figures 4A,B** should be more difficult to remember for Chinese listeners.

Other culturally dependent responses mentioned by Demorest and Morrison (2016) such as, expectation, preference, and tension can be potentially linked to information content as well. Regarding expectation, information content is a direct consequence of predictive failure and has been shown to account well for human pitch expectations (Pearce, 2005; Hansen and Pearce, 2014). Regarding preference, perceived groove and experienced pleasure have been hypothesized to depend on the right balance between predictability and unpredictability (Witek et al., 2014). Furthermore, influential proposals have postulated close ties between expectation and both emotional responses to music (Huron, 2006) and musical meaning (Meyer, 1957). Regarding tension, melodic expectation has recently been linked to expressive performance, which in turn was linked to perceived tension (Gingras et al., 2015).

5.4. General Discussion

While it is commonly assumed that the metrical accent of a beat, as derived from formal hierarchical descriptions of meter (Lerdahl and Jackendoff, 1983), is proportional to the probability of onsets at those beats, recent findings by Holzapfel (2015) and London et al. (2016) challenge this view. London et al. (2016) suggested that onset frequency need not be correlated with metrical accent for effective communication of meter. Instead, they argue, it is the recurrence and stability of rhythmic figures in the context of specific meters that may play a key role in the relation between rhythm and meter.

The results we presented show that models which take into account the preceding context of musical events, thus possessing the potential to learn the typical unfolding of multiple characteristic rhythmic patterns under different meters, are generally better at predicting rhythms and reconstructing annotated meters from note onsets alone. These findings, we would argue, provide further support for the idea the relationship between rhythm and meter is not only characterized by the distribution of note onsets, but also by characteristic rhythms and statistical properties of succession of interval between events.

The model we presented learns a generative model of rhythms from an annotated corpus. The supervised aspect of this approach challenges the cognitive plausibility of our model. Humans develop a feel for meter in their own culture without someone explicitly informing them about the "right" metrical interpretation. Nevertheless, situated exposure to rhythm almost always happens within a context containing an abundance of multi-sensory information related to the rhythmic practice. Within the music itself, other instruments, expressive timing and dynamics may provide strong metrical cues. In the environment, being rocked to music as an infant, participating in dancing or observing other people dance all contribute to the multi-sensory context by which rhythm perception is shaped. While not entirely putting concerns related to the supervised aspect of our approach to rest, metrical annotations in our training data can potentially be seen as capturing some of the information communicated in situated exposure to rhythms.

We have only considered event onset times in the present study, but other musical aspects such as melodic repetition are known influence the perception of meter as well (Hannon et al., 2004). A full account of meter perception should take these aspects into account. Our model could be a good starting point for such an account: due to the implementation of the model in IDyOM, it is possible to link metrical viewpoints with melodic viewpoints and incorporate melodic aspects into the generative model.

Another limitation of the current model is its relatively simple representation of metrical structure. Time signatures fall short in capturing the structural complexity of perceived meter. The model treats metrical categories as independent generative models and structural similarities between meters remain unexploited. The model is limited in its interpretation of rhythms into metrical categories by the categories observed in training data. In future work, we will seek to address these limitations by extending the model's representation of metrical structure.

The model introduced here represents an extension of previous work in probabilistic modeling of music (Conklin and Witten, 1995; Pearce, 2005). It is worth pointing out that the predictive mechanisms on which the model presented here is based, are domain independent (Pearce and Wiggins, 2004). The PPM* sequence prediction methods we employ can be applied to any domain that can be represented as structured sequences of symbols. Indeed, they were originally proposed in the field of text compression, but have proven to be useful in cognitive models of melodic expectation as well (Pearce, 2005; Pearce and Wiggins, 2012).

In summary, we have presented a computational probabilistic model meter perception, grounded in a predictive coding perspective of perception. The model has the potential to simulate musical expectations resulting from the perception of meter, shaped by previous exposure. The results show that the model can interpret simple rhythms in meters that agree with annotated time signatures and that it generates the hypothesized effects of enculturation. Simulations such as the ones presented here, can be used to generate theoretical predictions for cross-cultural studies of rhythm perception. Future research will determine the extent to which the learning processes implemented by our model capture aspects of those at work in human listeners.

AUTHOR CONTRIBUTIONS

BW designed and implemented the model. The model builds on IDyOM, which was designed and implemented by MP; MP and HH improved the model; HH, MP, and BW designed the simulations; BW performed the research and analyzed the data; BW wrote the paper; HH and MP improved the paper.

FUNDING

HH is supported by a Distinguished Lorentz fellowship granted by the Lorentz Center for the Sciences and the Netherlands Institute for Advanced Study in the Humanities and Social Sciences (NIAS) and a Horizon grant of the Netherlands Organization for Scientific Research (NWO). BW and MP also

received support from the EPSRC Digital Music Platform Grant held at Queen Mary (EP/K009559/1). MP is supported by a grant from the UK Engineering and Physical Science Research Council (EPSRC, EP/M000702/1).

ACKNOWLEDGMENTS

We thank the reviewers for their helpful comments which significantly improved the presentation of this work.

REFERENCES

Barlow, H. B. (1961). "Possible principles underlying the transformations of sensory messages," in *Sensory Communication*, ed W. A. Rosenblith (Cambridge, MA: MIT Press), 217–234.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–253. doi: 10.1017/S0140525X12000477

Cleary, J. G., and Witten, I. H. (1984). Data compression using adaptive coding and partial string matching. *IEEE Trans. Commun.* 32, 396–402. doi: 10.1109/TCOM.1984.1096090

Conklin, D., and Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *J. New Music Res.* 24, 51–73. doi: 10.1080/09298219508570672

Creel, S. C. (2011). Specific previous experience affects perception of harmony and meter. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 1512–1526. doi: 10.1037/a0023507

Creel, S. C. (2012). Similarity-based restoration of metrical information: different listening experiences result in different perceptual inferences. *Cogn. Psychol.* 65, 321–351. doi: 10.1016/j.cogpsych.2012.04.004

Demorest, S. M., and Morrison, S. J. (2016). "Quantifying culture: the cultural distance hypothesis of melodic expectancy," in *The Oxford Handbook of Cultural Neuroscience 1st Edn.*, Chapter 12, eds J. Y. Chiao, S.-C. Li, R. Seligman, and R. Turner (Oxford: Oxford University Press).

Demorest, S. M., Morrison, S. J., Nguyen, V. Q., and Bodnar, E. N. (2016). The influence of contextual cues on cultural bias in music memory. *Music Percept.* 33, 590–600. doi: 10.1525/mp.2016.33.5.590

Egermann, H., Pearce, M. T., Wiggins, G. A., and McAdams, S. (2013). Probabilistic models of expectation violation predict psychophysiological emotional responses to live concert music. *Cogn. Affect. Behav. Neurosci.* 13, 533–553. doi: 10.3758/s13415-013-0161-y

Friston, K. J. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 360, 815–836. doi: 10.1098/rstb.2005.1622

Gingras, B., Pearce, M. T., Goodchild, M., Dean, R. T., Wiggins, G. A., and McAdams, S. (2015). Linking melodic expectation to expressive performance timing and perceived musical tension. *J. Exp. Psychol. Hum. Percept. Perform.* 42, 594–609. doi: 10.1037/xhp0000141

Hannon, E. E., Snyder, J. S., Eerola, T., and Krumhansl, C. L. (2004). The role of melodic and temporal cues in perceiving musical meter. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 956–974. doi: 10.1037/0096-1523.30.5.956

Hannon, E. E., Soley, G., and Ullal, S. (2012). Familiarity overrides complexity in rhythm perception: a cross-cultural comparison of American and Turkish listeners. *J. Exp. Psychol. Hum. Percept. Perform.* 38, 543–548. doi: 10.1037/a0027225

Hannon, E. E., and Trehub, S. E. (2005a). Metrical categories in infancy and adulthood. *Psychol. Sci.* 16, 48–55. doi: 10.1111/j.0956-7976.2005.00779.x

Hannon, E. E., and Trehub, S. E. (2005b). Tuning in to musical rhythms: infants learn more readily than adults. *Proc. Natl. Acad. Sci. U.S.A.* 102, 12639–12643. doi: 10.1073/pnas.0504254102

Hansen, N. C., and Pearce, M. T. (2014). Predictive uncertainty in auditory sequence processing. *Front. Psychol.* 5:1052. doi: 10.3389/fpsyg.2014.01052

Holzapfel, A. (2015). Relation between surface rhythm and rhythmic modes in Turkish Makam music. *J. New Music Res.* 44, 25–38. doi: 10.1080/09298215.2014.939661

Honing, H. (2013). "Structure and interpretation of rhythm in music," in *The Psychology of Music, 3rd Edn.*, ed D. Deutsch (London: Academic Press), 369–404.

Huron, D. B. (1999). *Music Research Using Humdrum: A User Guide*. Stanford, CA: Center for Computer Assisted Research in the Humanities.

Huron, D. B. (2006). *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, MA: The MIT Press.

Lerdahl, F., and Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.

Locke, D. (1982). Principles of offbeat timing and cross-rhythm in southern Eve dance drumming. *Ethnomusicology* 26, 217–246. doi: 10.2307/851524

London, J. (2012). *Hearing in Time, 2nd Edn*. New York, NY: Oxford University Press.

London, J., Polak, R., and Jacoby, N. (2016). Rhythm histograms and musical meter: a corpus study of Malian percussion music. *Psychon. Bull. Rev.* 24, 474–480. doi: 10.3758/s13423-016-1093-7

Longuet-Higgins, H. C., and Lee, C. S. (1982). The perception of musical rhythms. *Perception* 11, 115–128. doi: 10.1068/p110115

Longuet-Higgins, H. C., and Steedman, M. J. (1971). On interpreting bach. *Mach. Intell.* 6, 221–241.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge, MA: Cambridge University Press.

Manning, C. D., and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

Meyer, L. B. (1957). Meaning in music and information theory. *J. Aesthet. Art Crit.* 15, 412–424. doi: 10.2307/427154

Moffat, A. (1990). Implementing the PPM data compression scheme. *IEEE Trans. Commun.* 38, 1917–1921. doi: 10.1109/26.61469

Morrison, S. J., and Demorest, S. M. (2009). "Cultural constraints on music perception and cognition," in *Cultural Neuroscience: Cultural Influences on Brain Function, vol. 178 of* Progress in Brain Research, ed J. Y. Chiao (Amsterdam: Elsevier), 67–77.

Olshausen, B. A., and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. doi: 10.1038/381607a0

Omigie, D., Pearce, M. T., and Stewart, L. (2012). Tracking of pitch probabilities in congenital amusia. *Neuropsychologia* 50, 1483–1493. doi: 10.1016/j.neuropsychologia.2012.02.034

Omigie, D., Pearce, M. T., Williamson, V. J., and Stewart, L. (2013). Electrophysiological correlates of melodic processing in congenital amusia. *Neuropsychologia* 51, 1749–1762. doi: 10.1016/j.neuropsychologia.2013.05.010

Palmer, C., and Krumhansl, C. L. (1990). Mental representations for musical meter. *J. Exp. Psychol. Hum. Percept. Perform.* 16, 728–741. doi: 10.1037/0096-1523.16.4.728

Pearce, M. T. (2005). *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. Ph.D. thesis, City University, London.

Pearce, M. T., Conklin, D., and Wiggins, G. A. (2005). "Methods for combining statistical models of music," in *Computer Music Modeling and Retrieval: Second International Symposium, CMMR 2004* (Esbjerg, Denmark), Revised Papers (Berlin; Heidelberg: Springer), 295–312. doi: 10.1007/978-3-540-31807-1_22

Pearce, M. T., Müllensiefen, D., and Wiggins, G. A. (2010). The role of expectation and probabilistic learning in auditory boundary perception: a model comparison. *Perception* 39, 1367–1392. doi: 10.1068/p6507

Pearce, M. T., and Wiggins, G. A. (2004). Improved methods for statistical modelling of monophonic music. *J. New Music Res.* 33, 367–385. doi: 10.1080/0929821052000343840

Pearce, M. T., and Wiggins, G. A. (2012). Auditory expectation: the information dynamics of music perception and cognition. *Top. Cogn. Sci.* 4, 625–652. doi: 10.1111/j.1756-8765.2012.01214.x

Rao, R. P. N., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580

Savage, P. E., Brown, S., Sakai, E., and Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proc. Natl. Acad. Sci. U.S.A.* 112, 8987–8992. doi: 10.1073/pnas.1414495112

Schaffrath, H., and Huron, D. (1995). *The Essen Folksong Collection in the Humdrum Kern Format*. Menlo Park, CA: Center for Computer Assisted Research in the Humanities.

Smith, E. C., and Lewicki, M. S. (2006). Efficient auditory coding. *Nature* 439, 978–982. doi: 10.1038/nature04485

Stobart, H., and Cross, I. (2000). The Andean anacrusis? Rhythmic structure and perception in Easter songs of Northern Potosí, Bolivia. *Br. J. Ethnomusicol.* 9, 63–92. doi: 10.1080/09681220008567301

Temperley, D. (2004). An evaluation system for metrical models. *Comput. Music J.* 28, 28–44. doi: 10.1162/0148926041790621

Temperley, D. (2007). *Music and Probability*. Cambridge, MA: MIT Press.

Temperley, D. (2009). A unified probabilistic model for polyphonic music analysis. *J. New Music Res.* 38, 3–18. doi: 10.1080/09298210902928495

Temperley, D. (2010). Modeling common-practice rhythm. *Music Percept.* 27, 355–376. doi: 10.1525/mp.2010.27.5.355

Vuust, P., and Witek, M. A. G. (2014). Rhythmic complexity and predictive coding: a novel approach to modeling rhythm and meter perception in music. *Front. Psychol.* 5:1111. doi: 10.3389/fpsyg.2014.01111

Witek, M. A. G., Clarke, E. F., Wallentin, M., Kringelbach, M. L., and Vuust, P. (2014). Syncopation, body-movement and pleasure in Groove music. *PLoS ONE* 9:e94446. doi: 10.1371/journal.pone.0094446

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 04 May 2017
doi: 10.3389/fpsyg.2017.00662

Perception of Leitmotives in Richard Wagner's Der Ring des Nibelungen

David J. Baker[1]* *and Daniel Müllensiefen*[2]

[1] *Music Cognition and Computation Lab, School of Music and Dramatic Arts, Louisiana State University, Baton Rouge, LA, USA,* [2] *Music, Mind and Brain Lab, Department of Psychology, Goldsmiths, University of London, London, UK*

The music of Richard Wagner tends to generate very diverse judgments indicative of the complex relationship between listeners and the sophisticated musical structures in Wagner's music. This paper presents findings from two listening experiments using the music from Wagner's *Der Ring des Nibelungen* that explores musical as well as individual listener parameters to better understand how listeners are able to hear leitmotives, a compositional device closely associated with Wagner's music. Results confirm findings from a previous experiment showing that specific expertise with Wagner's music can account for a greater portion of the variance in an individual's ability to recognize and remember musical material compared to measures of generic musical training. Results also explore how acoustical distance of the leitmotives affects memory recognition using a chroma similarity measure. In addition, we show how characteristics of the compositional structure of the leitmotives contributes to their salience and memorability. A final model is then presented that accounts for the aforementioned individual differences factors, as well as parameters of musical surface and structure. Our results suggest that that future work in music perception may consider both individual differences variables beyond musical training, as well as symbolic features and audio commonly used in music information retrieval in order to build robust models of musical perception and cognition.

Keywords: musical memory, leitmotives, opera, symbolic notation, computational Modeling

OPEN ACCESS

Edited by:
Geraint A. Wiggins,
Queen Mary University of London, UK

Reviewed by:
Greg Poarch,
Westfälische Wilhelms-Universität Münster, Germany
Andrei Radu Teodorescu,
Tel Aviv University, Israel

***Correspondence:**
David J. Baker
dbake29@lsu.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *24 October 2016*
Accepted: *12 April 2017*
Published: *04 May 2017*

Citation:
Baker DJ and Müllensiefen D (2017) Perception of Leitmotives in Richard Wagner's Der Ring des Nibelungen. Front. Psychol. 8:662. doi: 10.3389/fpsyg.2017.00662

1. INTRODUCTION

While Richard Wagner and his music have been the topic of a wide range of musicological and music theoretic research (Bailey, 1977; Deathridge and Dahlhaus, 1984; Dreyfus, 2012), the compositional techniques Wagner developed and their effect on listeners has not received nearly as much attention from the music psychology community. This may be due to the fact that Wagner's music does not make use of tonality in the traditional sense, but rather has been aptly described by David Huron as "contracadential" and very harmonically sophisticated (Huron, 2006). Huron notes that the complexity in Wagner's music may be attributed to its cadential content in that his cadences are "almost entirely divorced from perceptual or formal segmentation" (Huron, 2006, p. 338) making his music difficult to process for listeners who do not have prior listening experience.

In addition to the difficulty delineating cadential structures in his music, Wagner also composed his melodic material in order to avoid the regularity that is found in other 19th century composers (Dahlhaus, 1980; Grey, 2007). This conscious choice to write melodic material that seems to be endless and avoids easy segmentation often leads to difficulties for listeners, which results in thwarted and delayed expectations of musical events. Despite these inherent difficulties in parsing his cadential and melodic material, the continued popularity of his music for people at various

points in history (Magee, 1988) seems to suggest that listeners from a wide range of backgrounds are able to process and enjoy the complex auditory scenes in his music.

Initial work investigating how listeners are able to hear salient musical material in Wagner's music was carried out by Deliège (1992) in order to demonstrate the principles of musical cue abstraction (Deliège and Mélen, 1997). Cue abstraction is rooted in Gestalt schematization processes inspired by the work of Lerdahl and Jackendoff (1983) and uses grouping and similarity-difference principles in order to predict where listeners will perceive musical boundaries as well as salient musical events. Deliège's studies on the perception of Wagner's music focused primarily on leitmotives, which are short musical ideas that can be used to refer to people, places, or ideas related to the musical narrative (Hacohen and Wagner, 1997).

Leitmotives are ideal cues for studying salient musical events because they can exist in a multitude of permutations that are all perceived as the same cognitive entity. For example, the *Schwert-Motiv*, while often played in the major mode on the trumpet, can also be orchestrated with varying mode, range, and timbre in order to successfully convey the correct musical emotion the composer intended. Despite these changes, the leitmotive is often recognized as the same categorical entity as demonstrated in **Figures 1**, **2**.

Using leitmotives as cues, Deliège demonstrated her cue abstraction principles, which model real-time music listening, were able to accurately predict salient musical events in non-tonal music (Deliège, 1992). Her initial findings showed higher leitmotive recognition rates in participants with more musical training, indicating that listener background played a significant role in the identification of salient musical events. Deliège has also demonstrated the success of the cue abstraction mechanism with the music of Bach (Deliège, 1996), Berio (Deliège and El Ahnmadi, 1990), and Boulez (Deliège, 1989).

Morimoto, Kamekawa, and Marui extended research on leitmotives by investigating the effect of extra-musical verbal information on the memorization and recognition of leitmotives. They found that exposing listeners to different types of verbal information in relation to musical material and the narrative did not result in any significant differences in the ability to recognize and memorize leitmotives (Morimoto et al., 2009). In a similar way, Albrecht and Frieler (2014) investigated how additional visual information (i.e., the events on the opera stage) might contribute to an individual's leitmotive recognition rate. They found that seeing and hearing the opera actually decreased an individual's ability to identify leitmotives in the auditory signal and hence suggests that visual information can act as a distractor in terms of encoding leitmotives.

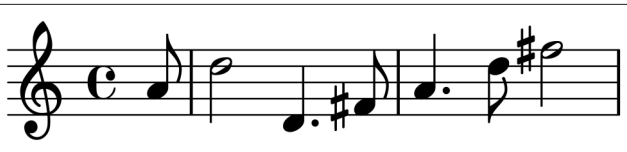


FIGURE 1 | The *Schwert-Motiv* in D major.

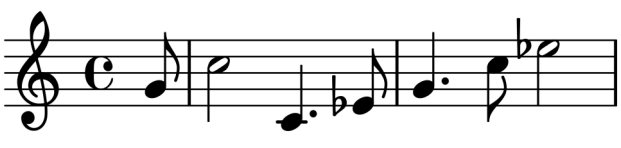


FIGURE 2 | The *Schwert-Motiv* in C minor.

Similar to much of existing work in music psychology, these previously mentioned studies investigating the perception of leitmotives categorized listeners based on their previous musical training. While musical training has been shown to be a factor that can contribute to performance in both tasks of perception (Besson et al., 2007; Williamson et al., 2010) and discrimination (Vuust et al., 2005) when investigating individual differences on musicality, much of this research unsystematically classifies participants into binary categories (such as "musicians" and "non-musicians"), primarily considering their years of formal musical training on an instrument as an indicator of their musical skills and experience. This somewhat arbitrary divide fails to consider other types of musical engagement or abilities other than instrumental skills (e.g., different types of perceptual abilities) which can also be deemed central to an individual's musicality (Levitin, 2012) or musical sophistication (Müllensiefen et al., 2014).

There is a lot of evidence from recent empirical research showing that scaled (i.e., continuous or ordinal as opposed to categorical or binary) measures of musical skills and experience represent good predictors in models of music perception and cognition (Chin and Rickard, 2012; Schaal et al., 2015; Bouwer et al., 2016), especially when considering musical background in the general population. While the aforementioned studies tend to reflect differences measuring individual's musical training, other studies have suggested that factors outside of musical training such as familiarity with the genre or style of the musical material (Tervaniemi, 2009; Hansen et al., 2016) as well as other non-performative abilities can play a crucial role in perceptual models (Bigand and Poulin-Charronnat, 2006). Though literature is sparse regarding perceptual models that takes into account genre familiarity, there are a number of studies that aim at mid-level features, such as schematic expectations (Eerola et al., 2009), and that do take into account listener backgrounds and musical acculturation that can be integrated in the modeling process via mechanisms such as statistical learning (Krumhansl et al., 2000).

We hypothesized that it might be possible to measure a listener's previous exposure to the music of Richard Wagner and use that measure as a predictor for their ability to recognize and remember cues in Wager's music. A previous study by Müllensiefen et al. (2016) found evidence that an individual's knowledge of and affinity for Wagner's music predicts memory accuracy for leitmotives in an experimental setting. In this particular experiment, expertise for Wagner's music was a stronger predictor than the amount of musical training for the participants' performance in the melodic memory experiment. These results suggested that an individual's prior exposure and understanding of a genre, and in particular Wagner's music,

does in fact play a significant role in the understanding of complex musical passages and the extraction of and memory for leitmotives.

In addition to individual differences between listeners in terms of general musical expertise and familiarity with Wagner's music in particular, features of the musical material itself are certainly also responsible for the degree to which the cognitive decoding of Wagner's music can be successful. Numerous studies from 1970s onwards have demonstrated how structural features of music can facilitate or hinder cognitive processing (Dowling, 1971, 1972; Dowling and Fujitani, 1971; Cuddy et al., 1979). However, much of this work made use of artificially constructed musical stimuli with the primary aim to control the features of the musical material used in the experimental setup , but sometimes at the expense of the ecological validity and generalizability to real music.

More recent work from music informatics and systematic musicology has suggested computational measures that produce feature descriptions of real music excerpts in symbolic encoding that can be used successfully in models of music perception and cognition (Pearce and Wiggins, 2006; Müllensiefen and Halpern, 2014; Collins et al., 2015; Vempala and Russo, 2015; Wiggins and Forth, 2015) Hence, one aim of this study is to employ computational measures of musical structure with leitmotives from Wagner's music and assess to what degree they are predictive of cognitive behavior. Complimentary to structural features of leitmotives via symbolic encoding, we also aim to assess how the similarity in sound between individual leitmotive excerpts and their occurrence in a musical context contributes to perceptual and cognitive decoding. There is a growing body of research demonstrating the usefulness of sound and audio features developed within the music information retrieval (MIR) framework for describing the development of general preferences and taste over time (Serra et al., 2013; Mauch et al., 2015), cognitive attributes like the catchiness of pop songs (Burgoyne et al., 2013; Van Balen, 2016), or perceived emotional content (Friberg et al., 2014). Specifically, in this study we assess similarity by comparing chromagram data derived from audio excerpts (Müller and Wapnewski, 1992; Mauch et al., 2015).

In summary, this study intends to assess how features of the compositional structure and audio similarity on one hand, as well as individual musical sophistication and expertise with regard to Wagner's music on the other affect recognition memory for leitmotives. Thus, the study aims to combine predictors reflecting features of the musical material and traits of the listener in a single model of perception and memory for leitmotives in Wagner's music. Specifically, we hypothesize that knowledge and affinity for Wagner's music music can be interpreted as proxies for familiarity with his leitmotive technique and should therefore have positive effects on leitmotive processing and memory. In addition, general musical training should also aid processing and memory on the experimental task, consistent with findings from previous experiments (Dowling, 1978, 1986; Harrison et al., 2016). The ability to speak German may also provide a processing advantage in this experiment since the German vocals might provide extra clues toward the decoding of leitmotives and musical events in the auditory scene. Wagner's leitmotives are often paired with certain terms or ideas from the libretto that we believe could enable participants who speak German to encode the musical structure of the leitmotives together with semantic connotations. This ability to bind multiple features and aspects of an a object at the encoding stage could then support retrieval processes in the recognition task. This assumption is in line with evidence from experimental studies that have shown a similar differential memory advantage of presenting music and text together (Serafine et al., 1984, 1986). Accounting for German speaking abilities was also incorporated into the design in order to account for any German text that could have been recognized in the exposure phase since recordings of live opera were used.

In terms of features of structural complexity we expect more complex leitmotives to be processed and remembered worse (Harrison et al., 2016). Finally, we hypothesize that the similarity in terms of sound (i.e., audio features) between an individual leitmotive and any similar sound but not identical parts in a longer passage can act as distractor and hence decrease memory performance.

We employ a cross-over experimental design that makes use of two scenes from Wagner's *Ring des Nibelungen*. The design allows us to use leitmotives that were the lures in the memory test of Experiment I to serve as correct responses in Experiment II and vice versa. Thus, the findings from both experiments can potentially replicate each other and therefore the design helps to disentangle incidental features of the leitmotives used as experimental stimuli from the parameters of interest (i.e., compositional structure and audio similarity) that should have the same effect in both experiments.

2. METHODS

2.1. Overall Design

This study consisted of two experiments that used the identical experimental design and procedures: In both experiments an approximately 10 min scene was played to participants followed by a surprise memory test for 20 leitmotives, some of which were present in the scene previously played (*old* items) and others that had not been present in the scene (*new* items). The two experiments were set up to replicate the findings from each other and thus reduce the effects of incidental features and hence ensure a greater robustness of the overall findings. The 10 *new* items in Experiment I were used as *old* items in Experiment II and the 7 *old* items in Experiment I were *new* items in Experiment II. The passages used were picked for their overlap in musical material, but due to using ecologically valid stimuli an even split on leitmotives was not possible.

2.2. Overall Procedure

Participants were tested in small groups. Upon arriving at the experiment participants signed a consent form and received the experimental instructions, which instructed them to listen attentively to a 10 min passage from a live recording of *Der Ring des Nibelungen* and subsequently answer some questions about the music. Participants listened to music via a pair of stereo speakers sitting at distances from 1 to 4 m from the speakers and via an initial sound check it was confirmed that

the volume of the audio was set to a comfortable level for all participants. After the exposure phase participants were handed a response sheet and started the test phase. Here, participants were played 20 short leitmotives for each of which they had to indicate the perceived pleasantness of the leitmotive on a 7-point scale, a binary indication (yes/no), whether this particular leitmotive occurred in the 10 min passage from the exposure phase, and a corresponding confidence rating on a 7-point scale. In addition, they also rated valence and arousal expressed by the leitmotive based on the Russell's circumplex circle of emotion (Russell, 1980). Questions were set up on their response sheet in the the order listed above and participants were asked to fill out the questions in order. Shorter leitmotives were repeated up to 3 times with a 3 s pause between repetitions, such that each leitmotive item in the test phase lasted approximately 20 s and was followed by a silent gap of 10 s before the next leitmotive was played. In total participants had approximately 30 s to make all five ratings (pleasantness, explicit memory, confidence, valence, arousal) and were told to complete their ratings before the next leitmotive was played. The order of leitmotives was randomized across two different lists to mitigate any order effects. Following the test phase participants completed a set of questionnaires assessing their musical background and Wagner affinity and expertise. Ethical approval was obtained from the Goldsmiths Psychology Department's Ethics Board.

2.3. Overall Materials

2.3.1. Self-report Measures

The self-report measures filled out at the end of each experimental session required participants to rate the familiarity with the passage in the exposure phase on a 7-point scale, their German speaking and writing abilities on 7-point scales, the amount of musical training they had received via the Training sub-scale of the Gold-MSI (Müllensiefen et al., 2014), as well as 14 questions assessing their affinity to Richard Wagner's music, each using a 5-point scale. In addition they also completed a 14-item objective multiple choice test assessing objective knowledge of *Der Ring des Nibelungen* and various facts relating to the life of Richard Wagner. Data from the Wagner affinity questionnaire was analyzed using factor analysis and each participant was assigned a corresponding factor score as described in Müllensiefen et al. (2016). Data from Wagner knowledge multiple choice test was scored using an item response model that generated an ability estimate for each participant (Müllensiefen et al., 2016).

2.3.2. Measures of Musical Structure and Sound

In order to assess each leitmotive's structural complexity, leitmotives were transcribed as a short monophonic melody into a symbolic music format and converted to a numerical tabular format suitable for melodic feature extraction using the FANTASTIC software toolbox (Müllensiefen, 2009). Four features that each capture a different aspect of melodic complexity and that had been used successfully to model cognitive behavior on melodic discrimination tests were extracted (see Müllensiefen, 2009; Harrison et al., 2016, for details): (1) Interval entropy, defined via the relative frequency of each melodic interval in the leitmotive, (2) Length, defined as the number of notes (3) Tonalness, defined as the highest of the 24 correlation coefficients as generated by the Krumhansl-Schmuckler key finding algorithm (Krumhansl, 2001). (4) Local step wise contour, defined as the mean absolute difference between adjacent values in the pitch contour vector of a melody.

These four features correlated highly across the 20 leitmotives, which suggested that they index a common dimension. Hence, principal component analysis (PCA) was used to aggregate the four features and derive a single measure of melodic complexity. The unidimensional PCA model using all four features explained 60% of the variance in the data, with Length having a relatively high uniqueness (0.59) value compared to the three other features (all values < 0.36). As a result, Length was removed and a unidimensional PCA model was run on the remaining 3 features which achieved to explain 70% of the variance in the data. PCA scores were derived from this model for each leitmotive and were used in the subsequent analysis to represent structural (i.e., melodic) complexity.

To assess audio similarity we used chromagram features (Mauch et al., 2015) that were extracted from the individual leitmotives on the recognition list. The audio data was extracted from the 10 min passage of the exposure phase of the experiment using Sonic-Annotator (Cannam et al., 2010). Chromagram features were then compared and the best alignment for each leitmotive within the 10 min passage was identified using database thresholding as implemented in the audioDB search engine (Rhodes et al., 2010).

2.4. Experiment I

2.4.1. Design

The first experiment used a within-subjects design, with identical experimental conditions for all participants. The independent variables measured were musical training, German speaking skills, Wagner affinity, as well as objective Wagner knowledge. For each leitmotive, judgments of pleasantness, perceived conveyed valence, as well as arousal ratings were also taken to gather subjective assessments of the leitmotive stimuli for the models. Questions regarding the musical material were taken in real-time during the experiment in the order listed above and information regarding individual differences were taken after the listening portion of the experiment. Our item based model also incorporated a chroma measure that was indicative of how close the probed audio stimuli used in the experiment were to the audio used in the listening portion of the experiment. The dependent variable measured was whether or not a participant was able to correctly identify a leitmotive from a listening test, as well as the participant's subjective ratings of the musical material itself.

2.4.2. Participants

For the first experiment a convenience sample ($N = 100$) was used, with additional effort made to recruit participants with either familiarity or affinity for the music of Richard Wagner from across the greater London area. The experiment was advertised over a host of mediums including posters, email lists, Twitter and general word-of-mouth to find individuals familiar

with the music of Wagner. The sample was made up of 55 females (55%) and 45 males (45%) with a mean age of 28.7 (R = 18–65, SD = 11.82). Written consent was obtained from all participants and participants had the option of accepting £7 compensation for travel and time expenses.

2.4.3. Materials and Procedure

The musical stimuli of the first experiment were based off an earlier study by Albrecht and Frieler (2014). The scene was chosen for its narrative qualities and high concentration of leitmotive material. The audio used was taken from the second scene of the first act of *Siegfried* of the 1976 Pierre Boulez *Der Ring des Nibelungen* DVD recording at the Bayreuth Festspielhaus. This scene is colloquially referred to as the *Wanderer Scene*. Excerpts chosen for probes in the memory sequence were taken from the same Boulez recording.

Twenty probes containing the leitmotives were chosen after consulting the Burghold (1910) libretto as well as the Albrecht study. The 10 probes that occurred in the *Wanderer Scene* were chosen to mirror the initial Albrecht study, each occurring with various frequencies. The 10 probes used as lures were taken from a similar narrative passage from the same recording of *Götterdömmerung*. Leitmotives used as lures in the first experiment were consequently used as "targets" (i.e., leitmotives actually contained in the 10 min audio passage) in the second experiment. After the 20 leitmotives were chosen, renditions of each leitmotive were then taken from throughout the Boulez *Der Ring des Nibelungen* to serve as audio excerpts for the test phase. When possible, probes were chosen without simultaneous sounding vocals. Data was collected using a participant response sheet generated for the purpose of this experiment.

2.5. Experiment II

2.5.1. Participants

The second experiment also used a convenience sample (N = 31) with additional effort made to recruit participants with specialized Wagner knowledge. The sample was made up of 16 females (52%) and 15 males (48%) with a mean age of 25.19 (R = 18–65, SD = 8.91). Participants from Experiment I were excluded from participating in Experiment II.

2.5.2. Materials

Participants were played a 10 min excerpt prior to Siegfried's death scene from *Götterdämmerung*. The 20 leitmotives probes for the memory test were exactly the same as in Experiment II only that their assignment to targets (*old* items) and lures (*new* items) changed given the different passage in the exposure phase. While the number of leitmotive items labeled as *old* and *new* was split evenly in the first experiment, the constraint to use the same leitmotive items as for Experiment I, resulted in 13 items *old* and 7 *new* items for Experiment II.

3. RESULTS

Across both samples the individual difference measures of Wagner knowledge and Wagner affinity were highly correlated (r = 0.71) and in order to avoid issues with multi-collinearity within the linear regression models used for analysis, both measures were subjected to a PCA which explained 85% of the variance with one dimension. Component factor scores for each participant were derived from the PCA model and were labeled as Wagner expertise.

Data modeling proceeded in three steps. The first model uses all data from both experiments and models participant responses only in terms of individual differences measures. The second model then uses significant individual difference measures identified in the first model and assesses whether the measure of structural leitmotive complexity as well as the number of occurrences of the leitmotive in the exposure phase contribute to modeling participant responses with *old* items. Here, we first assess data from Experiment I and Experiment II separately, and if model coefficients are comparable, we subsequently combine the data from both experiments. In the third step, we model responses to the *new* items including any significant individual differences measures as well melodic complexity in addition to sound similarity. All models use participants' binary responses as to whether a leitmotive was present or not in the 10 min passage during the exposure phase, scored either correct or incorrect, as the dependent variable. At all steps the data was modeled using generalized mixed effects models using participants as random effects and all models were fit using the "lme4" (Bates et al., 2015) package implemented in the statistical computing software "R" (R Core Team, 2013).

3.1. Model I: Individual Parameters

The data from all participants from Experiments 1 and 2 (N = 131) was used for the construction of Model I. Predictor variables initially specified for Model I were the Wagner expertise score, the musical training score from the Gold-MSI, and self-reported German speaking ability. In addition we used leitmotive as a second random effect in addition to participants to accommodate the fact that some leitmotive items might be generally more or less difficult. The initial model is given in **Table 1** and shows that only Wagner expertise emerged as a significant predictor of leitmotive recognition ability, while neither the musical training score nor German speaking ability reached the common significance threshold of $p < 0.05$. As a result, only Wagner expertise was retained as a fixed effects predictor and the model was refit. The refit individual differences model had a predictive accuracy of 69.9% for the participant responses and showed a significantly ($p < 0.001$ on a likelihood ratio test) better fit to the data (BIC = 3,164) than a null model only including random effects for participants and leitmotives (BIC = 3,236). In addition, the fit was not significantly worse ($p = 0.146$) than

TABLE 1 | Model I: individual differences variables.

	Coefficient	Standard Error	*p*-value
Intercept	0.61	0.20	<0.002***
Wagner Expertise	0.39	0.06	<0.001***
Musical Training	0.01	0.004	0.14
German Speaking Ability	−0.02	0.03	0.40

****p < 0.001.*

the full model including all three individual differences measures (BIC = 3,176). Therefore, we only used Wagner expertise as an individual differences measure in the subsequent modeling stages.

3.2. Model II: Old Items

For modeling responses to the old items two separate models were constructed for the data from Experiment I ($N = 100$) and II ($N = 31$). In addition to the random effect for participants and Wagner expertise as fixed predictor, number of occurrences of each leitmotive (as determined by the first author) in the exposure phase and the PCA scores for structural complexity were also included as fixed effects. Model parameters for both models were computed using the Laplace approximated maximum likelihood estimates and their 95% confidence intervals were determined by likelihood profiling. Parameter estimates and confidence intervals for both models are given in **Table 2** which shows that for all three fixed effects parameters confidence intervals overlap substantially. Specifically, the parameter estimates for Model II are contained within the corresponding confidence intervals derived for Model I, indicating that the estimates derived from the two models are not significantly different from each other. After collapsing the data from both experiments we computed a full model including all main effects as well as interactions between the individual differences in Wagner expertise and the two experimental factors of times heard and structural complexity. This can be seen in **Table 3**. We then removed the non-significant interaction between times heard and Wagner expertise and obtained the final model as given in **Table 3**. When compared on the Bayesian Information Criterion fit index, this final model fit the data substantially better (BIC = 1,635) than a null model only including Wagner expertise (BIC = 1,675), a model only including main effects (BIC = 1,642) and a model including both interaction effects (BIC = 1,640). The final model had a predictive accuracy of 68.12% In line with with one of our hypotheses, Wagner expertise had a positive effect on memory performance, while structural melodic complexity had a negative effect. Not in line with our original hypotheses, the number of times a leitmotive occurred in the exposure phase had a negative effect on recognition rates. We provide a possible explanation for this finding below.

3.3. Model III: New Items

For modeling the responses to the *new* items we followed the same modeling strategy of firstly modeling the data from Experiment I and consequently the data from Experiment II separately. Building on the results from steps 1 and 2, we included Wagner expertise as well as structural complexity as fixed effects predictors and added sound similarity based on the chromagram measure as a third predictor seen in **Table 4**. We did not include the number of times the leitmotive occurred in the exposure phase as a predictor because this variable has a constant value of zero for new items. After collapsing the data from both experiments we computed a full model including all main effects as well as interactions between the individual differences in Wagner expertise and the two experimental factors of structural complexity and chroma distance. We removed the non-significant interaction between chroma distance and Wagner expertise and obtained the final model as given in **Table 5**. When compared on the Bayesian Information Criterion fit index, this final model fit the data substantially better (BIC = 1,598) than a null model only including Wagner expertise (BIC = 1,624), a model including both interaction effects (BIC = 1,604) and was comparable to a model only have main effects (BIC = 1,597). The final mode had a predictive accuracy of 69.45%.

TABLE 2 | Model II: old items, Modeling item level data from experiment I and II separately.

	Experiment I		Experiment II	
	Coefficient	CI	Coefficient	CI
Wagner Expertise	0.87	[0.17, 0.45]	0.57	[0.21, 0.95]
Times Heard	−0.03	[−0.04, −0.01]	−0.06	[−0.11, −0.02]
Structural Complexity	−0.39	[−0.52, −0.26]	−0.13	[−0.42, 0.14]

TABLE 3 | Model II: combining item level data from experiment I and II.

	Coefficient	Standard Error	*p*-value
Intercept	0.92	0.08	<0.001***
Wagner Expertise	0.38	0.07	<0.001***
Structural Complexity	−0.32	0.06	<0.001***
Times Heard	−0.03	0.01	<0.001***
Expertise Complexity Interaction	0.24	0.06	<0.001***

$^{***}p < 0.001$.

TABLE 4 | Model III: modeling of data for new items from experiment I and II.

	Experiment I Old Items		Experiment II Old Items	
	Coefficient	CI	Coefficient	CI
Wagner Expertise	0.44	[0.27, 0.62]	0.40	[−0.04, 0.87]
Chroma Distance	1.04	[0.68, 1.42]	−1.86	[−4.93, 1.08]
Structural Complexity	0.40	[0.18, 0.62]	0.30	[0.06, 0.54]

TABLE 5 | Model III: combined data for new items.

	Coefficient	Standard Error	*p*-value
Intercept	−0.23	0.17	0.19
Wagner Expertise	0.39	0.08	<0.001***
Structural Complexity	0.35	0.08	<0.001***
Chroma Distance	0.71	0.13	<0.001***
Wager Expertise Complexity Interaction	0.23	0.09	<0.01*

$^{*}p < 0.05$; $^{***}p < 0.001$.

In line with our hypotheses, Wagner expertise has a positive effect on memory performance for new items, i.e., the ability to identify new items as not having heard before. Unlike its effect in the old item model, structural melodic complexity has a positive effect on correctly responding to new items with "not heard before" as does distance in terms of chromagram features.

4. DISCUSSION

Consistent with our initial hypothesis, the results of both experiments demonstrate that these models of leitmotive memory performance are comprehensive in that they include both individual differences variables as well as symbolic and audio features of musical structure. Model I was able to reproduce results from previous work Müllensiefen et al. (2016) demonstrating that Wagner expertise was a significant predictor of a listener's memory for musical material. Of the three individual differences variables hypothesized to contribute to an individual's leitmotive recognition rate, only Wager expertise but not general musical training nor German speaking ability emerged as a significant predictor. One reason that musical training may not have emerged as a significant predictor in the individual differences model is that musical training and Wagner expertise are correlated. Using the mixed effects models it is not possible to model correlations between predictors and in this case the stronger predictor of Wagner expertise may be suppressing the weaker predictor of musical training, thus possibly explaining the different previous findings (Müllensiefen et al., 2016) due to a different modeling technique (structural equation modeling) that can handle correlated predictors. To our knowledge, this is one of the first analyses that has used a scaled measure of musical expertise other than musical training (i.e., stylistic expertise) which accounts for the largest amount of variance explained in a participant's response, though for an exception see Farrugia et al. (2016).

In addition to musical training not emerging as significant, German speaking abilities also did not reach significance, which might be attributed to either unintelligible diction from the Wagnerian singing that would not lead to more efficient encoding or from not having enough German speakers as a part of the sample. The results of the first model serve as initial evidence for a hypothesis assuming that there are more aspects of musical expertise that can be important for modeling music perception and cognition other than solely relying on musical training as an indicator for musical skills and expertise.

The second statistical model was able to confirm the hypothesis that measures of structural complexity of items in the test phase explain part of the variance in the participants' memory response data. This is consistent with other research using similar methodologies (Dewitt and Crowder, 1986; Croonen, 1994). More specifically, the second model demonstrated that the structural complexity of a leitmotive has a negative effect on an individual's ability to recognize musical material, while the amount of times heard surprisingly displayed a negative effect. The findings on structural complexity were not surprising in light of some literature with complexity serving as a predictor of memory recall (Harrison et al., 2016). The surprising finding of the negative relationship with times heard might be attributed to a variable not measured in this experiment that is related to perceptual salience.

In the passage used, the more perceptually salient motives occur less frequently than the others used in the excerpt. After re-examining the excerpt, we believe that the perceptually salient motives are those that are easier to detect and remember from the dense auditory scene. Those motives would be structurally simpler and in fact there is a clear negative correlation between our measures of complexity and the amount of times heard in the excerpt ($r = -0.25$), which means that simpler motives occur most often. In addition, the experimental design of the memory task introduced a correlation between structural length and complexity on one hand and the number of times that a motive was played in the test phase on the other hand, because shorter motives were repeated more often during the retrieval task. It is possible that these additional repetitions could also have facilitated memory retrieval. That said, measures of compositional complexity and simplicity are not all that contribute to perceptual saliency. Gestalt principles like *Prägnanz* or uniqueness with respect to a corpus are important as well. The aspect of uniqueness is connected to principles of statistical learning and can for example be measured by second order corpus features which have already proven to be powerful predictors in previous studies on melodic memory (Müllensiefen and Halpern, 2014). To follow up on this finding, future research will focus on investigating the extent to which compositional features reflecting perceptual salience or uniqueness can be used in respect to a large and appropriate corpus such as the Barlow and Morgenstern dictionary of operatic themes (Barlow and Morgenstern, 1966).

The third model aimed at explaining how listeners make memory decisions regarding musical material that they cannot recognize from a recent listening episode. It included measures of chroma distance and structural complexity as well as a significant interaction between Wagner expertise and structural complexity. Accounting for a small, yet significant proportion of the variance, the expertise and complexity interaction provides further evidence supporting the notion that listeners with different individual characteristics can react differently to the same musical stimulus features. In particular the interaction effect suggests an interpretation that listeners with high Wagner expertise benefit from the structural complexity of the leitmotives presented more strongly to make correct decisions about the novelty of the leitmotive item. Additionally Model III also includes a component that does not reflect compositional structure in a traditional music-theoretical sense, but rather deals directly with the sound itself. Interestingly the chroma distance variable exhibits the strongest effect among the predictors in the model ($b = 0.71$) and thus provides further evidence that measures that reflect properties of sound and the musical surface can make important contributions to models of music perception and cognition.

5. CONCLUSIONS

Overall, we believe the results from this experiment are able to help close the gap between experimental work that has relied heavily on artificial designs and musical stimuli for the sake

of experimental control on one hand and research attempts to capture music listening in a more ecological setting on the other. The music of Richard Wagner has been notorious in its reputation for being difficult to comprehend, but the results from this study suggest that parsing the musical surface of something like *Der Ring des Nibelungen* is a process that is accomplished through repeated listening and active engagement with the music that does not require specialized musical training. Hearing these complex musical ideas is open to anyone and being able to hear salient musical events in a dense musical texture does not seem to be dependent on an individual's musical training. We believe that this is further evidence and reason for beginning to move closer to musical perception modeling that firstly moves away from using solely musical training as a proxy for musical ability and secondly incorporates recent work done in music informatics to help more accurately model perception.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Goldsmiths Psychology Ethics Board, Goldsmiths Department of Psychology with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Goldsmiths Psychology Ethics Board.

AUTHOR CONTRIBUTIONS

DB- Experimental design, literature review, running of participants, draft manuscript, data cleaning. DM- Experimental design, data analysis, experimental guidance and overview.

FUNDING

Funding for this research was provided by the AHRC Digital Transformations Transforming Musicology in the Digital Economy Programme AH/L006820/1.

ACKNOWLEDGMENTS

The authors would like to thank members of the London Wagner Society as well as members of Fulham Opera.

REFERENCES

Albrecht, H., and Frieler, K. (2014). "The perception and recognition of wagnerian leitmotifs in multimodal conditions," in *International Conference of Students of Systematic Musicology* (London).

Bailey, R. (1977). The structure of the "ring" and its evolution. *Nineteenth Century Music* 1, 48–61.

Barlow, H., and Morgenstern, S. (1966). *A Dictionary of Opera and Song Themes, Including Cantatas, Oratorios, Lieder, and Art Songs*. New York, NY: Crown Publishers.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Besson, M., Schön, D., Moreno, S., Santos, A., and Magne, C. (2007). Influence of musical expertise and musical training on pitch processing in music and language. *Restorat. Neurol. Neurosci.* 25, 399–410.

Bigand, E., and Poulin-Charronnat, B. (2006). Are we experienced listeners? a review of the musical capacities that do not depend on formal musical training. *Cognition* 100, 100–130. doi: 10.1016/j.cognition.2005.11.007

Bouwer, F. L., Werner, C. M., Knetemann, M., and Honing, H. (2016). Disentangling beat perception from sequential learning and examining the influence of attention and musical abilities on erp responses to rhythm. *Neuropsychologia* 85, 80–90. doi: 10.1016/j.neuropsychologia.2016.02.018

Burghold, J. (1910). Richard wagner. the ring of the nibelung. *Text with Haupt a Chlichen Leitmotifs and Musical Examples*. Mainz:. B. Schott.

Burgoyne, J. A., Bountouridis, D., van Balen, J., and Honing, H. (2013). "Hooked: a game for discovering what makes music catchy," in *Proceedings of the 14th International Society for Music Information Retrieval Conference*, eds A. de Souza Britto Jr., F. Gouyon, and S. Dixon (Curitiba), 245–250.

Cannam, C., Jewell, M. O., Rhodes, C., Sandler, M., and d'Inverno, M. (2010). Linked data and you: bringing music research software into the semantic web. *J. New Music Res.* 39, 313–325. doi: 10.1080/09298215.2010.522715

Chin, T., and Rickard, N. S. (2012). The music use (muse) questionnaire: an instrument to measure engagement in music. *Music Percept. Interdisc. J.* 29, 429–446. doi: 10.1525/mp.2012.29.4.429

Collins, T., Meredith, D., and Volk, A. (2015). "Mathematics and Computation in Music," in *Proceedings of 5th International Conference, MCM 2015*, Vol. 9110 (London: Springer).

Croonen, W. (1994). Effects of length, tonal structure, and contour in the recognition of tone series. *Percept. Psychophys.* 55, 623–632.

Cuddy, L. L., Cohen, A. J., and Miller, J. (1979). Melody recognition: the experimental application of musical rules. *Can. J. Psychol.* 33:148.

Dahlhaus, C. (1980). *Between Romanticism and Modernism: Four Studies in the Music of the Later Nineteenth Century Number 1*. Berkeley, CA: University of California Press.

Deathridge, J., and Dahlhaus, C. (1984). *The New Grove Wagner*. New York, NY: WW Norton & Company.

Deliège, I. (1989). A perceptual approach to contemporary musical forms. *Contemporary Music Rev.* 4, 213–230.

Deliège, I. (1992). Recognition of the wagnerian leitmotiv: Experimental study based on an excerpt from das rheingold. *Musik Psychol.* 9, 25–54.

Deliège, I. (1996). Cue abstraction as a component of categorisation processes in music listening. *Psychol. Music* 24, 131–156.

Deliège, I., and El Ahnmadi, A. (1990). Mechanisms of cue extraction in musical groupings: A study of perception on sequenza vi for viola solo by luciano berio. *Psychol. Music* 18, 18–44.

Deliège, I., and Mélen, M. (1997). "Cue abstraction in the representation of musical form," in *Perception and Cognition of Music,* eds I. Deliege and J. A. Sloboda (Hove: Psychology Press Ltd.), 387–412.

Dewitt, L. A., and Crowder, R. G. (1986). Recognition of novel melodies after brief delays. *Music Percept.* 3, 259–274.

Dowling, W. J. (1971). Recognition of inversions of melodies and melodie contours. *Percept. Psychophys.* 9, 348–349.

Dowling, W. J. (1972). Recognition of melodic transformations: Inversion, retrograde, and retrograde inversion. *Percept. Psychophys.* 12, 417–421.

Dowling, W. J. (1978). Scale and contour: two components of a theory of memory for melodies. *Psychol. Rev.* 85:341.

Dowling, W. J. (1986). Context effects on melody recognition: Scale-step versus interval representations. *Music Percept.* 3, 281–296.

Dowling, W. J., and Fujitani, D. S. (1971). Contour, interval, and pitch recognition in memory for melodies. *J. Acoust. Soc. Am.* 49, 524–531.

Dreyfus, L. (2012). *Wagner and the Erotic Impulse*. Cambridge, MA: Harvard University Press.

Eerola, T., Louhivuori, J., and Lebaka, E. (2009). Expectancy in sami yoiks revisited: the role of data-driven and schema-driven knowledge

in the formation of melodic expectations. *Music. Sci.* 13, 231–272. doi: 10.1177/102986490901300203

Farrugia, N., Allan, H., Müllensiefen, D., and Avron, A. (2016). "Does it sound like progressive rock? a perceptual approach to a complex genre," in *Prog Rock in Europe: Overview of a Persistent Musical Style*, ed P. Gonin (Dijon: Editions Universitaire), 197–212.

Friberg, A., Schoonderwaldt, E., Hedblad, A., Fabiani, M., and Elowsson, A. (2014). Using listener-based perceptual features as intermediate representations in music information retrieval. *J. Acoust. Soc. Am.* 136, 1951–1963. doi: 10.1121/1.4892767

Grey, T. S. (2007). *Wagner's Musical Prose: Texts and Contexts*, Vol. 3. Cambridge: Cambridge University Press.

Hacohen, R., and Wagner, N. (1997). The communicative force of wagner's leitmotifs: Complementary relationships between their connotations and denotations. *Music Percept.* 14, 445–475.

Hansen, N. C., Vuust, P., and Pearce, M. (2016). "If you have to ask, you'll never know": effects of specialised stylistic expertise on predictive processing of music. *PLOS ONE* 11:e0163584. doi: 10.1371/journal.pone.0163584

Harrison, P. M. C., Musil, J. J., and Müllensiefen, D. (2016). Modelling melodic discrimination tests: descriptive and explanatory approaches. *J. New Music Res.* 45, 265–280. doi: 10.1080/09298215.2016.1197953

Huron, D. B. (2006). *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, MA: MIT Press.

Krumhansl, C. L. (2001). *Cognitive Foundations of Musical Pitch*. Oxford: Oxford University Press.

Krumhansl, C. L., Toivanen, P., Eerola, T., Toiviainen, P., Järvinen, T., and Louhivuori, J. (2000). Cross-cultural music cognition: cognitive methodology applied to north sami yoiks. *Cognition* 76, 13–58. doi: 10.1016/S0010-0277(00)00068-8

Lerdahl, F., and Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.

Levitin, D. J. (2012). What does it mean to be musical? *Neuron* 73, 633–637. doi: 10.1016/j.neuron.2012.01.017

Magee, B. (1988). *Aspects of Wagner*. Oxford; New York, NY: Oxford University Press on Demand.

Mauch, M., MacCallum, R. M., Levy, M., and Leroi, A. M. (2015). The evolution of popular music: Usa 1960–2010. *Open Sci.* 2:150081. doi: 10.1098/rsos.150081

Morimoto, Y., Kamekawa, T., and Marui, A. (2009). "Verbal effect on memorisation and recognition of wagners leitmotifs," in *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM 2009)* (Jyväskylä), 357–361.

Müllensiefen, D. (2009). *Fantastic: Feature Analysis Technology Accessing Statistics (in a Corpus): Technical Report v1*. Technical report, Goldsmiths University of London.

Müllensiefen, D., Baker, D., Rhodes, C., Crawford, T., and Dreyfus, L. (2016). *Recognition of Leitmotives in Richard Wagner's Music: An Item Response Theory Approach*. Cham: Springer International Publishing.

Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PLoS ONE* 9:e89642. doi: 10.1371/journal.pone.0089642

Müllensiefen, D., and Halpern, A. R. (2014). The role of features and context in recognition of novel melodies. *Music Percept.* 31, 418–435. doi: 10.1525/mp.2014.31.5.418

Müller, U., and Wapnewski, P. (1992). *Wagner Handbook*. Cambridge, MA: Harvard University Press.

Pearce, M. T., and Wiggins, G. A. (2006). Expectation in melody: The influence of context and learning. *Music Percept.* 23, 377–405. doi: 10.1525/mp.2006.23.5.377

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rhodes, C., Crawford, T., Casey, M., and d'Inverno, M. (2010). Investigating music collections at different scales with audiodb. *J. New Music Res.* 39, 337–348. doi: 10.1080/09298215.2010.516832

Russell, J. A. (1980). A circumplex model of affect. *J. Person. Soc. Psychol.* 39:1161.

Schaal, N. K., Banissy, M. J., and Lange, K. (2015). The rhythm span task: comparing memory capacity for musical rhythms in musicians and non-musicians. *J. New Music Res.* 44, 3–10. doi: 10.1080/09298215.2014.937724

Serafine, M. L., Crowder, R. G., and Repp, B. H. (1984). Integration of melody and text in memory for songs. *Cognition* 16, 285–303.

Serafine, M. L., Davidson, J., Crowder, R. G., and Repp, B. H. (1986). On the nature of melody-text integration in memory for songs. *J. Mem. Lang.* 25, 123–135.

Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., et al. (2013). *Roadmap for Music Information Research*. ed G. Peeters (Creative Commons BY-NC-ND 3.0 license).

Tervaniemi, M. (2009). Musicianssame or different? *Ann. N.Y. Acad. Sci.* 1169, 151–156. doi: 10.1111/j.1749-6632.2009.04591.x

Van Balen, J. (2016). *Audio Description and Corpus Analysis of Popular Music*. Ph.D. thesis, Utrecht University.

Vempala, N. N., and Russo, F. A. (2015). An empirically derived measure of melodic similarity. *J. New Music Res.* 44, 391–404. doi: 10.1080/09298215.2015.1080284

Vuust, P., Pallesen, K. J., Bailey, C., van Zuijen, T. L., Gjedde, A., Roepstorff, A., et al. (2005). To musicians, the message is in the meter: pre-attentive neuronal responses to incongruent rhythm are left-lateralized in musicians. *Neuroimage* 24, 560–564. doi: 10.1016/j.neuroimage.2004.08.039

Wiggins, G. A., and Forth, J. (2015). "IDyOT: a computational theory of creativity as everyday reasoning from learned information," in *Computational Creativity Research: Towards Creative Machines*, eds T. R. Besold, M. Schorlemmer, A. Smaill (New York, NY: Springer), 127–148. .

Williamson, V. J., Baddeley, A. D., and Hitch, G. J. (2010). Musicians' and nonmusicians' short-term memory for verbal and musical sequences: comparing phonological similarity and pitch proximity. *Mem. Cogn.* 38, 163–175. doi: 10.3758/MC.38.2.163

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 04 May 2017
doi: 10.3389/fpsyg.2017.00666

A Dynamical Model of Pitch Memory Provides an Improved Basis for Implied Harmony Estimation

Ji Chul Kim [1, 2]*

[1] Department of Psychological Sciences, University of Connecticut, Storrs, CT, USA, [2] Oscilloscape LLC, East Hartford, CT, USA

Tonal melody can imply vertical harmony through a sequence of tones. Current methods for automatic chord estimation commonly use chroma-based features extracted from audio signals. However, the implied harmony of unaccompanied melodies can be difficult to estimate on the basis of chroma content in the presence of frequent nonchord tones. Here we present a novel approach to automatic chord estimation based on the human perception of pitch sequences. We use cohesion and inhibition between pitches in auditory short-term memory to differentiate chord tones and nonchord tones in tonal melodies. We model short-term pitch memory as a gradient frequency neural network, which is a biologically realistic model of auditory neural processing. The model is a dynamical system consisting of a network of tonotopically tuned nonlinear oscillators driven by audio signals. The oscillators interact with each other through nonlinear resonance and lateral inhibition, and the pattern of oscillatory traces emerging from the interactions is taken as a measure of pitch salience. We test the model with a collection of unaccompanied tonal melodies to evaluate it as a feature extractor for chord estimation. We show that chord tones are selectively enhanced in the response of the model, thereby increasing the accuracy of implied harmony estimation. We also find that, like other existing features for chord estimation, the performance of the model can be improved by using segmented input signals. We discuss possible ways to expand the present model into a full chord estimation system within the dynamical systems framework.

Keywords: implied harmony, tonal melody, automatic chord estimation, pitch memory, dynamical system, neural oscillation, gradient frequency neural network

OPEN ACCESS

Edited by:
Naresh N. Vempala,
Ryerson University, Canada

Reviewed by:
Dipanjan Roy,
Allahabad University, India
Jane Elizabeth Bednarz,
Texa A&M University-Commerce, USA

***Correspondence:**
Ji Chul Kim
jichulkim21@gmail.com

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *12 November 2017*
Accepted: *12 April 2017*
Published: *04 May 2017*

Citation:
Kim JC (2017) A Dynamical Model of Pitch Memory Provides an Improved Basis for Implied Harmony Estimation. Front. Psychol. 8:666. doi: 10.3389/fpsyg.2017.00666

INTRODUCTION

Melody is a succession of pitched sounds arranged to form a coherent musical pattern (Bingham, 1910; Apel, 1969). In Western tonal melodies, coherence is often achieved by organizing melodic tones to imply harmonic progressions. Although tones in a melody sound successively in time, they can convey the sense of harmony, which is a relationship among simultaneously sounding pitches, by arpeggiating a chord and connecting chord tones via nonchord tones such as passing tones and neighbor tones (Schenker, 1956; Thomson, 1999). Psychological studies have shown that implied harmony is an important feature of the perception and cognition of tonal melodies (Cuddy et al., 1981; Tan et al., 1981; Trainor and Trehub, 1994; Holleran et al., 1995; Povel and Jansen, 2002).

Automatic chord estimation is a classic research area in music informatics aimed at identifying a sequence of chords that best matches the harmonic progression of a given music signal. Current

signal-based approaches commonly employ chroma-based features such as chromagram which carry information on the energy distribution across 12 pitch classes or chromas (Jiang et al., 2011; Cho and Bello, 2014). Thus, chord estimation using these features is based on the duration and intensity of tones without taking their temporal order into account, which is consistent with the prevalent view of tonality perception and key-finding mechanisms based on pitch-class distributions (Krumhansl, 1990; Krumhansl and Cuddy, 2010). Chroma distributions are expected to be a reliable basis for chord estimation when there are more chord tones than nonchord tones in the frame of analysis. This is generally the case for harmonized music with explicit chordal support but not necessarily for unaccompanied melodies with frequent nonchord tones. Indeed, nonchord tones are recognized as a common source of errors in automatic chord estimation (Pardo and Birmingham, 2002; Lee and Slaney, 2006).

Here we present a novel feature extractor for automatic chord estimation that selectively enhances chord tones over nonchord tones on the basis of human perception of pitch sequences. Instead of analyzing chroma distributions in the acoustic signal, we use a model of human short-term pitch memory to determine the relative perceptual salience of individual tones in the signal. Psychological experiments have shown that pitches within a whole-tone range inhibit each other so that short-term retention of a pitch deteriorates when it is followed by a pitch neighbor (Deutsch, 1972, 1973; Deutsch and Feroe, 1975). Also, it has been shown that the memory of a melodic interval based on a simple frequency ratio (e.g., the perfect fifth based on 3:2) is more stable than the memory of a melodic interval based on a more complex ratio (e.g., the tritone which is approximated by 45:32) (Schellenberg and Trehub, 1994, 1996a,b). These findings suggest that melodic steps (a semitone and a whole tone) and leaps (intervals greater than a whole tone) have distinct perceptual properties: A pitch is weakened when it is followed by a step, while it becomes more salient when it forms a consonant leap with another pitch. Therefore, the salience of melodic pitches is determined not only by their duration but also by their temporal order (Bharucha, 1984; Brown, 1988) since the latter determines the pattern of steps and leaps. The differentiation between chord tones and nonchord tones may arise from the pattern of cohesion and competition among melodic pitches in short-term auditory memory, such that salient pitches that cohere together are heard as chord tones whereas pitches suppressed by others serve as nonchord tones (Kim, 2011; Kim and Large, under revision).

In this paper, we test pitch interactions arising from the pattern of melodic steps and leaps as a basis for automatic chord estimation. To model the interaction of melodic pitches in auditory memory, we use a network of tonotopically tuned nonlinear oscillators. This is not an arbitrary choice of implementation. Rather, it is based on the observation that the two distinct types of pitch interaction discussed above—inhibition by pitch neighbors and coherence based on simple frequency relationships—correspond with the two characteristic behaviors of nonlinear systems: lateral inhibition and nonlinear resonance. The model, which is described below, is a dynamical system; it is run by numerically integrating a set of differential equations which specify the dynamics and interactions of its components. Therefore, it runs forward in time (i.e., it can potentially run in realtime) and does not involve any search procedures or optimization steps that require access to an entire time series. The model is driven by audio signals, and acoustic frequencies are transformed into a complex pattern of oscillations which we take as a measure of pitch salience. We test the model with unaccompanied tonal melodies and show that chord tones are selectively enhanced in the response of the model compared to the distribution of physical tone durations.

GENERAL MATERIAL AND METHODS

Model

We model short-term pitch memory with a network of tonotopically tuned nonlinear oscillators, which is known as a *gradient frequency neural network* (abbreviated as GrFNN and pronounced *griffin*; Large et al., 2010). Nonlinear oscillation is found in many parts of the auditory system, including critical oscillations in the cochlea (Camalet et al., 2000; Hudspeth et al., 2010) and mode-locked firing of auditory subcortical neurons (Large et al., 1998; Laudanski et al., 2010). We use a generic mathematical form of nonlinear oscillation, called the canonical model, which describes oscillatory activities with complex-valued state variables (Kim and Large, 2015). GrFNNs have been used successfully to model auditory neural processing (Lerud et al., 2014, 2015) as well as music cognition (Large et al., 2015, 2016).

Here we describe the structure and function of the short-term pitch memory model with an example. (The differential equations governing the dynamics of the model are given below, along with the parameter values used in this study, but understanding of the mathematical details is not required to comprehend the results and implications of this study.) The model consists of two layers of nonlinear oscillators tuned to a chromatic scale (**Figure 1**). Layer 1 is driven by an audio signal and performs frequency analysis. **Figure 2** shows the response of the model to a passage composed by J. S. Bach for solo violin. Layer 1 oscillators resonate to different frequencies so that they separate out individual frequencies in the signal. The parameters for Layer 1 oscillators were chosen to capture the critical oscillations observed in the cochlea (see Equation 1 below for more details).

Layer 2 is a model of short-term pitch memory. High-amplitude oscillations above the on-threshold (see below) are considered active pitch traces that are salient in auditory memory. Layer 2 receives input from Layer 1 and includes internal pairwise connections between all oscillators (see **Figure 1** and Equation 2 below). Through these connections, Layer 2 oscillators either inhibit or resonate with each other depending on their frequency relationships. Two oscillators inhibit each other if their natural frequencies are a semitone or a whole tone apart. So a Layer 2 oscillation is suppressed when its stimulus tone is followed by another tone within a whole-tone distance. For example, the memory trace for the second tone (D♯6) in the Bach melody is suppressed at the onset of the following tone (E6) which is a semitone apart (**Figure 2B**). When the natural frequencies are more than a whole tone apart, the oscillators resonate together by synchronizing in an integer

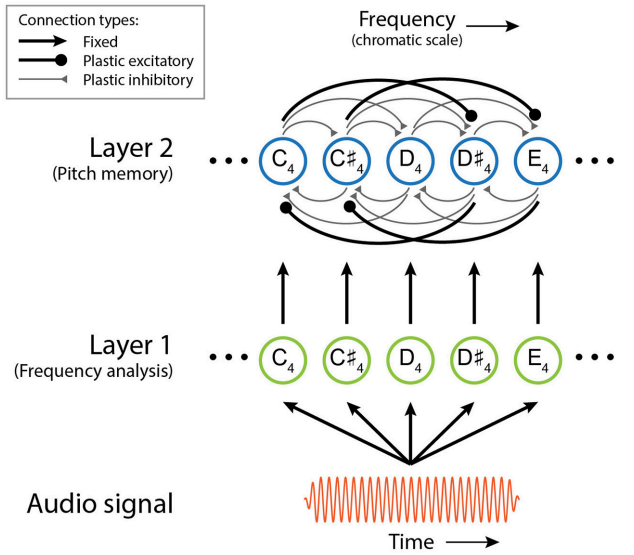


FIGURE 1 | Schematic of the dynamical model of short-term pitch memory. The colors and line widths used for different connection types are only for visual distinction and do not indicate their relative strengths.

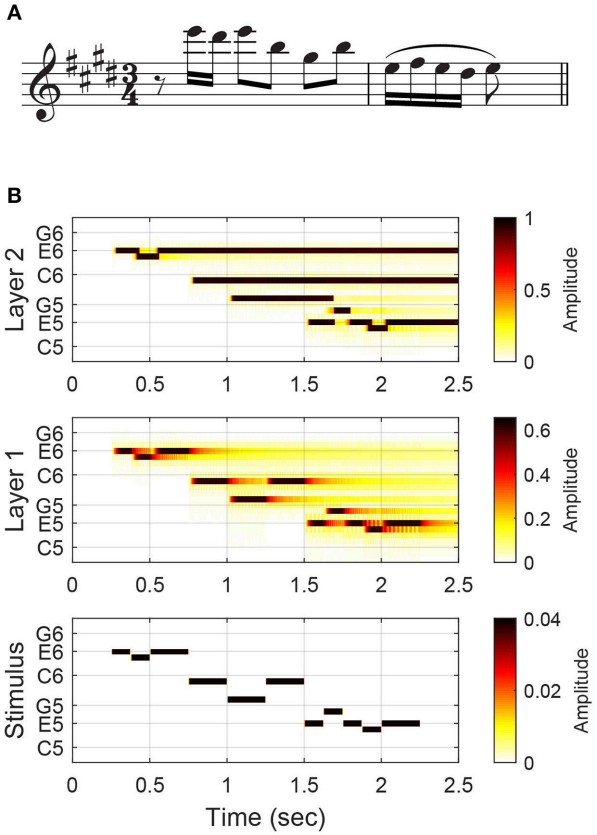


FIGURE 2 | The model's response to the opening of J. S. Bach's Violin Partita No. 3, BWV 1006, Prelude: (A) the musical score and **(B)** the amplitudes of Layer 1 and Layer 2 oscillators and stimulus tones. The stimulus (an audio signal) is depicted in a piano-roll representation. High-amplitude oscillations in Layer 2 (depicted with dark colors) are considered active pitch traces in auditory memory.

ratio (called mode-locking). Nonlinear resonance is stronger for simpler frequency relationships such as 2:1 (an octave) and 3:2 (a perfect fifth) so that oscillations driven by a consonant leap last longer than oscillations for a dissonant leap. For example, the oscillatory traces at E6 and B5, which are a perfect fifth apart, are sustained long beyond the physical duration of the tones (**Figure 2B**). The parameters for Layer 2 oscillators were chosen so that they have thresholds for turning on and off which simulates the persistence and loss of memory traces.

The pairwise connections between Layer 2 oscillators are governed by a Hebbian learning rule (Equation 3). The plastic connections model short-term adaptation in the auditory system rather than long-term learning. The connections strengthen and weaken quickly depending on the current amplitude and frequency relationship of their source and target oscillators. When two Layer 2 oscillators in a simple frequency relationship have high amplitudes at the same time, the plastic connections between them quickly strengthen and let the oscillators reinforce each other through nonlinear resonance (i.e., mode-locking). When two oscillators within a whole-tone range are activated simultaneously, the connections between them grow quickly but they introduce lateral inhibition so that the oscillator with higher amplitude (typically the one currently driven by a stimulus tone) suppresses the other oscillator. The plastic connections decay quickly as either of the oscillators goes below the off-threshold.

Let us discuss how the pitch memory model can improve the estimation of implied harmony by selectively enhancing chord tones over nonchord tones. Bach's pieces for solo instruments, such as the passage shown in **Figure 2A**, are well known for creating an impression of vertical harmony out of a single unaccompanied line (Davis, 2006). The oscillatory patterns formed in Layer 2 show how this may be possible (**Figure 2B**). The first group of notes (E-D♯-E) leaves one oscillatory trace at E6, with the trace for the neighbor tone (D♯6) confined to the time of physical sounding due to lateral inhibition. The next three notes (B-G♯-B) form consonant leaps, so their traces prolong together without inhibiting each other (note that the trace at B5 is sustained through a temporal gap). The last five notes form a turn figure made of only steps, so only the trace for the last note (E5) is extended. At the end of the passage, the oscillations at E6, B5 and E5 remain active. Along with the trace at G♯5, which prolongs beyond the note duration before being suppressed by the following F♯5, the active oscillatory traces suggest that the melody implies an E-major harmony. It is possible to estimate the chord from note durations (the chord tones take up 81% of total notated duration), but chord tones are made more salient in the response of the model (the chord tones take up 92% of total trace duration, excluding prolongations past the offset of the last note). Below we take the length of oscillatory traces as a measure of pitch salience and test if it can serve as a better basis for chord estimation than note durations.

Equations (1–3) specify the time evolution of each component in the dynamical model. (The readers may skip the equations

and proceed to the Material section.) Equation (1) describes the interaction of Layer 1 oscillators with an external signal.

$$\tau_1 \frac{dz_{1i}}{dt} = z_{1i}\left(\alpha_1 + \mathrm{i}2\pi f_i + \beta_{11}|z_{1i}|^2 + \frac{\epsilon_1\beta_{12}|z_{1i}|^4}{1-\epsilon_1|z_{1i}|^2}\right) + x(t), \tag{1}$$

where z_{1i} is a complex-valued state variable specifying the amplitude and phase of the ith oscillator in Layer 1, f_i is its natural frequency, $x(t)$ is a complex-valued external signal which can be obtained by applying the Hilbert transform to a real-valued audio signal, and the roman i is the imaginary unit. The parameters α_1, β_{11}, β_{12}, and ϵ_1 determine the intrinsic dynamics of the oscillators, and τ_1 is the time constant (see Kim and Large, 2015, for an analysis of all intrinsic dynamics available in the canonical model). The parameter values used are $\alpha_1 = 0$, $\beta_{11} = -0.1$, $\beta_{12} = -0.1$, $\epsilon_1 = 1$, and $\tau_1 = 0.0025$ (this is the critical Hopf regime, known to underlie cochlear dynamics; see Kim and Large, 2015).

Equation (2) determines the dynamics of Layer 2 oscillators (z_{2i}) which receive input from Layer 1 oscillators of identical natural frequencies (z_{1i}) as well as from all other oscillators in Layer 2 (z_{2j}).

$$\tau_{2i}\frac{dz_{2i}}{dt} = z_{2i}\left(\alpha_2 + \mathrm{i}2\pi + \beta_{21}|z_{2i}|^2 + \frac{\epsilon_2\beta_{22}|z_{2i}|^4}{1-\epsilon_2|z_{2i}|^2}\right) + c_{\mathrm{aff}}z_{1i} + \sum_{j\neq i}\sqrt{\epsilon_2}^{k_{ij}+m_{ij}-2}c_{ij}z_{2j}^{k_{ij}}\bar{z}_{2i}^{m_{ij}-1}, \tag{2}$$

where c_{ij} is a complex state variable for the plastic connection from the jth oscillator to the ith oscillator, and c_{aff} is the strength of afferent connections. k_{ij} and m_{ij} are integers that approximate the frequency ratio of the ith and jth oscillators (i.e., $k_{ij} : m_{ij} \approx f_i : f_j$), which corresponds to the ratio of mode-locking. The parameter values used are $\alpha_2 = -1.6$, $\beta_{21} = 2.2$, $\beta_{22} = -0.1$, $\epsilon_2 = 1$, $\tau_{2i} = 1/f_i$, and $c_{\mathrm{aff}} = 1.5$ (this is the subcritical double limit cycle regime which exhibits hysteresis with different on- and off-thresholds; see Kim and Large, 2015).

The evolution of plastic connections between Layer 2 oscillators (c_{ij}) is determined by a Hebbian learning rule,

$$\tau_{ij}\frac{dc_{ij}}{dt} = c_{ij}\left(\lambda_{ij} + \mu_{1ij}|c_{ij}|^2 + \frac{\epsilon_c\mu_{2ij}|c_{ij}|^4}{1-\epsilon_c|c_{ij}|^2}\right) + \sqrt{\epsilon_c}^{k_{ij}+m_{ij}-2}\kappa_{ij}z_{2i}^{m_{ij}}\bar{z}_{2j}^{k_{ij}}. \tag{3}$$

Different parameter values were used depending on the interval between the natural frequencies of the source and target oscillators. For a semitone difference: $\lambda_{ij} = -1$, $\mu_{1ij} = 0$, $\mu_{2ij} = -1$ and $\kappa_{ij} = -0.5$ (inhibitory). For a whole tone difference: $\lambda_{ij} = -1$, $\mu_{1ij} = 0$, $\mu_{2ij} = -1$ and $\kappa_{ij} = -1$ (inhibitory). For a difference greater than a whole tone: $\lambda_{ij} = -0.1$, $\mu_{1ij} = 0$, $\mu_{2ij} = -10000$ and $\kappa_{ij} = 0.02$ (excitatory). For all three cases: $\epsilon_c = 1$ and $\tau_{ij} = \frac{k_{ij}+m_{ij}}{k_{ij}f_j + m_{ij}f_i}$.

Material

We tested the dynamical model with tonal melodies from seven Mozart piano sonatas (K. 279, K. 280, K. 281, K. 282, K. 283, K. 331, and K. 545). We took the top voice from the expositions of the first movements in sonata form. For K. 311, which is a theme and variations, the melody was taken from the theme. We selected these melodies because they are accompanied by mostly unambiguous chordal support in the left hand. We relied on both the melody and the accompaniment to annotate each note in the melody with the underlying chord and whether the note is a chord tone or a nonchord tone. The Mozart melodies include ample nonchord tones (593 nonchord tones out of 2,020 notes, comprising 29% of total notes) compared to other collections we considered (e.g., nonchord tones represent only 7% of the notes in the vocal part of Schumann's *Dichterliebe*). This makes the Mozart melodies good materials to test for the differentiation between chord tones and nonchord tones. We used the annotations (based on both the melody and the accompaniment) to evaluate the model's responses to the unaccompanied melodies. The annotations should not be considered as the only possible harmonic interpretations since the harmony implied by a melody (without accompaniment) could differ from the harmony of the accompaniment (Temperley, 2007). Also, it is common knowledge that the same melody can be harmonized in many different ways. These potential discrepancies, however, would only make the model's predictions less accurate. Thus, the tests reported below should be considered conservative tests.

For each Mozart melody, we created an audio signal made of pure tones (complex-valued sinusoids) that match the notated pitches and durations in the score. An amplitude envelope was applied to each stimulus tone, with sustained amplitude of 0.04 and linear ramps of 5 ms at the onset and the offset. The use of pure tones, instead of complex tones, is due to the limitation of Layer 1 in the current form. Layer 2 is a model of short-term pitch memory which takes oscillations at individual *pitches* as input. Layer 1, however, separates individual spectral components in the audio signal rather than extracting individual pitches (or fundamental frequencies) from them. Instead of incorporating pitch estimation into the model (which requires more than frequency analysis; see, e.g., de Cheveigné, 2006), here we use audio signals containing only pure tones for which pitches can be obtained by frequency analysis alone. Currently we are developing a GrFNN pitch estimator, and the future versions of the present model will include a pitch estimator and thus be able to handle signals containing complex sounds.

Methods

For each stimulus signal, the model was run by numerically integrating Equations (1–3) using GrFNN Toolbox (Large et al., 2014), which is a software library for building and running GrFNN models. Before each integration, all oscillators and plastic connections in the model were set to random initial conditions with small amplitudes. The range of natural frequencies in the model was determined by the pitch range of the stimulus melody. The natural frequencies of the oscillators spanned from three semitones below the lowest note in the melody up to three

semitones above the highest note. For stable fixed-step numerical integration, the sampling frequency was set to 20 times the highest natural frequency in the model.

The duration of oscillatory traces in Layer 2 was taken as a measure of pitch salience. Trace duration was defined as the length of time from the moment a Layer 2 oscillation jumps above the on-threshold until either the moment it drops below the off-threshold or the next note onset at the same pitch or the offset of the last note in the signal (or the last note in the chord span for Test 2), whichever occurs first. So if a trace is extended into another trace at the same pitch, the trace duration for the first tone is counted only up to the onset of the second tone. For the parameter values used in this study, the on- and off-thresholds were 0.89 and 0.50 respectively. Note duration was defined as the length of time for which the stimulus tone stays above 50% of its maximum amplitude.

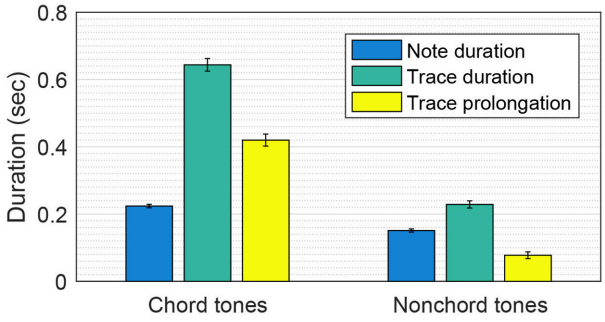


FIGURE 3 | Comparison of the trace prolongations for chord tones and nonchord tones in the Mozart melodies. Mean note duration, mean trace duration and mean trace prolongation (i.e., trace duration − note duration) are shown. The error bars indicate standard errors.

TEST 1: TRACE PROLONGATION FOR CHORD TONES AND NONCHORD TONES

To test whether chord tones are selectively emphasized in the model's response, we compared the trace durations for chord tones and nonchord tones. Given the high probability of nonchord tones being followed by a step (Bharucha, 1996), we predicted that the oscillatory traces driven by nonchord tones would mostly end soon after the note offsets while the traces for chord tones would often prolong beyond the note durations. We tested this prediction by comparing the difference between trace duration and note duration (hereafter, *trace prolongation*) for chord tones and nonchord tones.

Methods

The model was run for each of the Mozart melodies separately (see General Material and Methods above for details). For each note in the melodies (marked either as a chord tone or a nonchord tone), note duration, trace duration and trace prolongation (= trace duration − note duration) were determined. A *t*-test was performed to determine if chord tones and nonchord tones had significantly different trace prolongations.

Results and Discussion

The chord tones in the Mozart melodies had significantly longer trace prolongations than the nonchord tones [two-sample *t*-test: $t(2,018) = 12.07$, $p < 0.001$]. The mean trace prolongations for chord tones and nonchord tones were 420 and 76 ms, respectively (see **Figure 3**). This means that the chord tones were more emphasized in the pitch memory model than in the note durations. The note durations for chord tones and nonchord tones were also significantly different [mean durations: 224 and 151 ms; $t(2,018) = 8.57$, $p < 0.001$]. However, this difference does not explain the difference in trace prolongation because the trace prolongation for an isolated tone does not depend on the note duration, provided that the tone is long enough to activate an oscillatory trace (which is true for all notes in the Mozart melodies). Thus, longer trace prolongations for chord tones are attributed to the nonlinear interaction between oscillatory traces (i.e., inhibition and resonance) in conjunction with the fact that nonchord tones are followed by step more often (91% of the time in the Mozart melodies) than chord tones are (52%).

It is important to note that chord tones are selectively enhanced in the pitch memory model because of the regularities in the use of chord tones and nonchord tones in tonal music. A basic rule of counterpoint states that a nonchord tone (or a dissonance) must be resolved by step motion (Zarlino, 1558; Fux, 1725). The pitch traces for nonchord tones are prolonged to a lesser extent than the traces for chord tones because nonchord tones are mostly followed by a step whereas chord tones have no such restriction. If the opposite was true (i.e., chord tones were followed by a step while nonchord tones had no constraint), nonchord tones would be emphasized in the response of the model. Then, one could ask why chord tones and nonchord tones are used in certain ways, which is by no means limited to Western tonal music (Erickson, 1984; Thomson, 1999). It is reasonable to assume that the way melodic pitches interact in auditory memory has guided and constrained the way chord tones and nonchord tones are used in tonal music. The function of nonchord tones is to embellish chord tones without undermining their structural and perceptual prominence. Thus, one would want to limit the salience of nonchord tones while highlighting chord tones. Stepwise resolution of nonchord tones, which leads to the suppression of their pitch salience, may be viewed as a compositional practice evolved under the selective pressure by the principles of pitch organization in auditory memory.

TEST 2: TRACE DURATIONS WITHIN CHORD SPANS

The comparison of trace prolongations illustrates an important difference in the way chord tones and nonchord tones are used and perceived in tonal melodies, but it does not necessarily show that the prolonged traces contribute to better chord estimation. This is because the above analysis associates the entire length of a trace with the annotated function of the stimulus tone within the chord span in which its note duration falls. It is

possible that the oscillatory trace for a chord tone extends into the next chord span where it is not a chord tone, and this would compromise the accuracy of chord estimation. As shown in **Figure 4**, trace prolongations beyond the current chord span may strengthen or weaken the prominence of chord tones in the next chord span. For example, the trace at E5 starting in the first chord span prolongs into the second span where it remains a chord tone, thereby enhancing the representation of the chord tones. On the other hand, the trace at D5 that begins in the second chord span becomes a nonchord-tone trace in the next span. (It could be argued that this response is not necessarily wrong because the chord annotation is based on both the melody and the accompaniment, while the model is driven by the melody only. It is an empirical question, which is beyond the scope of this study, to what extent the model's response corresponds with the human perception of unaccompanied melodies.)

To investigate the effect of trace prolongation across chord spans, we compared the traces at chord pitches and nonchord pitches within individual chord spans regardless of the origin of the traces. The difference between the total trace durations for chord pitches and the total trace durations for nonchord pitches was taken as the perceptual salience of the annotated chord in the model's response. To evaluate the model's contribution to chord estimation over note durations, the difference in trace duration was then compared to the difference in total note duration between chord tones and nonchord tones in each chord span.

Methods

The simulation data obtained for Test 1 were used for the analysis of individual chord spans. For each annotated chord span, trace durations and note durations were summed for chord pitches and nonchord pitches separately. The chord boundaries used for calculating trace durations were shifted forward by 40 ms to reflect the typical rise time of Layer 2 oscillations after the stimulus onset. For each chord span, the differences between chord tones and nonchord tones in total trace duration and total note duration were calculated. A *t*-test was performed to determine whether the trace duration differences and the note duration differences are significantly different.

Results and Discussion

Figure 5 (top) shows the trace duration difference and the note duration difference for each chord span in the theme of K. 331. The graph reflects our observations above. For the second chord span, the trace duration difference is greater than the note duration difference (meaning chord pitches are more emphasized in the model response than in the note durations), while it is the opposite for the third chord span (chord pitches less prominent in the model). For K. 331, the mean trace duration difference between chord pitches and nonchord pitches was 1304 ms, and the mean note duration difference was 973 ms.

Considering all 405 chord spans in the seven Mozart melodies, trace duration differences and note duration differences were significantly different [paired-sample *t*-test: $t(404) = 6.21$, $p < 0.001$], with the mean values of 1056 ms (trace duration differences) and 567 ms (note duration differences) (see **Figure 6**). This suggests that, overall, the dynamical model's response can provide a better basis for chord estimation than note durations.

TEST 3: TRACE DURATIONS WITHIN SEGMENTED CHORD SPANS

Despite the overall advantage of trace duration over note duration, there are chord spans for which trace duration performs worse than note duration (see **Figure 5**, top). As discussed above, the prolongation of pitch traces across chord

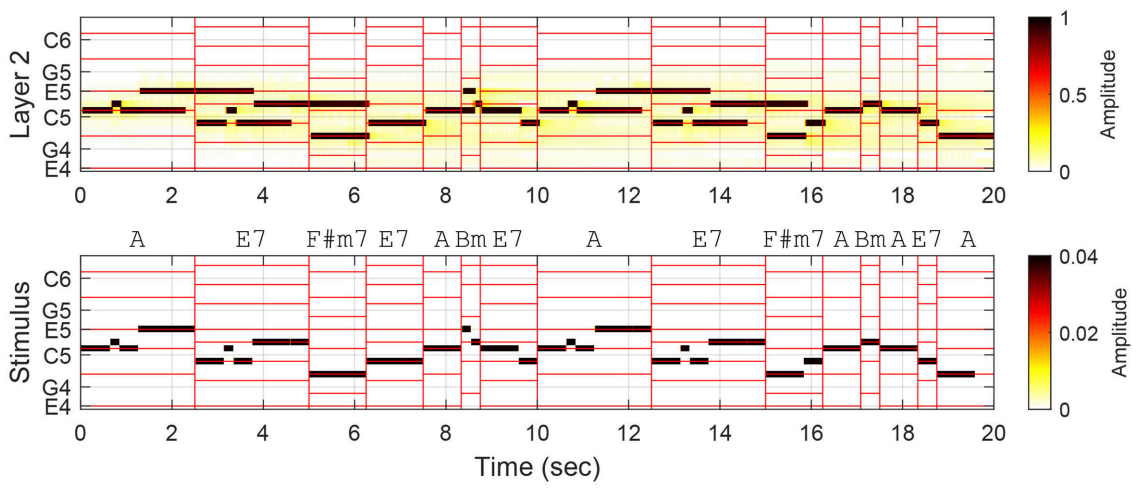


FIGURE 4 | Oscillatory traces formed in Layer 2 in response to the first two phrases (the first 15 chord spans) in Mozart Piano Sonata No. 11, K. 331, Theme. Vertical red lines demarcate chord spans, and horizontal lines indicate the pitches belonging to the chords. Chord annotations are based on both the melody and the accompaniment.

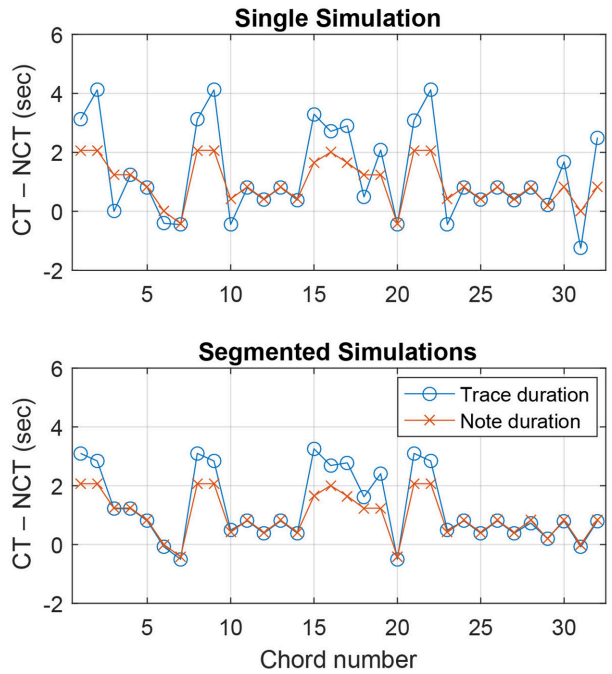


FIGURE 5 | Difference between chord pitches and nonchord pitches in total trace duration and total note duration within each chord span in Mozart Piano Sonata, K. 331, Theme. The top panel shows a single simulation run with the entire melody, and the bottom panel shows simulations for individual chord spans run separately. CT and NCT denote chord tones and nonchord tones.

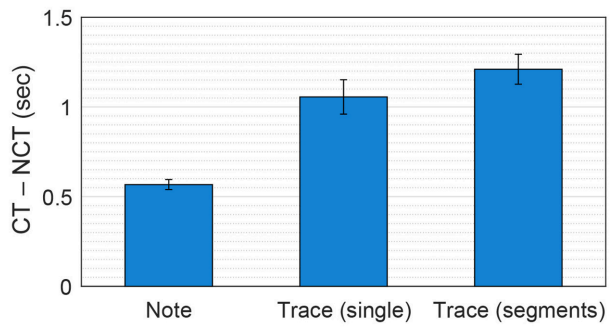


FIGURE 6 | Mean difference between chord pitches and nonchord pitches in note duration, trace duration in single simulations and trace duration in segmented simulations, averaged over all chord spans in the seven Mozart melodies. The error bars indicate standard errors. CT and NCT denote chord tones and nonchord tones.

boundaries could result in less accurate chord representations. This issue points to the importance of segmentation in chord estimation. Previous studies have shown that the accuracy of chord estimation can be improved by synchronizing analysis frames to the beat of the music being analyzed, which tends to align with harmonic changes (Bartsch and Wakefield, 2001; Bello and Pickens, 2005). We tested whether chord estimation based on the pitch memory model could be improved by using segmented stimulus signals. Instead of running the model for entire melodies, we chopped the melodies into individual chord spans and ran the model for each segment separately. This would prevent previous oscillatory traces from extending into the current chord span because each simulation starts anew from small random initial values.

Methods

A separate stimulus signal was prepared for each chord span in the Mozart melodies (total 405 segments; see General Material and Methods for the general procedures of stimulus preparation), and the model was run for each individual segment separately. As was done for Test 2, the total trace durations and total note durations for chord pitches and nonchord pitches were calculated for each chord span. A *t*-test was performed to determine if trace duration differences and note duration differences are significantly different in segmented chord spans.

Results and Discussion

Figure 5 (bottom) shows trace duration differences and note duration differences for the segmented simulations of K. 331. It can be seen that the trace duration difference is either comparable or greater than the note duration difference for all chord spans. Over all seven melodies, the trace duration differences for segmented simulations (1,211 ms on average) were significantly greater than those for single simulations in Test 2 [$t(404) = 3.16$, $p < 0.01$; see **Figure 6**]. This shows that, as was found for previous methods using chroma-based features, chord estimation based on the pitch memory model can benefit from processing each chord span separately.

GENERAL DISCUSSION

In this paper, we presented a first step toward automatic chord estimation based on nonlinear dynamics, which draws on research in music cognition and auditory neuroscience. As an alternative to the current methods of feature extraction for chord estimation, we used a dynamical model of short-term pitch memory to predict the relative salience of pitches in tonal melodies. We modeled cohesion and competition between melodic pitches as dynamic pattern formation in a gradient frequency neural network, which is a biologically realistic model of auditory neural processing. We tested the model with a collection of unaccompanied melodies and showed that it can provide better mid-level representations for chord estimation than the distribution of note durations which current chroma-based features are aimed to extract from the music signal. It was shown that chord tones are rendered more prominent in the model's response than in the note durations and that the advantage of the model can be increased by using segmented input signals.

The present study is an attempt to bridge music informatics with music cognition by developing a chord estimation method

based on the human perception of implied harmony. Much progress has been made in automatic chord estimation, with state-of-the-art systems employing cutting-edge techniques in signal processing and machine learning (see Cho and Bello, 2014; McVicar et al., 2014, for reviews). Recently, however, a plateau in performance was observed despite continuous incorporation of new data-driven methods which have proven to be successful in other machine learning domains (Humphrey and Bello, 2015). This calls for examination of the underlying assumptions of current chord estimation methods and also encourages incorporation of the findings in other related disciplines such as music cognition and auditory neuroscience. Here we showed that the pattern of pitch salience in the dynamical model of auditory short-term memory can provide a better feature for automatic chord estimation than the chroma distribution in the audio signal. The success of the present method demonstrates that human perception and underlying neural mechanisms can provide foundations for breakthroughs in music informatics research. It also warrants further investigation as to whether the dynamical models of auditory neural processing can improve the retrieval of other musical information.

The dynamical model of short-term pitch memory presented in this paper differs from previous models of echoic memory in which individual pitch traces, once initiated, decay monotonically independent of each other (e.g., Huron and Parncutt, 1993; Leman, 2000; Toiviainen and Krumhansl, 2003). In the present model, a pitch trace may sustain for a long time or be suppressed quickly at the offset of the stimulus tone depending on its interaction with other pitch traces, which is consistent with experimental findings on short-term pitch memory (Deutsch, 1972, 1973; Deutsch and Feroe, 1975; Schellenberg and Trehub, 1994, 1996a,b). The pitch dynamics observed in the present model also provides a psychological basis for the music-theoretical concept of *prolongation*, a central principle of the hierarchical organization of tonal music. In Schekerian analysis, prolongation refers to the ways in which a pitch or harmony remains active without physically sounding (Katz, 1935; Forte and Gilbert, 1982; Larson, 1997). The prolongation of pitch traces beyond note durations and the subordination of pitch traces to strong neighbors in the present model correspond directly with the idea of prolongation in music theory.

The dynamical model presented in this paper acts as a feature extractor that provides a novel mid-level representation for chord estimation. Hence, it does not perform chord estimation or labeling by itself. There are multiple ways to use the model for automatic chord estimation. For example, the current methods for estimating chords from feature representations (e.g., template matching and stochastic models) could be applied to the output of the present model. However, our ultimate goal is to expand the current model to perform chord estimation within the dynamical systems framework. This may be done by adding another layer of oscillators that holds information about common chord types by means of long-term Hebbian learning. The present model utilizes short-term plasticity to capture the interaction between pitch traces in short-term auditory memory. Adding long-term plastic connections to the model would lead to pattern formation in two different time scales, and the learning and recognition of common chord types could be modeled in terms of the interaction between layers with plasticity of different time scales.

The introduction of long-term plasticity also means the incorporation of the top-down influence of learned knowledge into the dynamical model. Cognitive psychologists have shown that listeners internalize regularities in tonal music through passive exposure and that the implicit knowledge thus acquired influences subsequent perceptions (Krumhansl, 1990; Tillmann et al., 2000; Pearce and Wiggins, 2012; Rohrmeier and Rebuschat, 2012). The model presented in this paper includes only afferent connections from the stimulus to Layer 1 and then to Layer 2, and the plastic connections adjust quickly to the current states of the oscillators. Thus, the response of the model reflects only the pattern of pitch salience in the short-term context. An extra layer with long-term plastic connections could carry information about frequently encountered chord types beyond the short-term context and modulate the activities in Layer 2 through efferent (top-down) connections. In this way, the influence of both short-term context and long-term knowledge could be accounted for within the dynamical systems framework.

We showed that the prominence of chord tones in the model's response could be raised by using segmented signals. This is because running the model separately for each segment prevents oscillatory traces from intruding into the next segment. The same effect can be achieved by deactivating (or resetting) oscillatory traces at segmentation boundaries while running the model continuously with the entire (unsegmented) signal. Segmentation would benefit chord estimation the most if it aligns with chord span boundaries. Above we used segmentations based on chord annotations, but this information is not available to a system performing automatic chord estimation (actually, that is the information such a system aims to obtain). One possible way to incorporate segmentation into the present model is to couple it with a rhythm model that synchronizes to a musical beat and meter (e.g., Large et al., 2015). In the same spirit as the use of beat-synchronized frames for chroma-based features, the pitch memory model could receive a modulatory signal from the rhythm model which deactivates pitch traces at the time of each downbeat. The pitch memory model, on the other hand, could provide input to the rhythm model at the time of harmonic change, which is an important cue for the perception of rhythm and meter (cf. Papadopoulos and Peeters, 2008).

Here we tested the dynamical model with unaccompanied melodies to focus on the differentiation of chord tones and nonchord tones in the absence of explicit chordal context. We found that the model selectively enhanced chord tones in the melodies, thus raising the probability of correct chord estimation. The results of this study prompt us to ask how well the model would handle music with multiple voices. We predict that the model would still show an advantage over raw pitch-class content. The presence of

vertical consonant intervals, which typically form between chord tones, would facilitate the suppression of nonchord tones. Also, we expect the model to capture pitch dynamics within individual voices as it did for single unaccompanied melodies. This prediction will have to be tested in future studies.

AUTHOR CONTRIBUTIONS

JK designed and ran the model and wrote the paper.

FUNDING

This work was supported by NSF BCS-1027761 and AFOSR FA9550-12-10388.

ACKNOWLEDGMENTS

The author wishes to thank Edward W. Large, Karl Lerud, Jung Nyo Kim, and two reviewers for their helpful comments and suggestions on earlier versions of the manuscript.

REFERENCES

Apel, W. (1969). *The Harvard Dictionary of Music, 2nd Edn*. Cambridge, MA: Belknap Press.

Bartsch, M. A., and Wakefield, G. H. (2001). "To catch a chorus: using chroma-based representations for audio thumbnailing," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY: IEEE), 15–18.

Bello, J. P., and Pickens, J. (2005). "A robust mid-level representation for harmonic content in music signals," in *Proceedings of the 6th International Conference on Music Information Retrieval* (London: Queen Mary, University of London), 304–311.

Bharucha, J. J. (1984). Anchoring effects in music: the resolution of dissonance. *Cogn. Psychol.* 16, 485–518. doi: 10.1016/0010-0285(84)90018-5

Bharucha, J. J. (1996). Melodic anchoring. *Music Percept.* 13, 383–400. doi: 10.2307/40286176

Bingham, W. V. D. (1910). Studies in melody. *Psychol. Rev. Monogr. Suppl.* 12, i–88. doi: 10.1037/h0093021

Brown, H. (1988). The interplay of set content and temporal context in a functional theory of tonality perception. *Music Percept.* 5, 219–249. doi: 10.2307/40285398

Camalet, S., Duke, T., Jülicher, F., and Prost, J. (2000). Auditory sensitivity provided by self-tuned critical oscillations of hair cells. *Proc. Natl. Acad. Sci. U.S.A.* 97, 3183–3188. doi: 10.1073/pnas.97.7.3183

Cho, T., and Bello, J. P. (2014). On the relative importance of individual components of chord recognition systems. *IEEE/ACM Trans. Audio Speech Lang. Process.* 22, 477–492. doi: 10.1109/TASLP.2013.2295926

Cuddy, L. L., Cohen, A. J., and Mewhort, D. J. K. (1981). Perception of structure in short melodic sequences. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 869–883. doi: 10.1037/0096-1523.7.4.869

Davis, S. (2006). Implied polyphony in the solo string works of J. S. Bach: a case for the perceptual relevance of structural expression. *Music Percept.* 23, 423–446. doi: 10.1525/mp.2006.23.5.423

de Cheveigné, A. (2006). "Multiple F0 estimation," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, eds D. Wang and G. J. Brown (Piscataway, NJ: IEEE Press; Wiley), 45–79.

Deutsch, D. (1972). Mapping of interactions in the pitch memory store. *Science* 175, 1020–1022. doi: 10.1126/science.175.4025.1020

Deutsch, D. (1973). Interference in memory between tones adjacent in the musical scale. *J. Exp. Psychol.* 100, 228–231. doi: 10.1037/h0035440

Deutsch, D., and Feroe, J. (1975). Disinhibition in pitch memory. *Percept. Psychophys.* 17, 320–324. doi: 10.3758/BF03203217

Erickson, R. (1984). A perceptual substrate for tonal centering? *Music Percept.* 2, 1–5. doi: 10.2307/40285278

Forte, A., and Gilbert, S. E. (1982). *Introduction to Schenkerian Analysis*. New York, NY: Norton.

Fux, J. J. (1725). *Steps to Parnassus. The Study of Counterpoint*. New York, NY: W. W. Norton & Company.

Holleran, S., Jones, M. R., and Butler, D. (1995). Perceiving implied harmony: the influence of melodic and harmonic context. *J. Exp. Psychol. Learn. Mem. Cogn.* 21, 737–753. doi: 10.1037/0278-7393.21.3.737

Hudspeth, A. J., Jülicher, F., and Martin, P. (2010). A critique of the critical cochlea: hopf–a bifurcation–is better than none. *J. Neurophysiol.* 104, 1219–1229. doi: 10.1152/jn.00437.2010

Humphrey, E. J., and Bello, J. P. (2015). "Four timely insights on automatic chord estimation," in *Proceedings of the 16th International Society for Music Information Retrieval Conference* (Málaga), 673–679.

Huron, D., and Parncutt, R. (1993). An improved model of tonality perception incorporating pitch salience and echoic memory. *Psychomusicology* 12, 154–171. doi: 10.1037/h0094110

Jiang, N., Grosche, P., Konz, V., and Müller, M. (2011). "Analyzing chroma feature types for automated chord recognition," in *Audio Engineering Society Conference: 42nd International Conference: Semantic Audio* (Ilmenau).

Katz, A. T. (1935). Heinrich Schenker's method of analysis. *Music. Q.* XXI, 311–329. doi: 10.1093/mq/XXI.3.311

Kim, J. C. (2011). *Tonality in Music Arises from Perceptual Organization*. Unpublished doctoral dissertation, Northwestern University.

Kim, J. C., and Large, E. W. (2015). Signal processing in periodically forced gradient frequency neural networks. *Front. Comput. Neurosci.* 9:152. doi: 10.3389/fncom.2015.00152

Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch*. New York, NY: Oxford University Press.

Krumhansl, C. L., and Cuddy, L. L. (2010). "A theory of tonal hierarchies in music," in *Music Perception*, Vol. 36, eds M. Riess Jones, R. R. Fay, and A. N. Popper (New York, NY: Springer), 51–87.

Large, E. W., Almonte, F. V., and Velasco, M. J. (2010). A canonical model for gradient frequency neural networks. *Phys. D Nonl. Phenom.* 239, 905–911. doi: 10.1016/j.physd.2009.11.015

Large, E. W., Herrera, J. A., and Velasco, M. J. (2015). Neural networks for beat perception in musical rhythm. *Front. Syst. Neurosci.* 9:159. doi: 10.3389/fnsys.2015.00159

Large, E. W., Kim, J. C., Flaig, N. K., Bharucha, J. J., and Krumhansl, C. L. (2016). A neurodynamic account of musical tonality. *Music Percept.* 33, 319–331. doi: 10.1525/mp.2016.33.3.319

Large, E. W., Kim, J. C., Lerud, K. D., and Harrell, D. (2014). *GrFNN Toolbox: Matlab Tools for Simulating Signal Processing, Plasticity and Pattern Formation in Gradient Frequency Neural Networks*. Available online at: https://github.com/MusicDynamicsLab/GrFNNToolbox

Large, E. W., Kozloski, J. R., and Crawford, J. D. (1998). "A dynamical model of temporal processing in the fish auditory system," in *Association for Research in Otolaryngology Abstracts Vol. 21*. (St. Petersburg, FL), 717.

Larson, S. (1997). The problem of prolongation in tonal music: terminology, perception, and expressive meaning. *J. Music Theor.* 41, 101–136. doi: 10.2307/843763

Laudanski, J., Coombes, S., Palmer, A. R., and Sumner, C. J. (2010). Mode-locked spike trains in responses of ventral cochlear nucleus chopper and onset neurons to periodic stimuli. *J. Neurophysiol.* 103, 1226–1237. doi: 10.1152/jn.00070.2009

Lee, K., and Slaney, M. (2006). "Automatic chord recognition from audio using a supervised HMM trained with audio-from-symbolic data," in *AMCMM '06 Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia* (Santa Barbara, CA: ACM Press), 11–20.

Leman, M. (2000). An auditory model of the role of short-term memory in probe-tone ratings. *Music Percept.* 17, 481–509. doi: 10.2307/40285830

Lerud, K. D., Almonte, F. V., Kim, J. C., and Large, E. W. (2014). Mode-locking neurodynamics predict human auditory brainstem responses to musical intervals. *Hear. Res.* 308, 41–49. doi: 10.1016/j.heares.2013.09.010

Lerud, K. D., Kim, J. C., Almonte, F. V., Carney, L. H., and Large, E.W. (2015). "A canonical nonlinear cochlear model," in *Association for Research in Otolaryngology Abstract*, Vol. 38 (Baltimore, MD), 211–212.

McVicar, M., Santos-Rodriguez, R., Ni, Y., and Bie, T. D. (2014). Automatic chord estimation from audio: a review of the state of the art. *IEEE/ACM Trans. Audio Speech Lang. Process.* 22, 556–575. doi: 10.1109/TASLP.2013.2294580

Papadopoulos, H., and Peeters, G. (2008). "Simultaneous estimation of chord progression and downbeats from an audio file," in *IEEE International Conference on Acoustics, Speech, and Signal Processing* (Las Vegas, NV: IEEE), 121–124.

Pardo, B., and Birmingham, W. P. (2002). Algorithms for chordal analysis. *Comput. Music J.* 26, 27–49. doi: 10.1162/014892602760137167

Pearce, M. T., and Wiggins, G. A. (2012). Auditory expectation: the information dynamics of music perception and cognition. *Top. Cogn. Sci.* 4, 625–652. doi: 10.1111/j.1756-8765.2012.01214.x

Povel, D.-J., and Jansen, E. (2002). Harmonic factors in the perception of tonal melodies. *Music Percept.* 20, 51–85. doi: 10.1525/mp.2002.20.1.51

Rohrmeier, M., and Rebuschat, P. (2012). Implicit learning and acquisition of music. *Top. Cogn. Sci.* 4, 525–553. doi: 10.1111/j.1756-8765.2012.01223.x

Schellenberg, E. G., and Trehub, S. E. (1994). Frequency ratios and the discrimination of pure tone sequences. *Percept. Psychophys.* 56, 472–478. doi: 10.3758/BF03206738

Schellenberg, E. G., and Trehub, S. E. (1996a). Children's discrimination of melodic intervals. *Dev. Psychol.* 32, 1039–1050. doi: 10.1037/0012-1649.32.6.1039

Schellenberg, E. G., and Trehub, S. E. (1996b). Natural musical intervals: evidence from infant listeners. *Psychol. Sci.* 7, 272–277. doi: 10.1111/j.1467-9280.1996.tb00373.x

Schenker, H. (1956). *Free Composition: Volume III of New Musical Theories and Fantasies, 2nd Edn*. Longman music series. New York, NY: Longman.

Tan, N., Aiello, R., and Bever, T. G. (1981). Harmonic structure as a determinant of melodic organization. *Mem. Cogn.* 9, 533–539. doi: 10.3758/BF03202347

Temperley, D. (2007). The melodic-harmonic 'divorce' in rock. *Popular Music* 26, 323–342. doi: 10.1017/S0261143007001249

Thomson, W. (1999). *Tonality in Music: A General Theory*. San Marino, CA: Everett Books.

Tillmann, B., Bharucha, J. J., and Bigand, E. (2000). Implicit learning of tonality: a self-organizing approach. *Psychol. Rev.* 107, 885–913. doi: 10.1037/0033-295X.107.4.885

Toiviainen, P., and Krumhansl, C. L. (2003). Measuring and modeling real-time responses to music: the dynamics of tonality induction. *Perception* 32, 741–766. doi: 10.1068/p3312

Trainor, L. J., and Trehub, S. E. (1994). Key membership and implied harmony in Western tonal music: Developmental perspectives. *Percept. Psychophys.* 56, 125–132. doi: 10.3758/BF03213891

Zarlino, G. (1558). *The Art of Counterpoint. Part Three of Le Istitutioni Harmoniche*. Music theory translation series. New Haven, CT: Yale University Press.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 26 April 2017
doi: 10.3389/fpsyg.2017.00639

Modeling Timbre Similarity of Short Music Clips

Kai Siedenburg[1]* *and Daniel Müllensiefen*[2]

[1] Department of Medical Physics and Acoustics, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany,
[2] Department of Psychology, Goldsmiths University of London, London, UK

There is evidence from a number of recent studies that most listeners are able to extract information related to song identity, emotion, or genre from music excerpts with durations in the range of tenths of seconds. Because of these very short durations, timbre as a multifaceted auditory attribute appears as a plausible candidate for the type of features that listeners make use of when processing short music excerpts. However, the importance of timbre in listening tasks that involve short excerpts has not yet been demonstrated empirically. Hence, the goal of this study was to develop a method that allows to explore to what degree similarity judgments of short music clips can be modeled with low-level acoustic features related to timbre. We utilized the similarity data from two large samples of participants: Sample I was obtained via an online survey, used 16 clips of 400 ms length, and contained responses of 137,339 participants. Sample II was collected in a lab environment, used 16 clips of 800 ms length, and contained responses from 648 participants. Our model used two sets of audio features which included commonly used timbre descriptors and the well-known Mel-frequency cepstral coefficients as well as their temporal derivates. In order to predict pairwise similarities, the resulting distances between clips in terms of their audio features were used as predictor variables with partial least-squares regression. We found that a sparse selection of three to seven features from both descriptor sets—mainly encoding the coarse shape of the spectrum as well as spectrotemporal variability—best predicted similarities across the two sets of sounds. Notably, the inclusion of non-acoustic predictors of musical genre and record release date allowed much better generalization performance and explained up to 50% of shared variance (R^2) between observations and model predictions. Overall, the results of this study empirically demonstrate that both acoustic features related to timbre as well as higher level categorical features such as musical genre play a major role in the perception of short music clips.

Keywords: short audio clips, music similarity, timbre, audio features, genre

OPEN ACCESS

Edited by:
Frank A. Russo,
Ryerson University, Canada

Reviewed by:
Blair Kaneshiro,
Stanford University, USA
Morwaread Mary Farbood,
New York University, USA

***Correspondence:**
Kai Siedenburg
kai.siedenburg@uni-oldenburg.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *24 October 2016*
Accepted: *10 April 2017*
Published: *26 April 2017*

Citation:
Siedenburg K and Müllensiefen D (2017) Modeling Timbre Similarity of Short Music Clips. Front. Psychol. 8:639. doi: 10.3389/fpsyg.2017.00639

1. INTRODUCTION

There is growing evidence that human listeners are able to instantly categorize short music clips containing complex mixtures of sounds, e.g., when scanning a radio dial or browsing through a playlist. Even more, the information contained in clips lasting only a few hundred milliseconds or less seems to be sufficient to perform tasks such as genre classification (Gjerdingen and Perrott, 2008; Mace et al., 2011; Plazak and Huron, 2011) or artist and song recognition (Schellenberg et al., 1999; Krumhansl, 2010).

More specifically, Gjerdingen and Perrott (2008) played participants audio excerpts of commercially available music at different lengths and asked them to indicate the genre of each excerpt. They found that 44% of participants' genre classifications of 250 ms excerpts were identical to the classifications by the same participants and of the same audio track when played for 3 s, demonstrating that listeners extract a considerable amount of information from very short excerpts. Results by Schellenberg et al. (1999) showed that even 100 ms excerpts could be matched to song title and artists with above-chance accuracy, and that time- varying high frequency information (> 1 kHz) was particularly important for correct identification. Similarly, Krumhansl (2010) showed that listeners are able to identify the artists and titles for 25% of a stimulus set consisting of 400 ms clips of popular music spanning four decades. Mace et al. (2011) were able to demonstrate that even at 125 ms length participants were able to achieve an accuracy of 54% on a genre recognition task which had a guessing level of 20%. At this timescale there are few, if any discernible melodic, rhythmic, harmonic or metric relationships to base judgements on. Though when musical-structural information is minimal, timbral information can still be rich.

Timbre is here understood as an umbrella term that denotes the bundle of auditory features (other than pitch, loudness, duration) that contribute to both sound source categories and sound quality (McAdams, 2013). In fact, timbre seems to be processed even from very short stimulus durations. For instance, Bigand et al. (2011) showed that variability in the spectral envelope can be processed from sounds as short as 50 ms. More recent results by Suied et al. (2014) have shown that listeners can even recognize timbre based on snippets as short as 16 ms (depending on the instrument family). In the latter study, performance increased monotonically with the length of the excerpts and plateaued at around 64 ms.

Building on this research, Musil et al. (2013) devised an individual differences test that investigates differences in the ability to extract information from short audio clips and to use it for similarity comparisons. This test forms part of the Goldsmiths Musical Sophistication battery of listing tests (Müllensiefen et al., 2014) and complements other individual differences tests that focus on melodic memory and beat perception abilities. The sound similarity test was designed to assess the ability to decode and compare complex musical sound textures and to be independent of temporal processing and memory capabilities and therefore only makes use of very short musical stimuli. While the test has been used in practice and proved to be fairly unrelated to other musical listening abilities (Müllensiefen et al., 2015), it has been difficult to build a model based on audio features that would describe participants' similarity judgements adequately (Musil et al., 2013).

On the contrary, there is a rich literature on audio features associated with computer-based instrument identification (Joder et al., 2009), genre classification (e.g., Andén and Mallat, 2011), the prediction of affective qualities (Laurier et al., 2009; McAdams et al., 2017), or more general aspects of the perception of audio excerpts (Alluri and Toiviainen, 2010). Audio features are most commonly derived from the Short-Time Fourier Transform of the music signal, from which spectral or temporal statistics are computed. A standard example are summary statistics such as the mean (i.e., centroid) or spread of short-time spectra, or the correlation of spectra across consecutive time windows (spectral *flux*). It is important to note that the utility of specific timbre descriptors as well as the size of feature sets varies considerably across computational and perceptual tasks. In effect, timbre description in psychology traditionally employs a handful of, say, less than 10 features, whereas many music information retrieval approaches rely on audio representations with a substantially higher dimensionality (Siedenburg et al., 2016a).

None of the psychological studies on short audio clips has used audio features to quantitatively model human perceptual responses to very short audio clips. For that reason, it is currently unclear to which extent simple categorization judgements can be predicted by low-level properties of the audio signal, as opposed to higher level concepts such as genre potentially inferred from the audio. But constructing a cognitively adequate model of audio similarity is not only useful for understanding what features and cues listeners extract and process from short audio clips. It can also serve as a first step for constructing future adaptive versions of individual differences tests of audio classifications that could allow a systematic scaling of difficulty of sets of audio clips by selecting clips that are more or less similar.

This paper aims to contribute toward the understanding of perceptual judgements of similarity for short music clips via a modeling approach. The present contribution is the first study to systematically quantify the extent to which similarity data of short musical excerpts can be explained by acoustic timbre descriptors. A notable feature of the current approach is that we not only evaluate the constructed statistical models in terms of their accuracy in describing a given set of observations, but also in their capacities to generalize to unseen data sets. The predictive accuracy of low-level timbre features is further compared with variables that encode meta information in the form of the genre and release date of songs.

This manuscript is organized as follows. In Section 2, we describe the experimental samples, stimuli, and procedures that provide the basis for our modeling study. In Section 3, the structure of the model is described in detail, in particular with regards to the audio features, normalization schemes, and statistical models of perceptual similarity. In Section 4, the presented models are comprehensively evaluated, before potential implications on timbre modeling are discussed in Section 5.

2. EXPERIMENTS

This study uses data from two separate experiments that used a sorting paradigm to assess the perceptual similarity of short music clips. In both cases the sorting paradigm was part of a larger test battery on several aspects of music perception (Müllensiefen et al., 2014). Only the data gathered via the similarity sorting paradigm is reported in this paper and has not been reported previously. The Ethics Board of Goldsmiths, University of

London approved the research undertaken and reported in the manuscript.

2.1. Participants

Sample I comprised responses from 137,339 participants who took part in the BBC Lab UK's online test *How Musical Are You?* in 2011 and 2012. The sample of participants is identical to the sample reported by Müllensiefen et al. (2014, Study 4), although the data from the sound similarity test has not been reported previously. In the training sample, 45.2% of the participants were female and mean age was 35.2 years (SD = 15). Participants were mainly UK residents (66.9%) but because the *How Musical Are You?* test was an open online application, the sample also included participants from other mainly Western and English-speaking countries (largest proportions: USA: 14.2%, Canada: 2.3%, Australia: 1.1%). The sample contained a large spread in terms of education (undergraduate degree/professional qualification: 34.1%, still in education: 23.4%, postgraduate degree: 19%, second school degree with around 18 years (e.g., British A-levels): 11.8%, first school degree around 16 years (e.g., British GCSE/O-levels): 7.5%, etc.) as well as in terms of the current profession of the participants (Other: 19.4%, Education/Training: 12.4%, Unemployed: 10.7%, Information technology: 7.1%, etc.). Only 1.8% stated "music" as their occupation. Participants in Sample I were tested with 400 ms excerpts.

Sample II comprised responses from 648 participants, collected via several experimental batteries that were run at Goldsmiths University of London between 2011 and 2014, all of which contained the sound similarity test using 800 ms excerpts. Participants came from a young student population (undergraduates as well as postgraduates) and were less diverse in terms of their educational and occupational background than participants in Sample I[1].

2.2. Stimuli

Prototypical but less well-known songs from four different genres were selected as experimental stimuli, as described by Musil et al. (2013). Because genre boundaries may be subjective and change over time (Gjerdingen and Perrott, 2008), we used the main four meta-genres identified by Rentfrow and Gosling (2003) as a guidance and selected the most prominent popular music style within each meta-category: jazz from reflective/complex, rock from intense/rebellious, pop from upbeat/conventional, and hip-hop from energetic/rhythmic. Additionally, following Krumhansl's (2010) finding that the approximate recording date of a song can be identified fairly accurately from short excerpts, specific decades were selected for each genre: 1960–70s for jazz, 1970–80s for rock, 1990–2000 for pop and hiphop. Exemplary songs for each of these genres were selected from the suggestions of prototypical songs given on the encyclopedic music datbase allmusic.com. In order to avoid the recognition of specific overly well-known tunes, songs were only selected if they were not present in the all-time top-100 Billboard charts and had never reached the top rank on the UK Billboard charts. However, two of the selected songs (*The Sign*, *I Wanna Love You Forever*) had reached first and third ranks of the US Hot-100 Billboard charts, respectively. Hence, we cannot rule out the possibility that individual participants might have recognized the songs of individual excerpts. Aiming for representative sound fragments, excerpts from each song were chosen such that the excerpt did not contain any human voice, there were at least two recognizable notes in the excerpt, and the fragment represented as much a possible the maximal timbral diversity (i.e., maximum number of instruments) of the song. In addition, the excerpt was preferably taken from a repeated section of the song. A table with all song titles, artists, and the corresponding genre is given in Table 1 of the Appendix (Supplementary Materials).

Excerpts were extracted directly from .wav files taken from the original CD recordings and stored at an audio sampling rate of 44.1 kHz. For the computation of audio features, all clips were converted to mono by summing both stereo channels. For the two experiments, excerpts of lengths 400 ms (Sample I) and 800 ms (Sample II) were used, extracted from different locations in the song, to which a 20 ms fade-in and fade-out was added. We needed to work with different stimulus durations in Samples I and II because in the original sound similarity sorting task (Müllensiefen et al., 2014), genre was used as a proxy for sound similarity, based on the fact that songs that belong to the same genre are often characterized by similarities in sound (e.g., see Rentfrow et al., 2011). In the absence of a perceptual-computational model of sound similarity at the stage of designing the experimental task, genre was the best proxy available to select groups of songs that would sound similar and at the same time different from other groups of songs, thus allowing to tentatively score the performance on the sorting task of each participant. But from the analysis of the behavioral data obtained for the 400ms excerpt set it became clear that many participants scored close to chance level. After piloting different clip lengths, a duration of 800ms then seemed to produce a distribution of performance scores that better allowed to characterize inter-individual differences.

2.3. Experimental procedure

The experimental paradigm was similar to the one used by Gingras et al. (2011) and Giordano et al. (2010). The participants' task was to listen to 16 short excerpts and to sort them into four groups of four items each by their similarity in sound. We deliberately avoided the term "genre" in the instructions and did not specify the nature of the sound similarity. Excerpts were identified by icons on a computer screen, while groups corresponded to boxes. Participants could listen to an excerpt by hovering over its icon, and could move icons around by clicking and dragging. Participants were allowed to listen to each clip as many times as they wished and change their sorting solution as often as necessary. There was no time constraint for the task and participants submitted their sorting solution when they felt that it could not be amended further. Only the final sorted state was recorded and used for subsequent analysis.

[1]Because participants of Sample II were aggregated from several individual experiments, unfortunately it was impossible at this stage to track participants' individual demographic information.

2.4. Data Characteristics

Pairwise perceptual similarity was defined as the relative number of times two clips were placed in the same group by participants. This measure is obtained by dividing the absolute number of times two clips were placed in the same group by the respective number of participants in each sample. The corresponding distribution of similarities with range between zero and one is shown in **Figure 1** (left panel).

Recall from Section 2.1 that the demographics of the participant populations from Samples I and II were not matched. In order to rule out potentially confounding effects of demographics on the similarity data, we drew subsamples of Sample I that better matched the demographics of the college-student population of Sample II. Among the 137,399 participants, there were 32,329 participants specifically with age between 18 and 24 years. Thereof 18,639 participants stated "At university" as occupational status, 3,199 participants stated "Education/Training" as their occupation, and 1,957 participants belonged to both categories. However, Pearson correlations between the similarities derived from these subsamples and the set of all participants were very strong, all $r(118) > 0.992$ ($p < 0.001$), which speaks against a pertinent influence of demographics.

Note that the diagonal entries of the similarity matrices depicted in **Figure 1** (two rightmost panels) play a distinct role. In fact, they derive from representing the data in matrix form and not from participants' direct classifications themselves (who only encountered distinct clips). The value of the diagonal entries of the matrix automatically equals one, regardless of participants' responses (because every clip trivially shares its own group). However, their inclusion in the model bears the danger of inflating figures of merit such as the coefficient of determination R^2. Because by simply differentiating identical and non-identical clips with a binary variable, one can readily obtain highly significant fits with the similarity data. For that reason, we took a conservative stance and only considered non-identical pairs for the following modeling, corresponding to the lower triangular dissimilarity matrix without diagonal entries (accordingly, the distribution of similarities depicted in the left panel of **Figure 1** only represents non-identical pairs). This makes the interpretation of R^2 coefficients more meaningful, but also reduced their magnitude by more than 20% points on average.

3. MODEL STRUCTURE

Modeling the similarity data comprised three main stages: (i) feature extraction from the audio clips, (ii) feature normalization, and the (iii) modeling of pairwise similarities of features. More specifically, we used two sets of audio features, both of which contained 24 features. Both sets were normalized in five different ways (but the normalized features were not pooled). The resulting pairwise distances of clips' audio features were then used as predictor variables in a latent-variable linear regression technique, namely partial least-squares regression (PLSR). Specifically, PLSR attempts to find the multidimensional direction (i.e., the latent variables) in the space of the predictor variables that best explains the maximal variance of the dependent variables. **Figure 2** visualizes the three modeling stages. The basic model structure is similar to the timbre dissimilarity model presented by Siedenburg et al. (2016b), but complements stage i) with an additional set of features, considers an array of normalization schemes in stage ii), and applies the model to the case of short music clips instead of isolated instrument tones.

3.1. Feature Sets

The two feature sets were: (i) a set of 24 timbre descriptors and (ii) 12 Mel-frequency cepstral coefficients (MFCCs) as well as 12 of their Δ-coefficients. In addition we also combined both sets to obtain a third feature set (iii) with 48 features.

3.1.1. Timbre Descriptors

We used the Timbre Toolbox (v1.2, Peeters et al., 2011), a large set of audio descriptors that describes the acoustic structure of audio signals with a focus on timbre. For the current purpose, we selected 24 out of its 164 descriptors. This selection possessed great overlap with the 34 descriptors used in Siedenburg et al. (2016b), which had provided a fairly robust model of musical timbre dissimilarity of isolated musical tones, each played individually on instruments of the Western orchestra. In contrast to the isolated tone case, however, ten of the twelve temporal

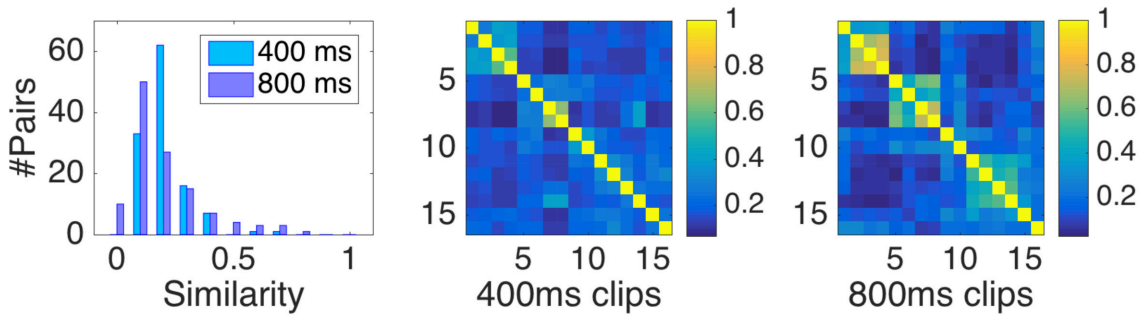


FIGURE 1 | (Left panel) Distribution of similarity data, here defined as the relative number of shared classifications of two clips. **(Middle and right panel)** pairwise similarities for Samples I (400 ms) and Sample II (800 ms).

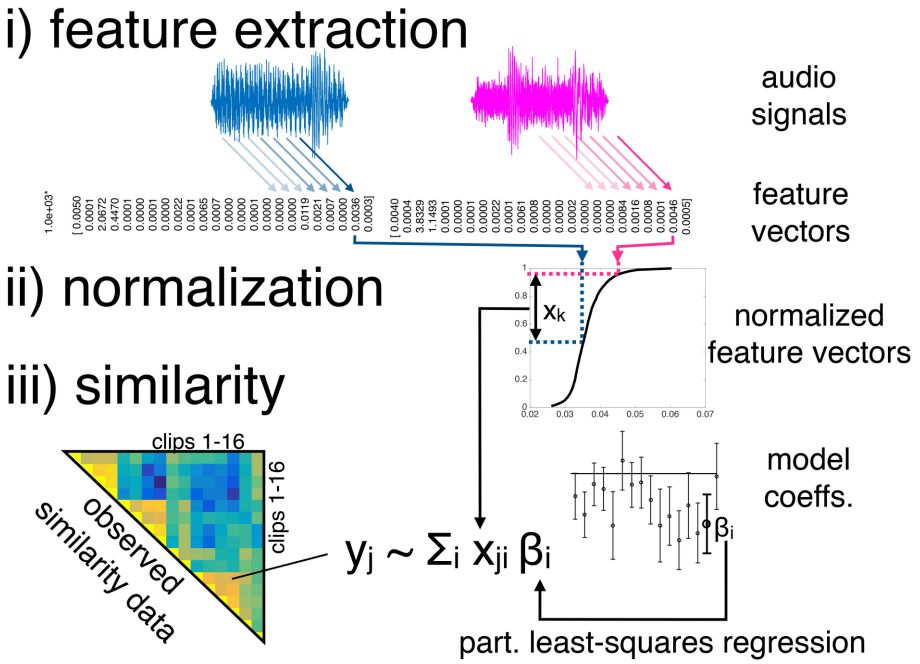


FIGURE 2 | Basic model structure. Note that every feature dimension is processed separately before being joined in the regression model.

descriptors were not taken into account for the description of clips, because it could be assumed that measures of attack or release-duration would not differ in any meaningful way across the currently used clips, given that they were extracted from the midsts of songs and contained dense musical textures.

Spectral shape descriptors were computed from an ERB-spaced Gammatone filterbank decomposition of the signal. They were measured with (fairly common) settings of 25 ms time frames with 1/2 overlap and summarized via the median and interquartile range as measures of central tendency and variability, respectively. Spectral descriptors included the first four moments of the spectral distribution, such as the spectral centroid that has been shown to correlate with perceived brightness (McAdams, 2013). Additional descriptors of the spectral distribution such as the decrease and flatness were also included, measuring spectral slope with an emphasis on lower frequencies and the peakiness of the spectrum, respectively, but also measures of spectrotemporal variation, relevant to capture spectrotemporal variability (the so-called spectral *flux*) (McAdams et al., 1995). We included four descriptors that were based on the time domain representation of the signal: the frequency and amplitude of energy modulation over time, and the median and interquartile range of the zero crossing rate. A full list of the descriptors is given in Table 2 in the Appendix (Supplementary Materials).

3.1.2. Mel-Frequency Cepstral Coefficients

As an alternative set of features, we considered the commonly-used Mel-frequency cepstral coefficients (MFCCs, Eronen, 2001) and their temporal derivatives. MFCCs are derived via a discrete cosine transform of the log-transformed power of Mel spectra. MFCCs thus represent the shape of an audio signal's spectral envelope: going up from lower to higher coefficients, MFCCs encode increasingly finer scales of spectral detail. MFCCs are standard in various tasks in audio content analysis and music information retrieval and have also been proposed as descriptors for timbre perception (see the review in Siedenburg et al., 2016a). In the current study, we used the first 12 MFCCs and their corresponding 12 ΔMFCCs, i.e., their first derivative over time. Both were computed for 25 ms time frames (1/2 overlap) of the audio signal, and the resulting time series was summarized by the median. These features were provided by the MIRtoolbox (v1.6.1, Lartillot and Toiviainen, 2007).

3.2. Feature normalization

In order to regularize the often idiosyncratic distributions of the raw feature values, five normalization schemes were considered:

N1) None (i.e., using raw feature values),
N2) Range normalization to [0, 1],
N3) Z-scores with zero mean and unit standard deviation,
N4) Rank transformation according to the test set: replacing a feature value by the fraction l/L, with l being the feature value's rank within the test set of size L,
N5) Rank transformation according to a corpus: replacing a feature value by the fraction l'/L', with l' being the feature value's rank within the corpus of size L'.

The corpus was obtained by extracting clips from a freely-available audio data set sampled at 44.1 kHz (Homburg et al.,

2005). We selected 110 songs for each of the four meta-genres of the current test set (jazz, rock, pop, hiphop), from which we extracted ten 800 ms clips each. The resulting 4,400 clips constituted our corpus. All of the above mentioned features were extracted from each clip of the corpus and used for the corpus-based ranking.

3.3. Similarity Modeling via Partial Least-Squares Regression

Per clip, each feature provided one scalar value. For any pair of clips, feature-wise distances were obtained by taking the absolute difference of the pair's respective feature values. These distances were summarized in a design matrix X of size $m \times n$, where $m = 120 = 16 \cdot 15/2$ denotes the number of pairs, and n denotes the number of features. As outlined above, in a first step we used three sets with (i) $n = 24$ timbre features (from the Timbre Toolbox, TT), (ii) $n = 24$ (Δ)MFCCs, and (iii) $n = 48$ features in the combined set.

In order to handle collinearity of predictors (Peeters et al., 2011), we used partial least-squares regression (PLSR, Geladi and Kowalski, 1986; Wold et al., 2001). PLSR is a regression technique that projects the predicted and observed variables onto respective sets of latent variables, such that the resulting variables' mutual covariance is maximized. More precisely, given a dependent variable y and an design matrix X, PLSR generates a latent decomposition such that $X = TP' + E$ and $y = Uq' + F$ with loadings matrices P ($n \times k$) and q ($1 \times k$), and components ("scores") T ($m \times k$) and U ($m \times k$) plus error terms E and F. The matrix W^* ($n \times k$) comprises the predictors' weights, such that $T = XW^*$. The regression coefficients for the original design matrix can be obtained by $\beta = W^*q'$ (cf., Wold et al., 2001), which yields a link to the generic multiple linear regression (MLR) design via $y = X\beta + F$. The decomposition maximizes the covariance of T and U, which yields latent variables that are optimized to capture the linear relation between observations and predictions. In this sense, PLSR also differs from principal component analysis (PCA) followed by MLR, as for instance used by Alluri et al. (2012), since PCA does not specifically adapt the latent decomposition to the dependent variable of interest.

In order to prevent overfitting of the response variable, the model complexity k can be selected via cross-validation. We used a model with $k = 2$ latent components, which yielded minimal 8-fold cross-validation errors in a majority of the model and evaluation conditions. We used the implementation provided by the plsregress.m function as part of MATLAB version R2015b (The MathWorks, Inc., Natick, MA), which applies the SIMPLS algorithm (De Jong, 1993).

The importance of individual predictors in the PLSR model was assessed by bootstrapping, which eventually allowed us to construct sparse regression models. For each of the two training conditions, the significance of the individual model coefficients β_i ($i = 1, ..., n$) was estimated by bootstrapping the 95% confidence interval of the coefficients (Efron and Tibshirani, 1994; Mehmood et al., 2012). We used a percentile-type method, that is, from the 16 clips per stimulus set, the similarity data of four randomly selected clips (drawn with replacement) were left out from the sample (yielding on average around 60% of the data points intact). This process was repeated 1,000 times. For every coefficient β_i the resulting 0.025 and 0.975 percentiles were taken as confidence boundaries. If confidence intervals did not overlap with zero, a predictor's contribution was considered to be significant, and the respective feature was selected for the sparse model.

4. MODEL EVALUATION

The goal of the subsequent model evaluation was to identify from among the three different feature sets and the five different normalization schemes an accurate and robust model of the perceptual similarity data. We place a special focus not only on the question how accurately a statistical model can be fitted to training data, but also on how well the model generalizes to a new set of perceptual data gathered from a different set of audio excerpts. This question is addressed by including sparse models in the subsequent evaluations that are known to generalize better to new datasets (Friedman et al., 2009) and by permuting the data from Sample I and Sample II as training and testsets. This means, each model is both fitted and tested on the datasets from Sample I (400 ms clips) and Sample II (800 ms clips). This results in 2 $\times$ 2 evaluation conditions per model. This evaluation setup also allows us to investigate the question how well a model describes the data set it was fitted to and to what degree it might be overfitted to the training data.

The evaluation proceeds in four steps. We first present results for the three feature sets in combination with all five normalization conditions. Secondly, we select a subset of the most relevant features from each model via bootstrapping and recompute the performance of the resulting sparse models. Thirdly, we consider the role of meta information such as genre and the release date of recordings. Finally, we discuss the role of individual acoustic features.

4.1. Results: The Effect of Feature Sets and Normalization Schemes

Table 1 presents the squared Pearson correlation coefficients R^2 for the three full feature sets and five normalization schemes, corresponding to the proportion of variance shared between the model predictions and empirical observations. The results indicate that the perceptual similarities of the 400 ms and 800 ms clips were both predicted with fairly similar accuracy. The combination of the two feature sets, TT+MFCC, yielded the highest R^2 values on training sets as could be expected from the larger pool of features to draw from. However, there are obvious differences between model fits derived on the training sets and model generalization to novel test sets, which suggests that all models considered at this point generalize rather poorly to unseen data. In fact, successful generalization is a rare exception with only four out of 30 predictions of unseen data yielding correlations that are significant at the $\alpha = 0.01$ level. Generalization of models based on MFCCs was particularly poor and did not provide a single significant correlation on a novel test set.

TABLE 1 | R^2 coefficients as performance indicators for full models derived from combining five normalization schemes (N1–N5) and three feature sets (TT, Timbre Toolbox; MFCC, MFCC coefficients and MFCC delta coefficients), each evaluated in the two training and testing conditions from 400 (I) and 800 ms (II) clips.

			TEST									
			N1		N2		N3		N4		N5	
			Raw		Range		z-scores		*r*-test		r-corpus	
			I	II	I	II	I	II	I	II	I	II
TRAIN	TT	I	–	–	0.07	0.08	0.07	0.06	0.33	–	0.11	0.09
		II	–	–	–	0.18	–	0.19	–	0.22	–	0.25
		Mean		–		0.08		0.08		**0.14**		0.11
	MFCC	I	0.08	–	0.21	–	0.20	–	0.21	–	0.23	–
		II	–	0.15	–	0.20	–	0.20	–	0.22	–	0.24
		Mean		0.06		0.10		0.10		0.11		**0.12**
	TT+MFCC	I	–	–	0.25	0.06	0.25	–	0.47	–	0.27	–
		II	–	–	–	0.24	–	0.26	–	0.39	–	0.33
		Mean		–		0.14		0.13		**0.22**		0.15

Note that R^2 coefficients < 0.06 that correspond to non-significant correlations at $p > 0.01$ are not displayed for the sake of clarity. The mean for each combination of normalization scheme and feature set across the four training-test evaluation conditions is given in the last row of each cell (non-significant correlations are considered a zero entry in the computation of the mean). Best average performance per feature set is given in bold font.

In terms of the normalization schemes, models using the test-set-based ranking (N4) produced the highest performance values overall. In particular for the combined feature set TT+MFCC, it yielded the best fit to the training data, potentially indicating that participants rely on relative differences within a specific acoustic context (namely the test set), rather than on absolute differences of acoustic features. **Figure 3** (top left panel) shows the scatterplot of the corresponding TT+MFCC (N4) model in all four evaluation conditions, graphically depicting how model fits decrease from when training and test dataset are identical to when datasets for model training and test differ. This decrease in model fit may be interpreted as an indicator of model overfitting. Hence, in the next evaluation step we aim to achieve better generalization performance and avoid overfitting by applying feature selection.

4.2. Feature Selection

We applied the feature selection approach described in Section 3.3 to obtain sparse models. This naturally led to different configurations of significant predictors per model and evaluation condition, which are displayed in **Figure 4**. The plot shows that the selection was fairly consistent across the different feature sets, in the sense that the combined feature set TT+MFCC roughly comprised the features already selected for the individual sets TT and MFCC. For the 400 ms clips, an average of 2.2, 3.0, and 4.6 significant variables were retained for the TT, MFCC, and TT+MFCC features sets, respectively (averaged across the five normalizations). For 800 ms clips, an average number of 4.4, 2.2, and 5.6 features were retained for the three respective feature sets. However, note that the set of features selected for the 400 and 800 ms clips is quite different. In particular for the test-ranked normalization (N4), the two sets do not share any common member.

Table 2 shows the results in all conditions for the sparse models[2]. There are 14/30 significant correlations for unseen test data, which is an improvement compared to the full models (4/30), yet still surprisingly low overall. The mean performance of sparse models (across all four evaluation conditions) was rather similar to the performance of the full models, which means the increase in generalization performance was traded against a decrease of accuracy on the training sets. The best model was obtained by the combined model TT+MFCC with the test-rank normalization (N4), with an average fit of $R^2 = 0.29$ on the training data and $R^2 = 0.14$ on novel test data. The detailed scatterplots of predictions and observations are shown in **Figure 3**.

Figure 4 (right panel) gives an overview and summary of the behavior of all models considered so far in terms of accuracy and generalization capacities: the x-axis displays the mean R^2 coefficients on novel test sets (i.e., the average of the off-diagonal values of the previously presented tables) and the y-axis represents the fit on the training sets (i.e., average of diagonal values). Generally, we are interested in a reasonable trade-off between both measures, which currently only appears to be achieved by the sparse TT+MFCC models with test-rank normalization (N4). In this sense, the figure illustrates two important methodological results: (a) The combined feature set TT+MFCC is superior to both TT and MFCC, (b) sparse variable

[2]The number of PLSR latent components k was naturally reduced to $k = 1$, if only one feature was selected by bootstrapping in the respective condition.

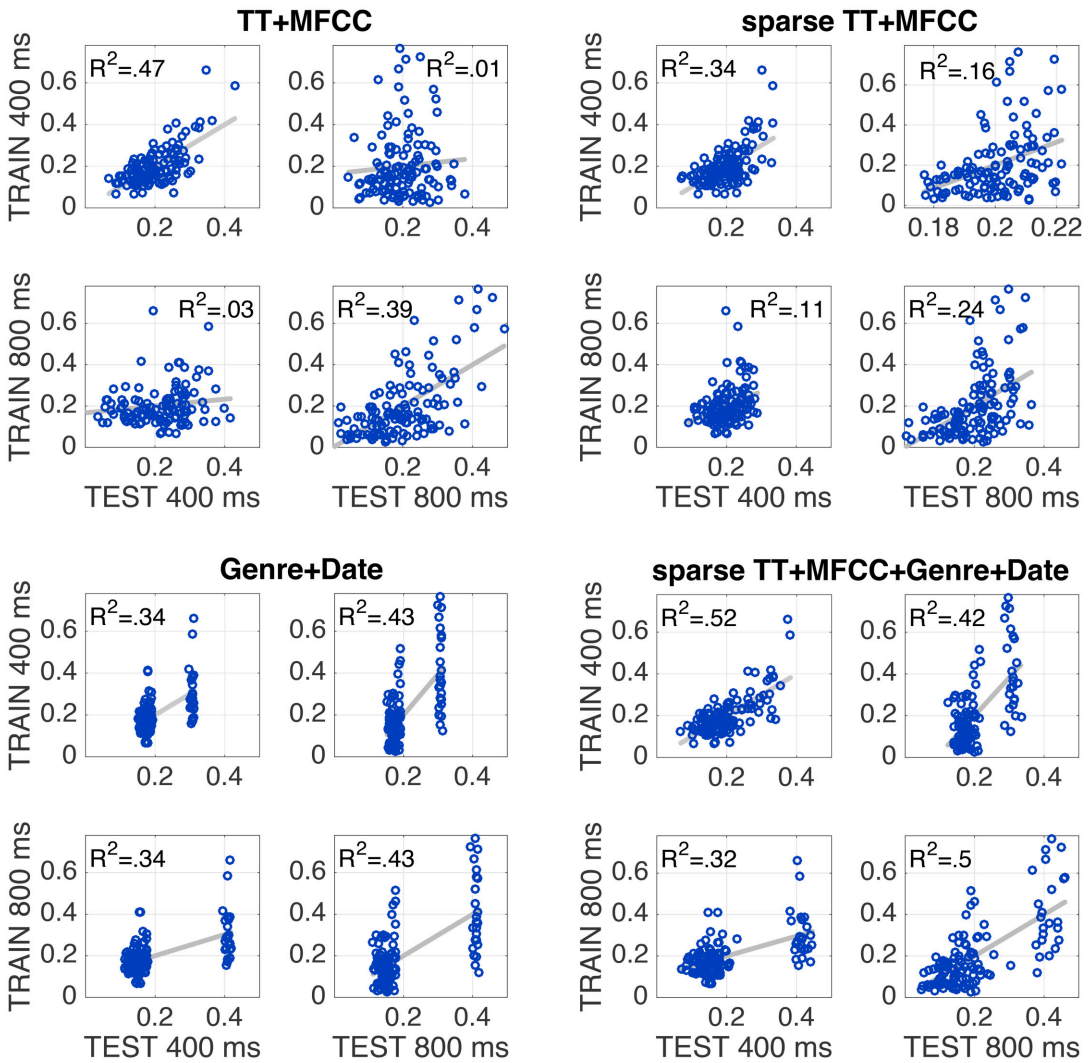


FIGURE 3 | Every individual plot shows the correspondence between model predictions (x-axis) and empirically observed similarities (y-axis). Every panel shows models that were trained to 400 ms (panel top row) or 800 ms clips (panel bottom row), and tested on the same two sets (left vs. right columns). Top left panel shows full feature set; top right: sparse feature selection; bottom left: non-acoustic variables only; bottom right: sparse model together with non-acoustic variables. All models utilize test-ranked features (N4).

selection is a means to trade accuracy on the training set against a greater ability of the models to generalize to unseen datasets.

4.3. The Role of Genre and Release Date

In a final step, we included non-acoustic information as predictor variables that were taken from the meta-data of the clips. Specifically, we considered the categorical variable of genre as well as the songs' release dates. However, is it important to keep in mind that the concept of genre is notoriously ambiguous (Craft et al., 2007). In the current case, genre was correlated not only with the release date of recordings, but also with recording techniques, instrumentation, and thus also with qualitative timbral similarity modeled by the continuously varying audio features utilized here. Therefore, this step was of exploratory nature and attempted to set the prediction results of the acoustic model into relation with approaches relying on meta information.

Genre was coded as binary predictor G indicating whether two clips shared the same genre (G = 0) or not (G = 1). As **Figure 3** (bottom right panel) demonstrates, adding these two predictors to the model with the best generalization performance, the sparse test-ranked-normalized TT+MFCC model, yielded a substantial increase in model performance of at least 18 percentage points in R^2. At the same time, the model that solely utilizes meta information (Genre+Date) robustly partitions the underlying pairs into fairly similar vs. dissimilar pairs. The computational analyses presented in the present paper thus confirm the efficiency of genre as a proxy for the selection of stimuli (as described in Section 2.2), with genre explaining the vast majority of the variability in the behavioral data.

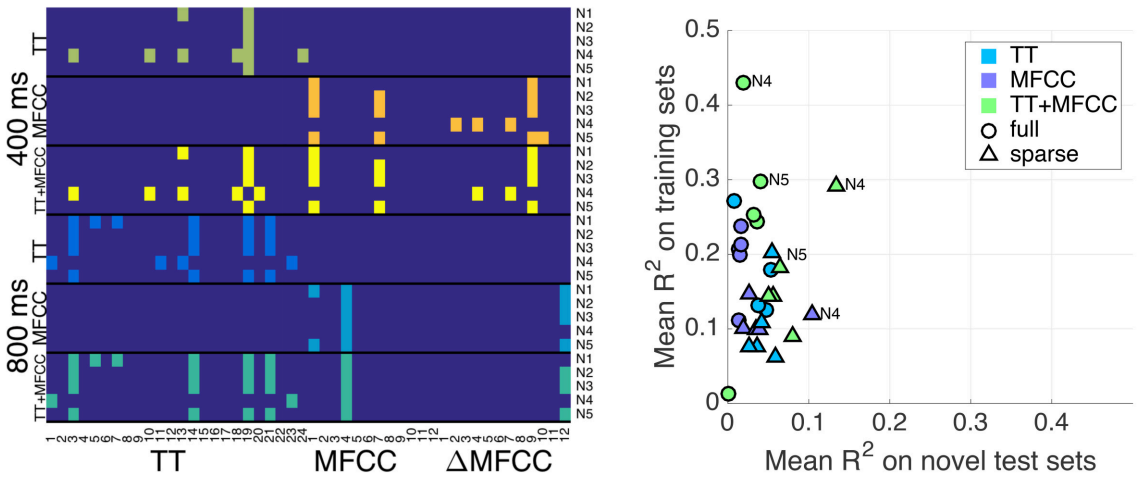


FIGURE 4 | (Left panel) Predictors selected for the sparse models. Pixels in colors other than dark blue refer to selected variables. The x-axis shows features, namely TT descriptors (1–24, see Table 2 in Supplementary Materials), and the 12 MFCCs and 12 ΔMFCCs. The y-axis concatenates all conditions including the 400 ms and 800 ms clips and the feature sets TT, MFCC, and TT+MFCC. Normalizations N1–N5 are annotated on the right. **(Right panel)** Overview of results from the solely acoustic models. Plot shows the mean fit to the training set (y-axis) and the mean fit on novel sets (x-axis). Circles correspond to the full models, triangles to sparse sets of features. A few normalizations are annotated.

TABLE 2 | Performance of sparse models.

			TEST									
			N1		N2		N3		N4		N5	
			Raw		Range		z-scores		*r*-test		r-corpus	
			I	II	I	II	I	II	I	II	I	II
TRAIN	Sparse TT	I	–	0.11	–	0.06	–	–	0.24	–	–	0.07
		II	–	0.11	–	0.14	–	0.14	0.10	0.16	–	0.20
		Mean		0.05		0.05		0.03		**0.12**		0.07
	Sparse MFCC	I	0.08	–	0.09	0.06	0.09	0.06	0.12	0.12	0.11	–
		II	–	0.12	–	0.11	–	0.11	0.09	0.12	–	0.18
		Mean		0.05		0.06		0.07		**0.11**		0.07
	Sparse TT+MFCC	I	0.07	0.11	0.12	0.07	0.12	0.06	0.34	0.16	0.13	0.08
		II	–	0.11	–	0.16	–	0.16	0.11	0.24	–	0.24
		Mean		0.07		0.09		0.09		**0.21**		0.11

R^2 coefficients are shown for the five normalizations (N1–N5) and three feature sets, each evaluated in the two training and testing conditions from 400 (I) and 800 ms (II) clips. Correlations with $p > 0.01$($R^2 < 0.06$) are not displayed for the sake of clarity. Condition means are concatenated below. Best average performance per feature set is given in bold font.

Moreover, the Genre+Date model here yielded better R^2 values in generalization than the model that relies on both acoustic and meta variables. The latter, indeed surprising finding could be taken as evidence for that listeners from Sample I and II used different weightings of acoustic information in their responses, potentially due to the different lengths of excerpts.

4.4. Role of Individual Features

By virtue of the parsimony of the sparse models, it is possible to take a more detailed look at the individual weightings of predictor variables. Here, we consider the exemplary case of the TT+MFCC (+Genre+Date) models with test-set ranking (N4). **Figure 5** shows the (standardized) regression coefficients β, which reflect the relative importance of the individual predictors for the prediction of similarity ($y = X\beta + F$).

For the models that included both acoustic and meta descriptors, the plots indicate that the genre descriptor was the most heavily weighted variable for both stimulus sets, and the date of release showed a by far smaller influence. Note that we found a very similar relation between the effect sizes of both variables for the model solely using genre and date information (the coefficients of which are not shown here). Regarding acoustic descriptors, the selections for both stimulus sets represent both spectral and spectrotemporal information: the spectral

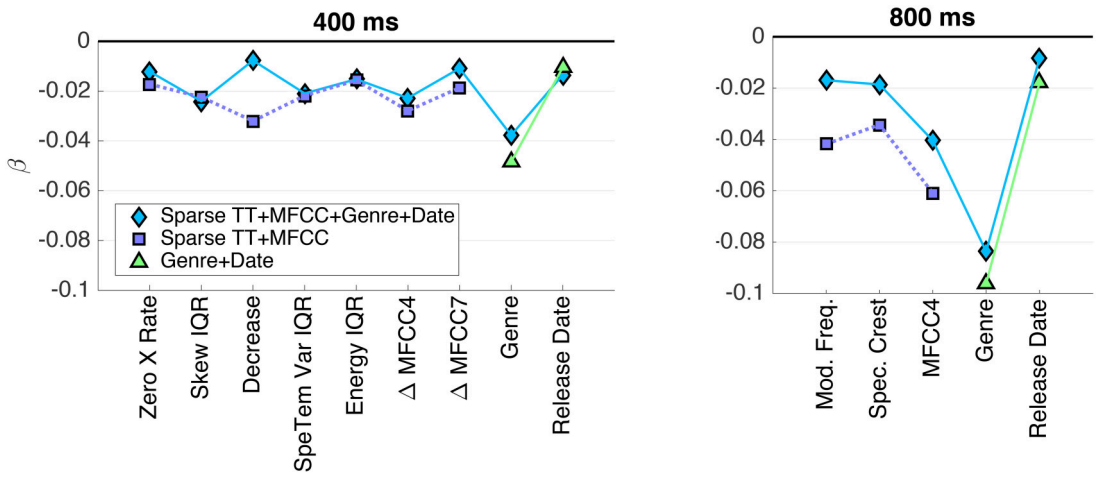


FIGURE 5 | PLSR coefficients β for the sparse model with test-rank normalization (N4), coefficients of the same model including genre and release date, as well as genre and release only. (Left panel) 400 ms clips; **(right panel)** 800 ms clips.

envelope distribution is represented by features such as Crest, Decrease, and MFCCs, whereas (spectro-)temporal modulations are represented by Modulation Frequency, Spectrotemporal Variation, and Δ MFCCs. Specifically, MFCC no. 4 was by a large margin the most important acoustic feature for the 800 ms clips, whereas Δ MFCC no. 4 was among the most important ones for the 400 ms clips.

From a more general stance, the presented evaluation, using five normalization conditions and three acoustic feature sets, indicates that one should not overestimate the universality of distinct acoustic features. In fact, the best model configuration did not share any features across the two stimulus sets. A plausible hypothesis could be that the duration of clips plays a pertinent role in the ways in which listeners compile and weight acoustic information from short music clips.

5. DISCUSSION

5.1. Summary

The main aim of this study was the development of the first audio-feature-based model for the prediction of human sound similarity judgements of short audio excerpts. We used partial least-squares regression in order to map from acoustic to perceptual similarity. Before entering the regression model, acoustic dissimilarities were normalized by using five schemes: (N1) raw feature values, (N2) range-normalization, (N3) z-scores, (N4) rank-transformation according to test set, and (N5) rank transformation according to a corpus. We then followed an exhaustive combinatorial approach that combined these five normalization schemes with two important candidate feature sets, the Timbre Toolbox (Peeters et al., 2011) and MFCCs, each of which contributed with 24 audio features. Importantly, each candidate model was assessed on the dataset it was fitted to, as well as on a set of novel audio excerpts. Our results indicate that combining both feature sets resulted in the most powerful model, in particular when being used with a test-set based rank transformation (N4). And even the sparse models with their drastically reduced numbers of features generally contained members from both features sets. This speaks for the complementary nature of Timbre Toolbox descriptors and MFCCs when it comes to the description of the similarity of music clips.

In line with the well-documented behavior of sparse models in terms of better generalization (e.g., Friedman et al., 2009, Ch. 16.2.2), we also found a trade-off between model performance on the training set vs. the models' enhanced ability to generalize to a new dataset. The best performing sparse model achieves an R^2 of up to 0.34 when evaluated on the dataset is was trained on and an R^2 of up to 0.16 when evaluated on a new dataset. This result for the first time provides evidence that a significant portion of the variance in the similarity perception of short music clips can be explained by acoustic features related to timbre. The fact that including only two variables encoding meta-information, and most importantly musical genre, substantially increased model performance (up to R^2 values of 0.52) suggests that the models based on acoustic features do not capture all information that participants are able to extract from the short audio clips and use for the similarity grouping. This finding also implies that great care should be taken in order to control for the effects of variables such as genre in future studies of sound and music similarity.

This last result is analogous to the importance of categorical information in the timbral dissimilarity of isolated instrument tones reported by Siedenburg et al. (2016b), where the addition of sound-source and instrument-family-related variables to a model based on acoustic features significantly improved the prediction of dissimilarity ratings (also see, Lemaitre et al., 2010). In this respect, the current results suggest that even if instructed to focus on low level auditory features (i.e., "the sound"), participants' responses are affected by higher level concepts such as genre. Although differences in timbral qualities, here measured by continuously varying audio features, likely constitute an important part of genre, genre categories might also be inferred from higher-level stylistic musical features such as rudimentary rhythmic or pitch-related information that are still

discernible in some of the clips. The current results then suggest that the inference of higher level concepts such as emotion or genre from short audio clips is based on more than timbral qualities, but rather on a complex mixture of acoustical, musical, and categorical (or higher-level) types of features. Notably, the exact weightings of these variables may vary with the duration of the excerpts. From the opposite perspective, the modeling infrastructure built up here could of course be applied to exploring the acoustic features utilized by humans in explicit genre identification tasks.

5.2. Limitations of the Current Study and Future Perspectives

This study represents the first rigorous attempt to build quantitative models that describe the perception of short audio excerpts based on audio feature extraction. Whereas, we have achieved encouraging accuracies on the training data, there is clearly room for improvement in future studies, in particular when it comes to generalization performance. A limitation of the current study is the fact that the two datasets differed in terms of the length of the excerpts (400 vs. 800 ms) whereas the modeling approach assumed that the same features are equally suitable for clips of both lengths. But this assumption might not be necessarily true. Hence, a future replication of this study should include different datasets with clips of the same lengths. Potentially, this might also help to achieve better generalization results. Specifically, it would be necessary to confirm the performance accuracy of the model with the best generalization performance (i.e., the sparse version of the TT+MFCC feature set using test-set-based rank normalization plus meta information) on a completely new dataset. New audio excerpts could be selected from a corpus according to their similarity predicted by the model, allowing us to generate precise hypotheses about the number of times the new excerpts are grouped together in the grouping paradigm. Using fully randomized approaches for determining the clips' starting points in the song, as proposed by Thiesen et al. (2016), would likely add further robustness to the experimental design.

It is also worth noting that the similarity data used in this study were derived from a grouping paradigm that required participants to make categorical decisions and it is unclear whether this specific paradigm introduced any sort of bias into the data. However, several other experimental paradigms can be used to obtain similarity data from participants and might be employed in future studies alongside the grouping paradigm (Giordano et al., 2011). These include pairwise similarity ratings on fine-grained scales, rankings of clips in relation to an anchor stimulus, triadic comparisons (Allan et al., 2007) or similarity comparisons of two pairs of clips.

In order to capture relevant additional information contained in the audio clips beyond timbral features, future investigations could also include mid-level features that describe aspects of the rhythmic, harmomic and pitch patterns (e.g., Müller, 2015). However, a systematic study of the participants' strategies used for arriving at perceptual categorizations of short audio clips would be a most helpful starting point for selecting features for subsequent modeling. The *thinking-aloud method* (Kuusela and Paul, 2000) commonly used in HCI research and other areas could be highly instrumental here to obtain qualitative insights into the cognitive processes employed when perceiving short audio clips.

Eventually, a reliable computational model of the perceptual similarity of short audio clips can serve as the basis for a refined individual differences test that assesses the ability to make fine-grained distinction between short musical excerpts. A computational model is necessary in order to create a test that is adaptive and homes in on the individual participant's ability level for judging sound similarities. In the case of the grouping paradigm, the computational model would be used for automatically selecting sets of clips that are easy vs. difficult to group, i.e., that differ in their within/between-group similarity. But the scientific value of a test that tracks and predicts an individual's ability to make similarity judgements lies not only in potential use as a new testing tool. Significant additional value comes from the cognitive insights gained from applying music information retrieval techniques to model complex perceptual processes.

AUTHOR CONTRIBUTIONS

Both authors KS and DM contributed equally to the design of the data modeling and writing of the manuscript. KS was mainly responsible for the acoustic feature models and data analysis. DM was responsible for data collection at Goldsmiths, University of London.

ACKNOWLEDGMENTS

We thank Bruno Gingras and Jason Musil for their work on data collection and pre-processing the experimental data. We also thank the BBC LabUK for help with recruitment and data collection of the Sample I data. We also wish to thank the two reviewers for helpful comments and advice.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2017.00639/full#supplementary-material

REFERENCES

Allan, H., Müllensiefen, D., and Wiggins, G. A. (2007). "Methodological considerations in studies of musical similarity," in *Proceedings of the 13th International Society for Music Information Retrieval Conference* (Vienna), 473–478.

Alluri, V., and Toiviainen, P. (2010). Exploring perceptual and acoustical correlates of polyphonic timbre. *Music Percept.* 27, 223–241. doi: 10.1525/mp.2010.27.3.223

Alluri, V., Toiviainen, P., Jääskeläinen, I., Glerean, E., Sams, M., and Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm.

NeuroImage 59, 3677–3689. doi: 10.1016/j.neuroimage.2011.11.019

Andén, J., and Mallat, S. (2011). "Multiscale scattering for audio classification," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, (Miami, FL), 657–662.

Bigand, E., Delbé, C., Gérard, Y., and Tillmann, B. (2011). Categorization of extremely brief auditory stimuli: domain-specific or domain-general processes? *PLoS ONE* 6:e27024. doi: 10.1371/journal.pone.0027024

Craft, A. J. D., Wiggins, G. A., and Crawford, T. (2007). "T.: How many beans make five? The consensus problem in music-genre classification and a new evaluation method for single-genre categorisation systems," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR07)* (Vienna).

De Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometri. Intel. Lab. Syst.* 18, 251–263. doi: 10.1016/0169-7439(93)85002-X

Efron, B., and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press.

Eronen, A. (2001). "Comparison of features for musical instrument recognition," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY* (Piscataway, NJ: IEEE), 19–22.

Friedman, J., Hastie, T., and Tibshirani, R. (2009). *The Elements of Statistical Learning*, 2nd Edn. Heidelberg: Springer.

Geladi, P., and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Anal. Chim. Acta* 185, 1–17. doi: 10.1016/0003-2670(86)80028-9

Gingras, B., Lagrandeur-Ponce, T., Giordano, B. L., and McAdams, S. (2011). Perceiving musical individuality: performer identification is dependent on performer expertise and expressiveness, but not on listener expertise. *Perception* 40, 1206–1220. doi: 10.1068/p6891

Giordano, B. L., Guastavino, C., Murphy, E., Ogg, M., Smith, B. K., and McAdams, S. (2011). Comparison of methods for collecting and modeling dissimilarity data: applications to complex sound stimuli. *Multivar. Behav. Res.* 46, 779–811. doi: 10.1080/00273171.2011.606748

Giordano, B. L., McDonnell, J., and McAdams, S. (2010). Hearing living symbols and nonliving icons: category specificities in the cognitive processing of environmental sounds. *Brain Cogn.* 73, 7–19. doi: 10.1016/j.bandc.2010.01.005

Gjerdingen, R. O., and Perrott, D. (2008). Scanning the dial: the rapid recognition of music genres. *J. New Music Res.* 37, 93–100. doi: 10.1080/09298210802479268

Homburg, H., Mierswa, I., Möller, B., Morik, K., and Wurst, M. (2005). "A benchmark dataset for audio classification and clustering," in *Proceedings of the 6th International Society for Music Information Retrieval Conference* (London), 528–531.

Joder, C., Essid, S., and Richard, G. (2009). Temporal integration for audio classification with application to musical instrument classification. *IEEE Trans. Audio Speech Lang. Proces.* 17, 174–186. doi: 10.1109/TASL.2008.2007613

Krumhansl, C. L. (2010). Plink: "thin slices" of music. *Music Percept.* 27, 337–354. doi: 10.1525/mp.2010.27.5.337

Kuusela, H., and Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *Am. J. Psychol.* 113, 387–404. doi: 10.2307/1423365

Lartillot, O., and Toiviainen, P. (2007). "A Matlab toolbox for musical feature extraction from audio," in *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx)* (Bordeaux), 237–244.

Laurier, C., Lartillot, O., Eerola, T., and Toiviainen, P. (2009). "Exploring relationships between audio features and emotion in music," in *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM 2009)* (Jyväskylä) 260–264.

Lemaitre, G., Houix, O., Misdariis, N., and Susini, P. (2010). Listener expertise and sound identification influence the categorization of environmental sounds. *J. Exp. Psychol. Appl.* 16, 16–32. doi: 10.1037/a0018762

Mace, S. T., Wagoner, C. L., Hodges, D., and Teachout, D. J. (2011). Genre identification of very brief musical excerpts. *Psychol. Music* 40, 112–128. doi: 10.1177/0305735610391347

McAdams, S. (2013). "Musical timbre perception," in *The Psychology of Music*, ed D. Deutsch (San Diego, CA: Academic Press), 35–67.

McAdams, S., Douglas, C., and Vempala, N. N. (2017). Perception and modeling of affective qualities of musical instrument sounds across pitch registers. *Front. Psychol.* 8:153. doi: 10.3389/fpsyg.2017.00153

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. *Psychol. Res.* 58, 177–192. doi: 10.1007/BF00419633

Mehmood, T., Liland, K. H., Snipen, L., and Sæbø, S. (2012). A review of variable selection methods in partial least squares regression. *Chemometr. Intel. Lab. Syst.* 118, 62–69. doi: 10.1016/j.chemolab.2012.07.010

Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PLoS ONE* 9:e89642. doi: 10.1371/journal.pone.0089642

Müllensiefen, D., Harrison, P., Caprini, F., and Fancourt, A. (2015). Investigating the importance of self-theories of intelligence and musicality for students' academic and musical achievement. *Front. Psychol.* 6:1702. doi: 10.3389/fpsyg.2015.01702

Müller, M. (2015). *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Heidelberg: Springer.

Musil, J., El-Nusairi, B., and Müllensiefen, D. (2013). "Perceptual dimensions of short audio clips and corresponding timbre features," in *From Sounds to Music and Emotions. Lecture Notes in Computer Science*, Vol. 7900, eds M. Aramaki, M. Barthet, R. Kronland-Martinet, and S. Ystad (Berlin: Springer), 214–227.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (2011). The Timbre Toolbox: Extracting audio descriptors from musical signals. *J. Acoust. Soc. Am.* 130, 2902–2916. doi: 10.1121/1.3642604

Plazak, J., and Huron, D. (2011). The first three seconds listener knowledge gained from brief musical excerpts. *Musicae Sci.* 15, 29–44. doi: 10.1177/1029864910391455

Rentfrow, P. J., Goldberg, L. R., and Levitin, D. J. (2011). The structure of musical preferences: a five-factor model. *J. Personal. Soc. Psychol.* 100, 1139. doi: 10.1037/a0022406

Rentfrow, P. J., and Gosling, S. D. (2003). The do re mi's of everyday life: the structure and personality correlates of music preferences. *J. Personal. Soc. Psychol.* 84, 1236–1256. doi: 10.1037/0022-3514.84.6.1236

Schellenberg, E. G., Iverson, P., and McKinnon, M. C. (1999). Name that tune: identifying popular recordings from brief excerpts. *Psychon. Bull. Rev.* 6, 641–646. doi: 10.3758/BF03212973

Siedenburg, K., Fujinaga, I., and McAdams, S. (2016a). A comparison of approaches to timbre descriptors in music information retrieval and music psychology. *J. New Music Res.* 45, 27–41. doi: 10.1080/09298215.2015.1132737

Siedenburg, K., Jones-Mollerup, K., and McAdams, S. (2016b). Acoustic and categorical dissimilarity of musical timbre: evidence from asymmetries between acoustic and chimeric sounds. *Front. Psychol.* 6:1977. doi: 10.3389/fpsyg.2015.01977

Suied, C., Agus, T. R., Thorpe, S. J., Mesgarani, N., and Pressnitzer, D. (2014). Auditory gist: recognition of very short sounds from timbre cues. *J. Acoust. Soc. Am.* 135, 1380–1391. doi: 10.1121/1.4863659

Thiesen, F. C., Kopiez, R., Reuter, C., Czedik-Eysenberg, I., and Schlemmer, K. (2016). "In the blink of an ear: a critical review of very short musical elements," in *Proceedings of the 14th International Conference on Music Perception and Cognition* (San Francisco, CA), 147–150.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometr. Intel. Lab. Syst.* 58, 109–130. doi: 10.1016/S0169-7439(01)00155-1

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 13 April 2017
doi: 10.3389/fpsyg.2017.00587

Perceptually Salient Regions of the Modulation Power Spectrum for Musical Instrument Identification

*Etienne Thoret *, Philippe Depalle and Stephen McAdams*

Schulich School of Music, McGill University, Montreal, QC, Canada

The ability of a listener to recognize sound sources, and in particular musical instruments from the sounds they produce, raises the question of determining the acoustical information used to achieve such a task. It is now well known that the shapes of the temporal and spectral envelopes are crucial to the recognition of a musical instrument. More recently, Modulation Power Spectra (MPS) have been shown to be a representation that potentially explains the perception of musical instrument sounds. Nevertheless, the question of which specific regions of this representation characterize a musical instrument is still open. An identification task was applied to two subsets of musical instruments: tuba, trombone, cello, saxophone, and clarinet on the one hand, and marimba, vibraphone, guitar, harp, and viola pizzicato on the other. The sounds were processed with filtered spectrotemporal modulations with 2D Gaussian windows. The most relevant regions of this representation for instrument identification were determined for each instrument and reveal the regions essential for their identification. The method used here is based on a "molecular approach," the so-called bubbles method. Globally, the instruments were correctly identified and the lower values of spectrotemporal modulations are the most important regions of the MPS for recognizing instruments. Interestingly, instruments that were confused with each other led to non-overlapping regions and were confused when they were filtered in the most salient region of the other instrument. These results suggest that musical instrument timbres are characterized by specific spectrotemporal modulations, information which could contribute to music information retrieval tasks such as automatic source recognition.

Keywords: spectrotemporal modulation, musical timbre, Instrument identification, Modulation power spectrum, Bubble method

OPEN ACCESS

Edited by:
Frank A. Russo,
Ryerson University, Canada

Reviewed by:
Michael David Hall,
James Madison University, USA
Christoph Reuter,
University of Vienna, Austria

****Correspondence:***
Etienne Thoret
etienne.thoret@mcgill.ca

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *21 October 2016*
Accepted: *29 March 2017*
Published: *13 April 2017*

Citation:
Thoret E, Depalle P and McAdams S (2017) Perceptually Salient Regions of the Modulation Power Spectrum for Musical Instrument Identification. Front. Psychol. 8:587. doi: 10.3389/fpsyg.2017.00587

INTRODUCTION

Automatic musical instrument recognition is one of the more complex problems in musical informatics research. Work on how humans do this could provide important insights concerning how to get machines to do it, as well to improve automatic annotation algorithms, for example. Listeners' ability to recognize musical instruments has animated research for many years. From several points of view, either purely computational (Brown, 1999; Brown et al., 2001) or purely perceptual (McAdams, 1993, 2013), it has been shown that the acoustic signal encompasses many indices specific to each instrument, which contribute to their recognition. In order to understand what information is essential for algorithms or for perceptual recognition processes, mathematical representations of sound signals have been developed. In a discussion of the relation

between Music Information Retrieval (MIR) issues and music cognition issues, Aucouturier and Bigand (2013) stressed the importance of investigating and developing biologically inspired representations to better understand what signal information is relevant in MIR tasks (see also Siedenburg et al., 2016), and reciprocally, how MIR algorithms may help to better understand the processing underpinning perceptual tasks.

The simplest representation of a sound is its waveform, which corresponds to the sound pressure recorded by a microphone or the vibration that moves the tympanic membrane. This first type of representation leads to timbre descriptors that are relevant either from a computational point of view or that have been shown to significantly contribute to perceptual dissimilarity judgments. For instance, attack time has been shown to be a strong perceptual cue to distinguish sustained and impulsively excited instruments (Iverson and Krumhansl, 1993; McAdams et al., 1995), and has also been shown to be a relevant feature for instrument classification (Saldanha and Corso, 1964). Nevertheless, this representation doesn't reveal many of aspects of a sound, in particular its spectral content. In order to reveal the evolution of the spectral content over time, spectrograms of sounds have been used for some time (Koenig et al., 1946). Interestingly, this representation can be related to the transformation of mechanical waves into neural signals achieved at the cochlear level. Many sound descriptors have been derived from this kind of representation. One of the most well-known is certainly the average spectral centroid over the duration of a sound, which has been shown to correlate well with perceptual dimensions (e.g., Grey and Gordon, 1978; McAdams et al., 1995; Giordano and McAdams, 2010; Hjortkjær and McAdams, 2016).

Many experiments using identification, discrimination or dissimilarity-rating tasks have investigated the specific influence of temporal and spectral cues on timbre perception. Hall and Beauchamp (2009), for example, have shown in identification and discrimination tasks that listeners are more sensitive to the spectral envelope of musical instrument sounds than to the temporal envelope, and they are more sensitive to spectral envelope shape than to the spectral centroid *per se*. In a meta-analysis of 23 datasets from 17 published studies, Giordano and McAdams (2010) showed that confusions in identification tasks are related to perceived similarity between the same instruments. These experiments have stressed that perceptual results can be explained to a certain extent by audio descriptors computed from spectral and spectrotemporal descriptors that are plausibly used by the auditory system to identify a sound source such as a musical instrument.

Recent studies have emphasized the interest of another kind of representation, the Modulation Power Spectrum (MPS) (Elliott and Theunissen, 2009; Elliott et al., 2013). Basically, the MPS corresponds to the two-dimensional Fourier transform of a spectrogram and can be seen as a representation characterizing its temporal and spectral periodicities. This representation highlights the temporal and spectral regularities of a spectrogram. For musical sounds with tremolo (regular amplitude modulation) for example, the MPS will be composed of a local maximum at the tremolo frequency. Similarly, if the musical sound is perfectly harmonic, the MPS will be composed of a local maximum in the spectral modulation dimension. Interestingly, as with the waveform or the spectrogram, this representation can be associated with a processing stage in the auditory system. Indeed, some neuron populations in primary auditory cortex seem to respond selectively to specific spectrotemporal modulations, at least in the ferret (Shamma, 2001). The prominent role of these spectrotemporal modulations in the perception and classification of musical timbre has been suggested recently (Patil et al., 2012; Elliott et al., 2013; Hemery and Aucouturier, 2015; Patil and Elhilali, 2015). In particular, Patil et al. (2012) have shown that this kind of representation can be used in the automatic classification of musical instruments, but it also correlates with perceptual dissimilarity ratings between instruments. Nevertheless, it remains unknown whether specific aspects of spectrotemporal modulations are relevant for the recognition of musical instruments. If some ranges of spectrotemporal modulation are more relevant than others to recognize and identify musical instruments, this would shed light on a possible strategy used by auditory processes to identify specific sound sources such as musical instruments. From a purely computational point of view, this approach would enable us to envisage new timbre descriptors related to musical instruments in addition to those derived from temporal and time-frequency representations (Peeters et al., 2011). Note that these potential timbre descriptors based on the MPS representation should also be linked to the timbre descriptors defined on time-frequency representations. As the spectral modulations are a kind of decomposition of the spectral envelope, MPS-based timbre descriptors should be linked to descriptors such as the formants, the spectral centroid, higher-order statistical moments or mel-frequency cepstral coefficients. For more detail concerning audio descriptors related to timbre perception, see Pachet and Aucouturier (2004), Peeters et al. (2011), and Elliott et al. (2013).

Here we tackle these questions for sustained (blown and bowed) instruments (tuba, trombone, saxophone, clarinet, cello) and instruments producing impulsive (plucked and struck) sounds (viola pizzicato, guitar, harp, vibraphone, marimba). We aimed to determine which region of the MPS leads to the identification of these musical instruments. Based on a filtering method proposed by Elliott and Theunissen (2009) and on a "molecular" approach, the so-called "bubbles" method, proposed by Gosselin and Schyns (2001), we set up an identification task in which listeners had to recognize processed versions of original sounds composed from a small region of their MPS. This allows us to determine the relevance of the location of each bubble, i.e., corresponding to a 2D Gaussian window, of the MPS in the recognition of musical instrument sounds and then, by combining the responses for bubble regions, to compute a global mask that highlights the most salient MPS regions for each instrument and for all instruments combined. This approach allows us to identify the most salient regions of the MPS for instrument identification, and moreover, if instruments are confused with each other, to determine which regions of the MPS lead to the specific confusions. The bubble method was initially developed to identify which part of a face is used by the visual system to determine gender and whether the face was expressive

or not. Participants were asked to identify gender or categorize it as expressive or not from small parts of the face. A similar method has recently proved its efficacy in identifying which regions of the MPS are relevant for speech intelligibility (Venezia et al., 2016).

THE MODULATION POWER SPECTRUM OF MUSICAL SOUNDS

The MPS is defined here as the two-dimensional Fourier transform of the time-frequency representation (TFR) of a sound signal (Singh and Theunissen, 2003; Elliott and Theunissen, 2009). More specifically, the TFR $X(t,f)$ itself is defined here as the amplitude of the Fourier transform obtained with a Gaussian window and is commonly known as the magnitude of the Short-Term Fourier Transform (STFT) or the Gabor Transform. The MPS is the amplitude of the successive Fourier transforms along the STFT temporal and frequency axes. This MPS representation is composed of two dimensions: temporal modulations (in Hz) and spectral modulations (in cycles/Hz), see **Figure 1**.

The resolution of the MPS, denoted $MPS(s,r)$ with s and r being spectral and temporal modulations, respectively, is constrained by the resolution of the time-frequency representation $X\left(t,f\right)$, mainly characterized by the effective sizes of the temporal Gaussian windows and the overlap between two successive windows. They indeed define the upper and lower boundaries of the spectral and temporal modulations axes. Constrained by the uncertainty principle $\sigma_t \geq \frac{1}{2\pi\sigma_f}$ where σ_t and σ_f correspond to the uncertainties along the temporal and spectral modulation dimensions, respectively, we here choose $\sigma_t = 11.61$ ms and $\sigma_f = 21.53$ Hz leading to upper boundaries of 43 Hz and 23.22 cycles/Hz which correspond to values relevant for the auditory perception of sounds such as speech (Elliott and Theunissen, 2009).

EXPERIMENTS 1 AND 2

Materials and Methods

Participants

Thirty-one participants (12 females) with ages between 19 and 45 ($M = 24.4$, $SD = 5.7$) took part in the first experiment and 32 participants (14 females) with ages between 18 and 45 ($M = 24.2$, $SD = 5.7$) took part in the second experiment. All participants were musicians who had completed at least second-year university-level musical training in performance, composition or theory. Seventeen of the participants took part in both experiments (5 females). Participants provided informed consent, had normal hearing, and were compensated for their time.

Stimuli

The stimuli were five arpeggios generated from samples of the Vienna Symphonic Library. In the first experiment, five sustained instruments (trombone, tuba, saxophone, cello, and clarinet) playing three musical pitches: F#3 (with a fundamental frequency of 185.0 Hz), C4 (261.6 Hz), and F#4 (370.0 Hz) were chosen. This range of pitches doesn't involve large variations of timbre across the three different notes. In the second experiment, five impulsive instruments were chosen (vibraphone, marimba, harp, guitar, viola pizzicato) playing the same pitches. Based on other work in the lab (McAdams et al., 2016), we chose to separate sustained instruments from impulsive instruments as it would have been too obvious to distinguish them in an identification task. For each instrument, the three notes were equalized in loudness in a preliminary experiment. Their durations were all cut to 0.5 s with a 50-ms raised cosine fade-out amplitude envelope to avoid discrimination based on duration. The attack was preserved. Finally, arpeggios were generated by concatenating the three notes from the lowest to the highest.

In order to determine which regions of the MPS lead to the identification of musical instruments, we employed a technique for filtering instrumental sounds in the spectrotemporal modulation domain (see **Figure 1**). With this technique, a sound is processed by keeping only a small region of its MPS, this filtered version is reconstructed, and then whether the information that remains is relevant for the identification of the initial instrument is evaluated with listener testing. Hence, the MPS is first multiplied by a "bubble," a two-dimensional Gaussian MPS-filter frequency response $G_{(\mu_s,\sigma_s),(\mu_r,\sigma_r)}(s,r)$ where μ_s, μ_r and σ_s, σ_r are the means and standard deviations in the scale and rate dimensions, respectively:

$$G_{(\mu_s,\sigma_s),(\mu_r,\sigma_r)}(s,r) = \exp\left(-\frac{1}{2}\left(\frac{s-\mu_s}{\sigma_s}\right)^2\right) \exp\left(-\frac{1}{2}\left(\frac{r-\mu_r}{\sigma_r}\right)^2\right) \quad (1)$$

It must be noted that the MPS and the filter G are composed of four quadrants with positive and negative spectral and temporal modulations. For the sake of simplicity and as the filter is perfectly symmetric in amplitude and phase in the spectral and temporal modulation dimensions, only positive values are presented in what follows. The MPS-filtered TFR $Y(t,f)$ can then be easily reconstructed by a 2D inverse Fourier transform of the processed MPS: $MPS(s,r) \cdot G_{(\mu_s,\sigma_s),(\mu_r,\sigma_r)}(s,r)$. Note that $Y(t,f)$ is magnitude only, lacks the phase, and thus does not allow for perfect reconstruction of the waveform directly from standard reconstruction technique such as the overlap-add method (OLA; Rabiner and Schafer, 1978). Therefore, we instead used Griffin and Lim's (1984) algorithm in a MATLAB implementation provided by Slaney (1994) in order to iteratively build a signal, the STFT magnitude of which is as close as possible to the $Y(t,f)$ in a quadratic sense. Twenty-five iterations lead to a correct reconstruction of the waveform for an acceptable computation time. **Figure 1** summarizes the whole analysis-filtering-synthesis process. Practically speaking, the quality of the reconstruction is evaluated by computing the averaged relative log-error ratio ϵ in percent between the desired spectrogram $Y(t,f)$ and the STFT

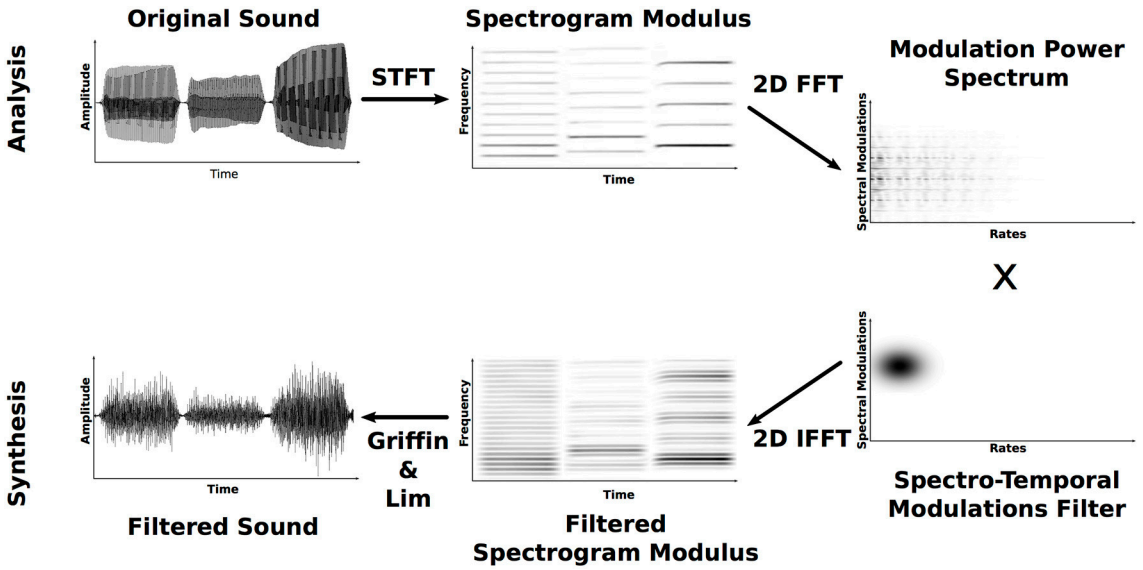


FIGURE 1 | Analysis–filtering–synthesis process. Only the top-right quadrant of the MPS and of the filter are represented.

magnitude of the reconstructed waveform $Y_b(t,f)$:

$$\epsilon = 100\frac{1}{N_f N_t}\sum_{t_i=1}^{N_t}\sum_{f_i=1}^{N_f}\left|\frac{\log\left(Y\left(t_i,f_i\right)\right)-\log(Y_b\left(t_i,f_i\right))}{\log(Y\left(t_i,f_i\right))}\right| \quad (2)$$

where N_f and N_t are the numbers of frequency and time bins, respectively.

The stimulus files were normalized at −3 dB relative to 16-bit amplitude resolution. In the first experiment, the peak level of the stimuli ranged from 58 to 71 dB SPL (A-weighted). In the second experiment, the peak level of the stimuli ranged from 63 to 70 dB SPL (A-weighted). Stimuli were classically sampled at 44,100 Hz with 16-bit resolution.

Apparatus

Both experiments took place in an IAC model 120act-3 double-walled audiometric booth (IAC Acoustics, Bronx, NY). Stimuli were presented over Sennheiser HD280Pro headphones (Sennheiser Electronics GmbH, Wedemark, Germany) using a Macintosh computer (Apple Computer, Inc., Cupertino, CA) with digital-to-analog conversion on a Grace Design m904 monitor system (Grace Digital Audio, San Diego, CA). The experimental interface was programmed in the Max7 audio software environment (Cycling '74, San Francisco, CA) and data collection was programmed in Matlab (The Mathworks, Inc., Natick, MA) interacting via the User Data Protocol (*udp*).

Procedure

Participants first completed a standard pure-tone audiogram to ensure normal hearing with hearing thresholds of 20 dB HL or better at octave-spaced frequencies in the range of 250–8,000 Hz (Martin and Champlin, 2000; ISO 389–8, 2004). The task was 5-Alternative Forced Choice (5-AFC). In each trial, the participants were asked to recognize the instrument that played the arpeggios among the five instruments. They were asked to answer as quickly as possible after hearing the sounds in order that they answer the most intuitively when the sounds were degraded by the filtering process. The experiment began with a training session of 15 trials (5 instruments × 3 repetitions) during which the participants performed the task with the original, unprocessed sounds. After having completed the training session, the participants began the main experiment, which was composed of 480 trials (5 instruments × 96 filters). For each instrument, the MPS was filtered with 96 Gaussian filters $G_{(\mu_s,\sigma_s),(\mu_r,\sigma_r)}$ with the following standard deviations: $\sigma_r = 5\ Hz$ and $\sigma_s = 4\ cycles/Hz$ overlapping by 75% along each dimension (12 rates and 8 spectral modulations, see **Figure 2**). These standard deviations were determined by empirical tests in order to provide a good trade-off between accurate sampling and a reasonable number of filters for sampling the MPS. The averaged log-error ratio (cf. Equation 2) for the 480 sounds equaled 10.25%. Hence in each trial, one of the five instrument arpeggios was processed with one filter, and the participant had to recognize the original instrument. The order of presentation of the 480 trials was randomized for each participant.

Data Analysis

For all participants and for all five instruments, a confusion matrix was computed and association scores were tested against chance level with a one-tailed *t*-test. The *p*-values were adjusted with Bonferroni corrections for multiple testing. The subsequent data analysis was inspired by the so-called "bubbles" method proposed by Gosselin and Schyns (2001). In each trial, if the sound was properly associated with the instrument, the MPS filter was added to a CorrectMask matrix. Across all trials, each MPS filter was added to a TotalMask matrix. For each participant, a ProportionMask was derived by dividing

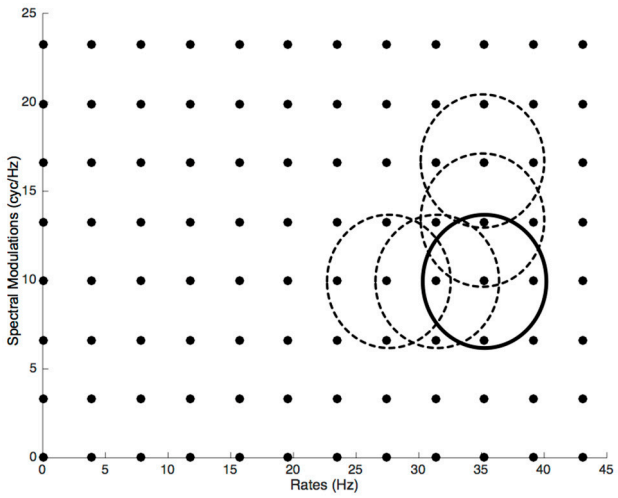


FIGURE 2 | Sampling of the Modulation Power Spectrum by 96 Gaussian filters in the scale-rate plane. The dots show the center value and the circles the standard deviation of the 2D Gaussian distribution.

CorrectMask by TotalMask. If no region had any special perceptual significance for recognition, ProportionMask would be homogeneous. To the contrary, if some regions were more important for recognition, they would have higher values than the other regions of the ProportionMask. Note that our method differs from that of Gosselin and Schyns (2001), which was initially used to determine the most salient parts of a face for gender and expressivity recognition. Although they used an adaptive method that adjusted the number of bubbles to converge on 75% correct recognition, here we only used single bubbles in order to determine their independent contribution to instrument identification. Given that MPS filters overlap each other, the resulting ProportionMasks represent the relative importance of each region of the MPS to the identification of that instrument. In order to determine which regions are the most relevant for the identification of each instrument, a one-tailed *t*-test between ProportionMask values and the averaged value of the ProportionMask ($\alpha = 0.05$) was applied for each instrument and across participants to compute a SalienceMask. Hence, the *p*-values of these tests were here used as a measure of the relevance of each spectrotemporal modulation value: the smaller the *p*-values, the more salient the spectrotemporal modulation. The statistical significance of each spectrotemporal modulation was also determined and corresponds to the DiagnosticMask of Gosselin and Schyns (2001). Here, we considered that a bin of the SalienceMask is significant when the *p*-value is lower than 0.05. The DiagnosticMask is a binary mask set to 1 or 0 when the SalienceMask is significant or not, respectively. The description of all of the masks described previously is summarized in **Table 1**.

In order to reveal the most salient spectrotemporal modulation regions, we first computed the SalienceMask for all instruments, and then for each instrument separately. In addition, when one instrument is significantly confused with another one, the same analysis is performed to generate a ConfusionMask by substituting the correctly associated mask in the CorrectMask with those from the instrument with which it has been confused. This mask reveals the spectrotemporal regions in which one instrument is incorrectly identified as another.

Results

Confusion Matrices

Tables 2, 3 present the averaged confusion matrices across participants from the two experiments. All instruments were recognized above chance in both experiments [$p < 0.001$– Trombone: $t_{(30)} = 12.84$, $d = 2.31$, Clarinet: $t_{(30)} = 16.28$, $d = 2.92$, Tuba: $t_{(30)} = 12.31$, $d = 2.21$, Cello: $t_{(30)} = 13.84$, $d = 2.48$, Saxophone: $t_{(30)} = 9.82$, $d = 1.76$ for Experiment 1, and $p < 0.001$—Viola Pizzicato: $t_{(31)} = 15.30$, $d = 2.70$, Guitar: $t_{(31)} = 8.02$, $d = 1.41$, Harp: $t_{(31)} = 11.49$, $d = 2.03$, Marimba: $t_{(31)} = 13.02$, $d = 2.30$, Vibraphone: $t_{(31)} = 10.57$, $d = 1.86$ for Experiment 2]. In addition, in Experiment 1, tuba, cello and saxophone were significantly confused with trombone [$t_{(30)} = 5.91$, $p < 0.001$, $d = 1.06$], saxophone [$t_{(30)} = 1.75$, $p < 0.05$, $d = 0.31$] and cello [$t_{(30)} = 3.84$, $p < 0.01$, $d = 0.69$], respectively. In the second experiment, the guitar, harp, marimba and vibraphone were significantly confused with harp [$t_{(31)} = 4.32$, $p < 0.001$, $d = 0.76$], guitar [$t_{(31)} = 3.69$, $p < 0.001$, $d = 0.65$], vibraphone [$t_{(31)} = 2.59$, $p < 0.01$, $d = 0.45$] and marimba [$t_{(31)} = 2.35$, $p < 0.05$, $d = 0.41$], respectively.

Perceptually Relevant Spectrotemporal Modulations

Figures 3, 4 present the SalienceMask for all instruments combined and for each instrument separately for Experiments 1 and 2. The yellowest regions of each plot are the most salient regions of the MPS. The *p*-values of the ProportionMasks are displayed. Concerning the sustained sounds and for all instruments combined (upper left plot of **Figure 3**), the most salient spectrotemporal modulations ranged from 0 to 30 Hz and from 0 to 18 cyc/Hz. The trombone, the clarinet and the cello also have their most relevant regions for low spectral and temporal modulations (**Figure 3**). The saxophone has its most salient region for temporal modulations comprised between 10 and 30 Hz. Concerning the tuba, the whole range of spectral modulations is relevant for its identification. For impulsive sounds and all instruments combined (upper left of **Figure 4**), the most salient spectrotemporal modulations ranged from 0 to 18 Hz and from 0 to 15 cyc/Hz. The harp and the vibraphone also have their most relevant regions for low spectral and temporal modulation. The viola pizzicato has its most salient MPS regions comprised between 10 and 30 Hz and 0 and 15 cyc/Hz. The marimba has its most salient regions for high rates (>15 Hz). The guitar has its most salient regions for high rates (>20 Hz) and high spectral modulations (>5 cyc/Hz). It is interesting to note that in both experiments, the most relevant spectrotemporal modulations for all instruments combined are centered on the same region, i.e., low spectral and temporal modulations.

If we consider the DiagnosticMask (see **Figure 5**), the most salient regions of the plane for all sustained instruments combined and all impulsive instruments combined represents

TABLE 1 | Summary of the different Masks computed for the analysis of the salient regions of the MPS for each instrument.

Mask	Description
CorrectMask	For one instrument, sum of the filters leading to correct identification.
TotalMask	Sum of all the filters.
ProportionMask	Ratio between the CorrectMask and the TotalMask.
SalienceMask	For each instrument, the *p*-value of a single-sample *t*-test against chance performance (0.2) of each bin of the CorrectMask.
ConfusionMask	Ratio between the sum of the filters leading to a wrong association of instrument A with instrument B and the sum of all filters.
DiagnosticMask	Binary mask associated with a SalienceMask or a ConfusionMask. For each bin, it equals 1 if the SalienceMask of ConfusionMask's bin is significant, i.e., $p < 0.05$, and equals 0 otherwise.

TABLE 2 | Confusion matrix in percent response averaged across participants for experiment 1 (sustained sounds).

	Trombone	Clarinet	Tuba	Cello	Saxophone
Trombone	**61*** **	2.6	20.6	3.5	12.3
Clarinet	3.6	**69.9*** **	7	9.6	9.9
Tuba	**34.8*** **	2.9	**54.5*** **	3.3	4.5
Cello	4.1	7.3	5.7	**59.6*** **	**23.3***
Saxophone	5	7.6	5.5	**30.9** **	**51*** **

*Association rates significantly above chance are shown in bold. ***p < 0.001; **p < 0.01; *p < 0.05.*

TABLE 3 | Confusion matrix in percent response averaged across participants for experiment 2 (impulsive sounds).

	Viola Pizz.	Guitar	Harp	Marimba	Vibraphone
Viola Pizz.	**69.8*** **	9.9	12.1	5.3	2.9
Guitar	9.1	**45.2*** **	**30.1*** **	7	8.6
Harp	16.6	**27.5*** **	**42.9*** **	7.3	5.7
Marimba	3	4.5	4	**61.9*** **	**26.5** **
Vibraphone	0.6	1.6	1.1	**30.1***	**66.6*** **

*Association rates significantly above chance are shown in bold. ***p < 0.001; **p < 0.01; *p < 0.05.*

38 and 22.9%, respectively. If we consider each instrument separately, the sustaining instrument that provides the largest salient area is the clarinet (45.5% of the MPS plane) followed by saxophone (38.9%), trombone (33.4%), cello (28.3%), and tuba (25.3%). The five impulsively excited instruments have salient areas of similar size, 27.4% for viola pizzicato, 29% for guitar, 24.7% for harp, 27.1% for marimba and 24.6% for vibraphone.

Interestingly, for instruments that were confused, the ConfusionMasks presented in **Figures 6**, **7** confirm that the salient regions of the SalienceMask lead to confusion when an instrument's MPS is filtered with spectrotemporal modulations in the most salient areas of the other instrument. For instance, the area leading to identifications of the cello stimulus as a saxophone corresponds to the most salient area of the saxophone and vice versa. The same phenomenon is observed for the marimba/vibraphone and harp/guitar pairs (see **Figure 7**) and to a certain extent for the trombone and the tuba (see **Figure 6**). These results confirm that these spectrotemporal areas are specific to the timbre of the confused instruments.

DISCUSSION

In this paper we sought to determine the most salient regions of the MPS for the identification of musical instruments producing either sustained or impulsive sounds. Based on the "bubbles" method developed by Gosselin and Schyns (2001), we have shown that globally the most salient spectrotemporal modulations are centered on low rates and low spectral modulations. Interestingly, when two instruments are confused, the spectrotemporal modulations enabling their discrimination do not overlap, suggesting that these regions are specific to these instruments. Moreover, note that confusions appear when the original sounds are filtered in the most salient regions of the instrument with which they are confused, reinforcing the idea that they are specific to the timbre of these instruments. Also, specific regions of the MPS other than the low spectral and temporal modulations are specific to some instruments, e.g., for the guitar. This does not concur with the general finding that globally low rates and low spectral modulations are relevant and suggests that when instruments were confused, listeners were focusing on a specific region of the MPS.

From a perceptual point of view, the fact that different regions of the MPS are more or less significant for the identification of different instruments suggests that these regions are specific to the timbre of these instruments. Counterintuitively, we could have thought that instruments sharing the same relevant region would be confused. However, the SalienceMasks reveal the region that allows for identification within the context of the sound set being tested. Two instruments can therefore have close SalienceMasks and even provide good recognition, suggesting that the SalienceMasks cannot be used as a measure of similarity between instruments. Conversely, when two instruments are confused, the fact that their salient spectrotemporal modulations don't overlap, and, even more, that their ConfusionMask falls within the region of the SalienceMasks of the other instrument, reinforces the idea that these two non-overlapping regions are specific to these instruments in this context. For example, according to these results, we can conclude that the SalienceMask of the saxophone corresponds to specific timbral properties of this instrument in comparison with those of the cello timbre with which it has been confused. Nevertheless, we suspect that if the cello had been removed from the instrument subset, the SalienceMasks of the saxophone would have been different. The same expectation would hold for the trombone/tuba, guitar/harp and marimba/vibraphone pairs as well.

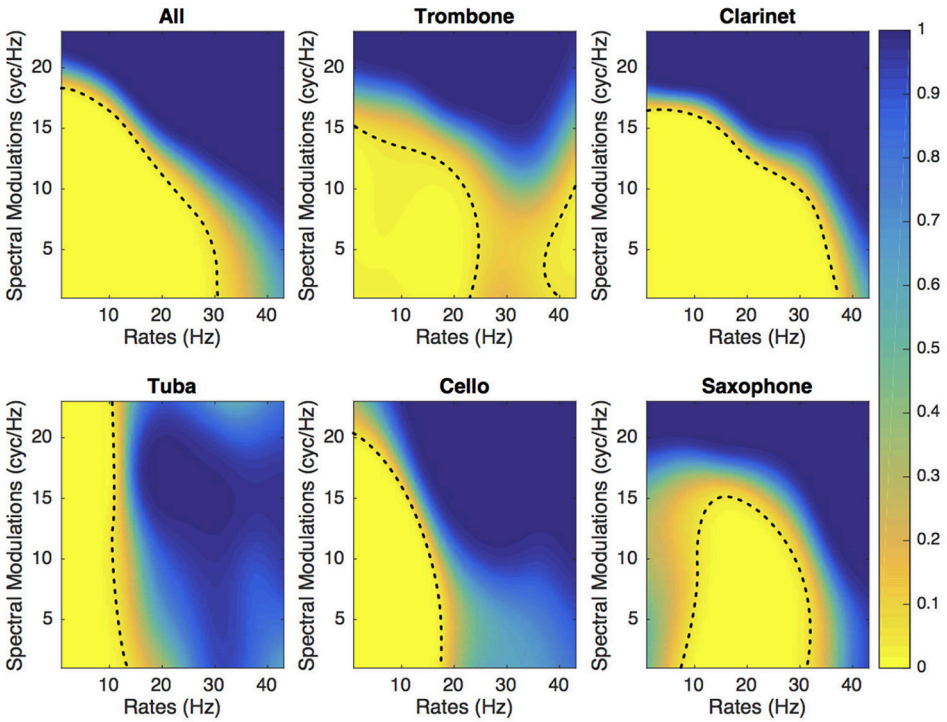


FIGURE 3 | Experiment 1. SalienceMask of the MPS for the five sustained instruments and all instruments combined. The dashed line represents the threshold at $p = 0.05$.

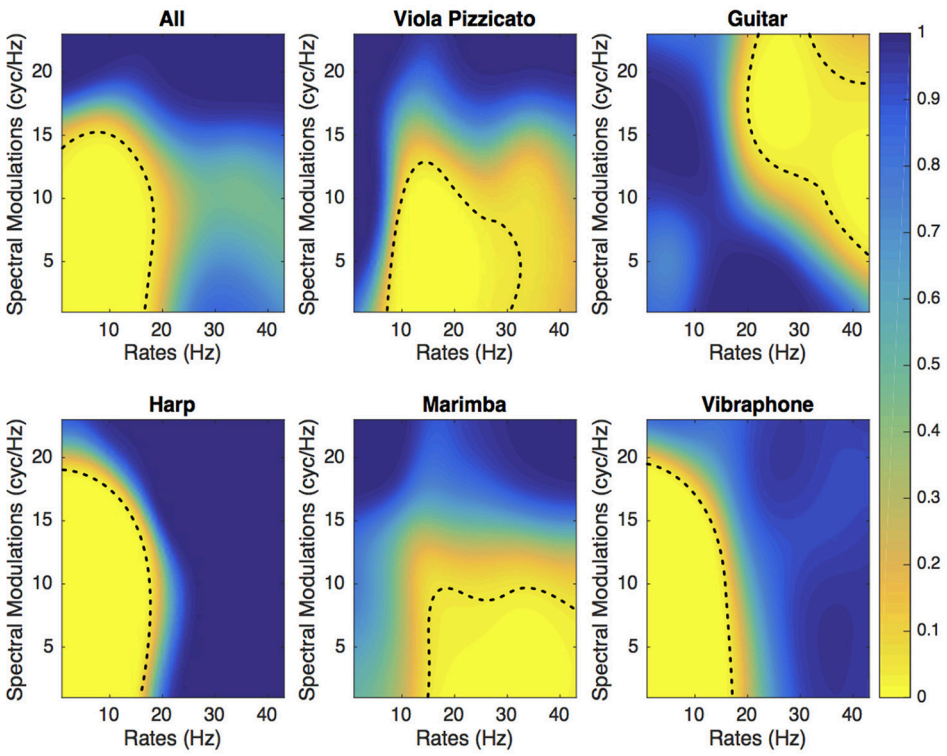


FIGURE 4 | Experiment 2. SalienceMask of the MPS of the five sustained instruments and all instruments combined. The dashed lines represent the thresholds at $p = 0.05$.

In order to fully validate that specific MPS regions are characteristic of some instruments, additional experimentation is needed. In particular, an identification experiment with the original sounds filtered by their SalienceMasks would evaluate whether it removes the confusions between the different instruments. From a cortical point of view, we may expect that this ability to focus on different regions of the MPS is possible due to the plasticity of the neurons in primary auditory cortex. Several studies have indeed revealed that neurons of this cortical network can reshape their sensitivity to different spectrotemporal modulations according to the needs of the tasks (Fritz et al., 2003; David et al., 2012; Slee and David, 2015). It is therefore possible in the context of each instrument subset that our cognitive processes can focus on different regions of the MPS in order to discriminate similar instrument sounds within a given stimulus context.

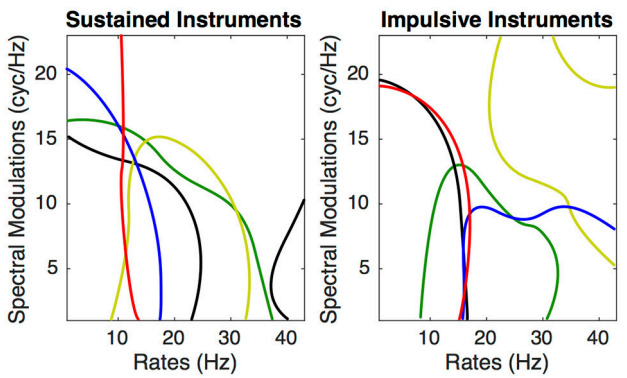


FIGURE 5 | DiagnosticMasks of the MPS of the five sustained instruments. Left: Tuba red, Clarinet green, Saxophone yellow, Cello blue, Trombone black. **Right**: Viola Pizzicato green, Guitar yellow, Harp red, Marimba green, Vibraphone black.

These results can also be considered in the light of the recent study of Isnard et al. (2016) who showed that severely impoverished sounds in the time-frequency domain—music, speech or environmental sounds—can still be recognized. In the same way, Suied et al. (2013) determined a perceptually sparse representation of speech sounds in the spectrotemporal modulation domain in order to determine the minimum acoustic information necessary to convey emotions in speech sounds. In line with this work, we have shown here that musical instrument sounds impoverished in the spectrotemporal modulation domain can still be recognized.

From a more general perspective, these two experiments are a first step toward determining new acoustic descriptors relevant to the perception of musical timbre. Even if the MPS appears to be less intuitive than the time-frequency representation, it must be noted that it is an ingenious way to describe the spectrum of a sound as it is invariant according to several transformations in the time-frequency domain. Here, we considered a spectrogram with a linear frequency scale for which the MPS is invariant by translation in the time-frequency domain. Hence we may expect to determine acoustical invariants that characterize musical instruments categories (McAdams, 1993) from these representations.

CONCLUSION

The results of this study shed light on the most relevant regions of the MPS for the identification of musical instrument timbre. From a perceptual point of view, this research provides a ground from which to investigate whether the MPS regions determined here could be used to determine new timbre descriptors and/or serve as a sound representation for automatic recognition algorithms. Moreover, comparison with other approaches to timbre such as multidimensional scaling might be an interesting perspective of this work, although Elliott et al. (2013) found

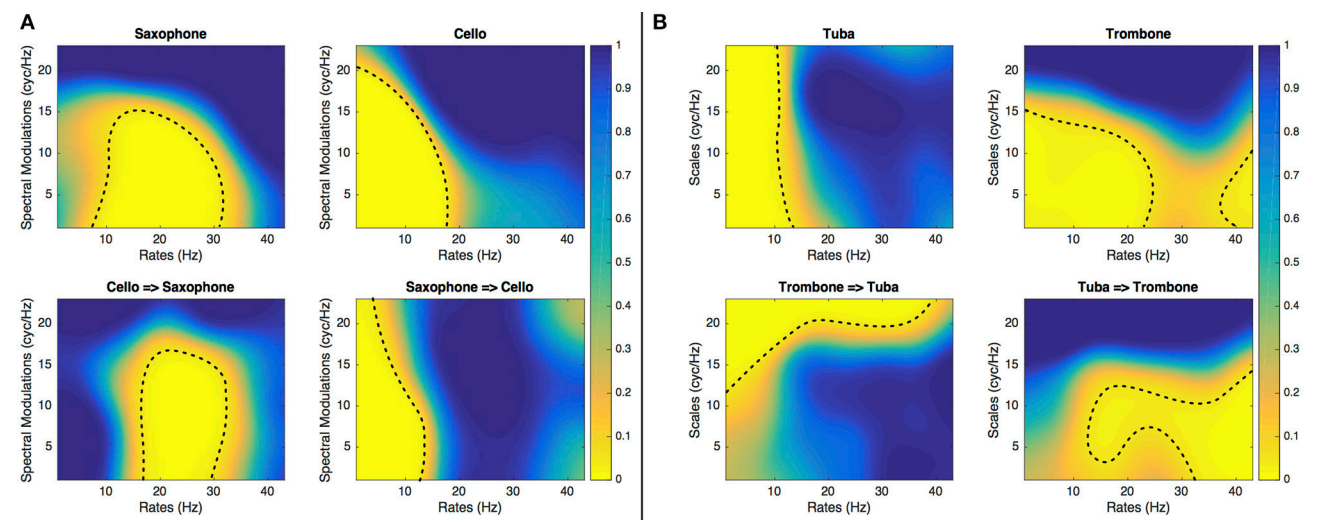


FIGURE 6 | Experiment 1. *ConfusionMasks* of the cello/saxophone **(A)** and trombone/tuba **(B)**. The dashed lines represent the thresholds at $p = 0.05$. The upper panels show the instrument identified as itself and the lower panels show the instrument identified as another instrument.

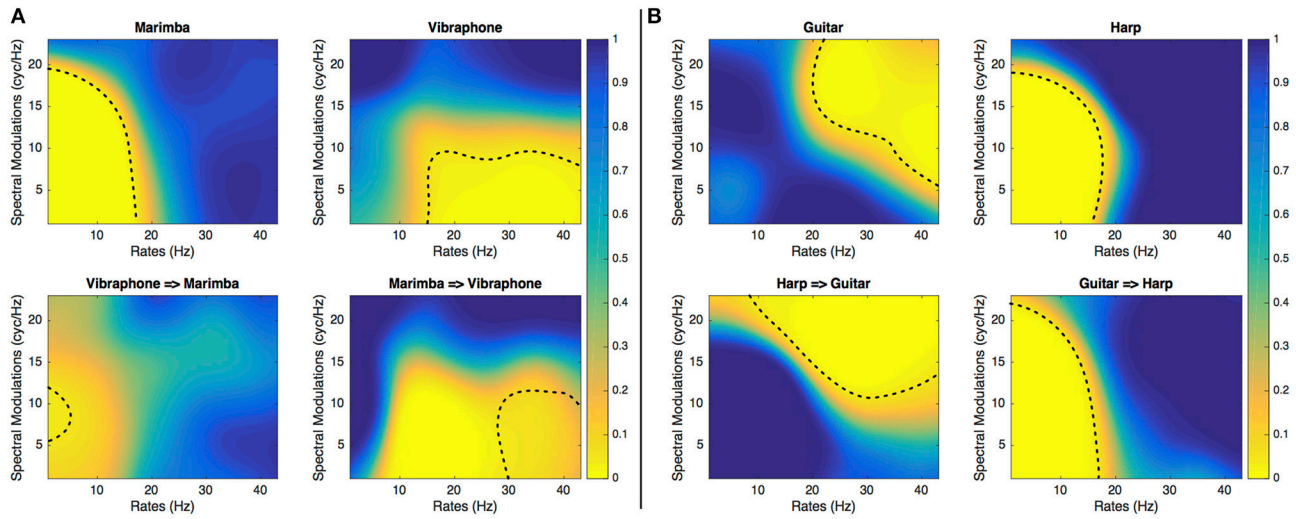


FIGURE 7 | Experiment 2. *ConfusionMasks* of the marimba/vibraphone **(A)** and guitar/harp **(B)**. The dashed lines represent the thresholds at $p = 0.05$. The upper panels show the instrument identified as itself and the lower panels show the instrument wrongly identified as another instrument.

fairly similar predictive power for MPS representations and combinations of unidimensional audio descriptors. Future research will focus on how this new approach is linked to the other conceptions of timbre. In particular, we can expect to link temporal modulations to the relevant aspects of the temporal envelope (e.g., the attack time) and similarly with spectral modulation and spectral envelope properties (e.g., formant and pitch). As the stimuli were composed of arpeggios, no specific analysis has been done on how filtering in the MPS domain might impact properties such as attack time for each note. It is for instance plausible that the filtering in the temporal modulation dimension may have impacted rise times. Moreover, other parameters such as the loudness of the filtered stimuli may have influenced the identification scores and could also be investigated in further experiments, although it isn't clear how to "control" for this factor given that the filtered signals in different regions of the MPS have differing amounts of energy. Finally, it might be of interest to compare the relevance of the MPS representation with other spectrotemporal modulation representations such as those used by Patil et al. (2012) or Andén et al. (2015) inspired by the plausible two-dimensional wavelet achieved at the level of the primary auditory cortex by spectrotemporal receptive fields (Shamma, 2001).

ETHICS STATEMENT

The protocol of this study was certified for ethics compliance by the McGill Research Ethics Board II with written consent from all subjects in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

ET, PD, and SM conceived and designed the experiments. ET performed the experiments. ET, PD, and SM analyzed the data. ET, PD, and SM wrote the paper.

FUNDING

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada awarded to SM (RGPIN-2015-05208, RGPAS-478121-15) and to PD (RGPIN-262808-2012) as well as a Canada Research Chair awarded to SM.

ACKNOWLEDGMENTS

The authors are thankful to Grace Wang for help running participants.

REFERENCES

Andén, J., Lostanlen, V., and Mallat, S. (2015). "Joint time-frequency scattering for audio classification," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)* (New York, NY: IEEE), 1–6.

Aucouturier, J. J., and Bigand, E. (2013). Seven problems that keep MIR from attracting the interest of cognition and neuroscience. *J. Intell. Inf. Syst.* 41, 483–497. doi: 10.1007/s10844-013-0251-x

Brown, J. C., Houix, O., and McAdams, S. (2001). Feature dependence in the automatic identification of musical woodwind instruments. *J. Acoust. Soc. Am.* 109, 1064–1072. doi: 10.1121/1.1342075

Brown, J. C. (1999). Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. *J. Acoust. Soc. Am.* 105, 1933–1941. doi: 10.1121/1.426728

David, S. V., Fritz, J. B., and Shamma, S. A. (2012). Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2144–2149. doi: 10.1073/pnas.1117717109

Elliott, T. M., and Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Comput. Biol.* 5:e1000302. doi: 10.1371/journal.pcbi.1000302

Elliott, T. M., Hamilton, L. S., and Theunissen, F. E. (2013). Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. *J. Acoust. Soc. Am.* 133, 389–404. doi: 10.1121/1.4770244

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat. Neurosci.* 6, 1216–1223. doi: 10.1038/nn1141

Giordano, B. L., and McAdams, S. (2010). Sound source mechanics and musical timbre perception: evidence from previous studies. *Music Percept.* 28, 155–168. doi: 10.1525/mp.2010.28.2.155

Gosselin, F., and Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Res.* 41, 2261–2271. doi: 10.1016/S0042-6989(01)00097-9

Grey, J. M., and Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *J. Acoust. Soc. Am.* 63, 1493–1500. doi: 10.1121/1.381843

Griffin, D., and Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust. Speech Signal Process* 32, 236–243. doi: 10.1109/TASSP.1984.1164317

Hall, M. D., and Beauchamp, J. W. (2009). Clarifying spectral and temporal dimensions of musical instrument timbre. *Can. Acoust.* 37, 3–22.

Hemery, E., and Aucouturier, J. J. (2015). One hundred ways to process time, frequency, rate and scale in the central auditory system: a pattern-recognition meta-analysis. *Front. Comp. Neurisci.* 9:80. doi: 10.3389/fncom.2015.00080

Hjortkjær, J., and McAdams, S. (2016). Spectral and temporal cues for perception of material and action categories in impacted sound sources. *J. Acoust. Soc. Am.* 140, 409–420. doi: 10.1121/1.4955181

Isnard, V., Taffou, M., Viaud-Delmon, I., and Suied, C. (2016). Auditory sketches: very sparse representations of sounds are still recognizable. *PLoS ONE* 11:e0150313. doi: 10.1371/journal.pone.0150313

ISO 389–8 (2004). *Acoustics – Reference Zero for the Calibration of Audiometric Equipment – Part 8: Reference Equivalent Threshold Sound Pressure Levels for Pure Tones and Circumaural Earphones (Tech. Rep.)*. Geneva: International Organization for Standardization.

Iverson, P., and Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbre. *J. Acoust. Soc. Am.* 94, 2595–2603. doi: 10.1121/1.407371

Koenig, R., Dunn, H. K., and Lacy, L. Y. (1946). The Sound Spectrograph. *J. Acoust. Soc. Am.* 18, 19–49. doi: 10.1121/1.1916342

Martin, F. N., and Champlin, C. A. (2000). Reconsidering the limits of normal hearing. *J. Am. Acad. Audiol.* 11, 64–66.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. *Psychol. Res.* 58, 177–192. doi: 10.1007/BF00419633 doi: 10.1007/BF00419633

McAdams, S., Tse, A., and Wang, G. (2016, July). "Generalizing the learning of instrument identities across pitch registers," in *Paper Presented at the 14th International Conference on Music Perception and Cognition* (San Francisco, CA).

McAdams, S. (1993). "Recognition of sound sources and events," in *Thinking in Sound: The Cognitive Psychology of Human Audition*, eds S. McAdams and E. Bigand (Oxford: Oxford University Press), 146–198.

McAdams, S. (2013). "Musical timbre perception," in *The Psychology of Music*, *3rd Edn.*, ed D. Deutsch (San Diego, CA: Academic Press), 35–67.

Pachet, F., and Aucouturier, J. J. (2004). Improving timbre similarity: how high is the sky. *J. Negat. Results Speech Audio Sci.* 1, 1–13.

Patil, K., and Elhilali, M. (2015). Biomimetic spectro-temporal features for music instrument recognition in isolated notes and solo phrases. *EURASIP J. Adv. Sig. Pr.* 2015:27. doi: 10.1186/s13636-015-0070-9

Patil, K., Pressnitzer, D., Shamma, S., and Elhilali, M. (2012). Music in our ears: the biological bases of musical timbre perception. *PLoS Comput. Biol.* 8:e1002759. doi: 10.1371/journal.pcbi.1002759

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (2011). The timbre toolbox: extracting audio descriptors from musical signals. *J. Acoust. Soc. Am.* 130, 2902–2916. doi: 10.1121/1.3642604

Rabiner, L. R., and Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice Hall.

Saldanha, E. L., and Corso, J. F. (1964). Timbre cues and the identification of musical instruments. *J. Acoust. Soc. Am.* 36, 2021–2026. doi: 10.1121/1.1919317

Shamma, S. (2001). On the role of space and time in auditory processing. *Trends Cogn. Sci.* 5, 340–348. doi: 10.1016/S1364-6613(00)01704-6

Siedenburg, K., Fujinaga, I., and McAdams, S. (2016). A comparison of approaches to timbre descriptors in music information retrieval and music psychology. *J. New Music Res.* 45, 27–41. doi: 10.1080/09298215.2015.1132737

Singh, N. C., and Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *J. Acoust. Soc. Am.* 114, 3394–3411. doi: 10.1121/1.1624067

Slaney, M. (1994). *An Introduction to Auditory Model Inversion.* Interval Technical Report IRC1994. Available online at: https://engineering.purdue.edu/%7emalcolm/interval/1994-014/

Slee, S. J., and David, S. V. (2015). Rapid task-related plasticity of spectrotemporal receptive fields in the auditory midbrain. *J. Neurosci.* 35, 13090–13102. doi: 10.1523/JNEUROSCI.1671-15.2015

Suied, C., Drémeau, A., Pressnitzer, D., and Daudet, L. (2013). "Auditory sketches: sparse representations of sounds based on perceptual models," in *International Symposium on Computer Music Modeling and Retrieval*, eds M. Aramaki, M. Barthet, R. Kronland-Martinet, and S. Ystad (Berlin; Heidelberg: Springer), 154–170.

Venezia, J. H., Hickok, G., and Richards, V. M. (2016). Auditory bubbles: efficient classification of the spectrotemporal modulations essential for speech intelligibility. *J. Acoust. Soc. Am.* 140, 1072–1088. doi: 10.1121/1.4960544

Vienna Symphonic Library. Available online at: http://vsl.co.at/en

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 08 February 2017
doi: 10.3389/fpsyg.2017.00153

Perception and Modeling of Affective Qualities of Musical Instrument Sounds across Pitch Registers

Stephen McAdams[1]*, *Chelsea Douglas*[1] and *Naresh N. Vempala*[2]

[1] Music Research, Schulich School of Music, McGill University, Montreal, QC, Canada, [2] Department of Psychology, Ryerson University, Toronto, ON, Canada

Composers often pick specific instruments to convey a given emotional tone in their music, partly due to their expressive possibilities, but also due to their timbres in specific registers and at given dynamic markings. Of interest to both music psychology and music informatics from a computational point of view is the relation between the acoustic properties that give rise to the timbre at a given pitch and the perceived emotional quality of the tone. Musician and nonmusician listeners were presented with 137 tones produced at a fixed dynamic marking (forte) playing tones at pitch class D# across each instrument's entire pitch range and with different playing techniques for standard orchestral instruments drawn from the brass, woodwind, string, and pitched percussion families. They rated each tone on six analogical-categorical scales in terms of emotional valence (positive/negative and pleasant/unpleasant), energy arousal (awake/tired), tension arousal (excited/calm), preference (like/dislike), and familiarity. Linear mixed models revealed interactive effects of musical training, instrument family, and pitch register, with non-linear relations between pitch register and several dependent variables. Twenty-three audio descriptors from the Timbre Toolbox were computed for each sound and analyzed in two ways: linear partial least squares regression (PLSR) and nonlinear artificial neural net modeling. These two analyses converged in terms of the importance of various spectral, temporal, and spectrotemporal audio descriptors in explaining the emotion ratings, but some differences also emerged. Different combinations of audio descriptors make major contributions to the three emotion dimensions, suggesting that they are carried by distinct acoustic properties. Valence is more positive with lower spectral slopes, a greater emergence of strong partials, and an amplitude envelope with a sharper attack and earlier decay. Higher tension arousal is carried by brighter sounds, more spectral variation and more gentle attacks. Greater energy arousal is associated with brighter sounds, with higher spectral centroids and slower decrease of the spectral slope, as well as with greater spectral emergence. The divergences between linear and nonlinear approaches are discussed.

Keywords: musical timbre, emotion, pitch register, musical instruments, valence, tension arousal, energy arousal, preference

OPEN ACCESS

Edited by:
Antonino Vallesi,
University of Padua, Italy

Reviewed by:
Giulio Pergola,
University of Bari, Italy
Eveline Vernooij,
University of Udine, Italy

***Correspondence:**
Stephen McAdams
stephen.mcadams@mcgill.ca

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *11 November 2016*
Accepted: *23 January 2017*
Published: *08 February 2017*

Citation:
McAdams S, Douglas C and Vempala NN (2017) Perception and Modeling of Affective Qualities of Musical Instrument Sounds across Pitch Registers. Front. Psychol. 8:153. doi: 10.3389/fpsyg.2017.00153

INTRODUCTION

The relationship between music and emotion has become a widely studied topic. Its existence is undeniable and multiple studies have revealed that for most people the predominant motivation for listening to and engaging in music is its emotional impact (Sloboda and O'Neill, 2001; Krumhansl, 2002; Juslin and Laukka, 2004). Although, there is an increasing amount of research on music and emotion, it remains difficult to draw decisive conclusions about how musical factors contribute to emotion in a piece. In addition to global structural factors such as mode, melody, harmony, tempo, and form (cf. Gabrielsson and Lindström, 2010 for a review), it is likely that acoustic factors of a sound can relay affective information as well. In this paper, we examine musical instrument timbre and the audio descriptors derived from the sound signal that contribute to its timbre in relation to a three-dimensional model of perceived affect. We also develop linear regression and nonlinear neural net models to establish a computational link between audio descriptors and perceived emotion ratings, providing a basis for a music informatics approach to the role of timbre in emotion perception.

A three-dimensional model of affect measures emotion as a function of valence, tension arousal, and energy arousal (Schimmack and Grob, 2000). This model likely provides a more complete representation of affect than the two-dimensional model with only valence and arousal because tension arousal and energy arousal have been shown to be two distinct measures of activation that should not be collapsed into a single measure (Schimmack and Reisenzein, 2002). Schubert (1999) completed a series of experiments applying a dimensional model of affect to music research and found the dimensional model to be a valid and reliable measure for research involving music and emotion. Furthermore, the three-dimensional model of affect has recently been applied to multiple perceived emotion and music studies (Ilie and Thompson, 2006; Eerola et al., 2009, 2012).

Emotion perception refers to a listener recognizing an expressed emotion, but does not necessitate the feeling of that emotion (Juslin and Västfjäll, 2008). When examining expressed emotion in an entire piece, researchers have mostly focused on pitch combination and order, as well as tempo, which has led to the understanding that structural factors play an important role in emotion perception in music listening. However, listeners' judgments of perceived emotion are not solely based on these structural elements. By altering factors such as amplitude, pitch register, pitch contour, temporal envelope, and filtering in synthesized tone sequences, over two-thirds of the variance in listener's perceived emotion ratings have been explained by the manipulation of the acoustic cues (Scherer and Oshinsky, 1977). Further research supports the notion that finer acoustic features, such as dynamics, articulation, spectrum, and attack character are also factors listeners consider when making emotion judgments (Juslin and Laukka, 2004; Gabrielsson and Lindström, 2010). The latter three factors are components that contribute to the timbre of a sound (McAdams et al., 1995). Furthermore, performers and composers reportedly use timbre as a means of communicating intended emotion to listeners (Holmes, 2011), and parallels involving timbral dimensions have been drawn between perceived emotion in music and in speech sounds (Juslin and Laukka, 2003).

Timbre is a multidimensional acoustic attribute that is composed of spectral, temporal, and spectrotemporal dimensions (McAdams et al., 1995). The term "timbre" refers to a set of perceptual attributes that listeners use to discriminate different sound sources in addition to pitch, loudness, duration, and spatial position. These attributes also contribute to source identity (McAdams, 1993). Additionally, the timbre of acoustic instruments varies with both pitch register and musical dynamics, i.e., a given instrument played in a low register can have a drastically different timbre when played in a high register, and at a given pitch, a change in dynamics (playing effort) is also accompanied by a change in timbre (Risset and Wessel, 1999; Marozeau et al., 2003; McAdams and Goodchild, forthcoming).

The notion that perceived emotion can be judged by non-structural acoustic features is supported by listeners' ability to make emotional judgments on sound samples of extremely short duration, and therefore, with limited acoustic information. In certain cases, as little as 250 ms of a musical excerpt holds enough information to perceive an emotional tone in a consistent manner across listeners (Peretz et al., 1998; Filipic et al., 2010), and even a single note provides listeners with enough cues to form an emotional judgment (Bigand et al., 2005). Furthermore, musical expertise has no impact on musical recognition and emotional judgments based on minimal acoustic information (Filipic et al., 2010). The ability to recognize emotion in such short stimuli emphasizes the importance of examining how individual acoustic factors, such as timbre, contribute to emotion perception in music.

Timbre has been identified as a musical feature correlated with perceived, discrete emotions. In general, bright sounds are associated with happiness, dull sounds with sadness, sharp sounds with anger, and soft sounds with both fear and tenderness (Juslin and Laukka, 2004). When one group of participants in a study by Huron et al. (2014) was asked to judge acoustic properties of 44 Western instruments and another group to judge those instruments' ability to express sadness, all judgments appeared to be based on participants' familiarity and knowledge of the instruments rather than on listening to their timbres. Using the acoustic property judgments, such as the darkness of sound, from the first group of participants as predictors for the sadness judgments of the second group of participants, Huron et al. concluded that acoustic properties of the instruments, the ability of the instrument to make small pitch movements, to play low pitches, and to play quietly predicted sadness judgments. Furthermore, Hailstone et al. (2009) studied timbre as a main factor contributing to emotion perception in music. Listeners heard melodies that possess a strong emotional intent and labeled them with an emotion category in a forced-choice paradigm. The melodies were played by one of four different instruments. There was a significant interaction between instrument and emotion judgment, suggesting that timbral cues may be more important for communicating some basic emotions than others. However, Hailstone et al.'s experiment only studied four instrument sounds and four synthetic sounds with novel melodies, which could possibly confound the emotional expression of the timbre alone.

Eerola et al. (2012) examined timbre in relation to a two-dimensional affect model of valence and energy arousal. The stimuli consisted of 110 recorded samples of musical instrument sounds. Pitch and duration were kept constant at D#4 (311-Hz fundamental frequency) and 1 s, respectively, across all stimuli, and loudness was equalized. Participants listened to the individually presented stimuli and gave affect and preference ratings. The rating scales included valence (pleasant/unpleasant), energy arousal (awake/tired), tension arousal (tense/relaxed), and preference (like/dislike). The three-dimensional model of affect (Schimmack and Grob, 2000) used to collect ratings was reduced to a two-dimensional model for analysis purposes due to a highly collinear relationship between the energy-arousal and tension-arousal dimensions. Furthermore, the valence and preference ratings in their study had a nearly perfect correlation, $r = 0.97$. They also investigated the acoustic cues contributing to affect ratings of individual musical instrument sounds drawn from woodwind, string, brass, and percussion families and equalized in pitch, duration, and loudness. They selected seven audio descriptors based on a principal component analysis of 26 descriptors from the MIRToolbox (Lartillot and Toiviainen, 2007). Valence ratings were primarily explained by a linear combination of the ratio of high- to low-frequency energy, temporal envelope centroid, and spectral skewness, with positive valence resulting from sustained sounds with more energy in the lower-frequency components. Energy-arousal ratings were more related to the ratio of high- to low-frequency energy, temporal envelope centroid, and attack slope, with energetic sounds having sharper attacks and more dominant high frequency components.

It is important to note here that Eerola et al.'s (2012) bipolar emotional valence scale was labeled from unpleasant to pleasant. This provides a clear methodological difference compared to Bigand et al.'s (2005) study in which emotional valence (positive/negative) and pleasantness were rated on two separate scales. The stimuli used in Bigand et al.'s Experiments 2 and 3 were orchestral excerpts of short duration (1 s), sometimes consisting of one single tone. The difference in definition of the scales may have contributed to a key difference in the findings, because Bigand's group found that the emotional valence dimension was not correlated with pleasantness judgments ($r = 0.08$), suggesting that happy music is not necessarily identified with pleasant emotions or sad music with unpleasant emotions. These findings lead to our hypothesis that perceived valence will not be completely correlated with preference ratings in the current experiment.

Pitch height (or register) is also a factor in the perceived affective quality of music and musical sounds. In an extensive review of the literature on musical cues to emotion, Gabrielsson and Lindström (2010) report somewhat inconsistent findings on the link between affective qualities and pitch height. Across the studies reviewed, higher pitch was variously associated with expressions such as happy, serene, dreamy, graceful, exciting, surprising, potent, angry, fearful, and active, whereas lower pitch was characterized as sad, dignified/solemn, vigorous, exciting, boring, and pleasant. These authors suggest that the apparent contradictions may depend on musical context. In the speech domain, higher pitch is associated with arousing and happy affect and a submissive manner and lower pitch with calming and sad affect and a more threatening manner (Frick, 1985), a principle that appears to carry over into music (Juslin and Laukka, 2003; Huron et al., 2006). Two studies have explicitly manipulated pitch height (in addition to other musical parameters) to determine its effect on emotion perception. Eerola et al. (2013) in an emotion category rating paradigm found that at lower pitch, ratings were higher for "scary" and "sad" and lower for "happy" and "peaceful," whereas at higher pitch, ratings were higher for "happy," intermediate for "sad" and "peaceful" and lowest for "scary." Ilie and Thompson (2006) used the 3D model of emotion and found that higher pitch was rated as more pleasant for music, but as less pleasant for speech, than was lower pitch. In an analysis of several thousand instrumental themes, Huron (2008) found that on average pitch height was slightly lower for minor-key than for major-key themes, indicating that composers intuitively use a pitch height in addition to mode to convey emotional tone.

The following experiment aims to further contribute to research regarding the role of timbre and pitch-register-related differences in timbre in affect perception in music by showing that participants' judgments regarding perceived affect vary systematically with timbral qualities of short instrument sounds across their pitch registers. First we examined affect ratings in relation to broader variables such as pitch register and instrument family with a linear mixed model analysis, thus extending Eerola et al.'s research by including different pitch registers. The role of tessitura in emotion perception is a little-studied but important issue, because orchestration treatises often mention the different qualities of instrument sounds in their different registers (e.g., Adler, 2002). Confining a study to a single pitch places some instruments in their optimal middle register and others in extreme low or high registers, which require greater playing effort and may by consequence affect their emotional qualities. By extending the registers, we also expected to find different patterns in the tension-arousal ratings compared to the energy-arousal ratings, supporting a three-dimensional model of affect, instead of a two-dimensional model. Additionally, we expected to find a difference in the perceived valence ratings compared to the preference ratings, highlighting a difference between perceived measures and felt measures. Finally, we expected a significant interaction between pitch register and instrument family for each of the perceived affect ratings, showing perceived emotion ratings may not be the same for all instruments across pitch registers.

To relate the perceptual results to timbral properties, we then examined the relationship between the perceived affect ratings and specific audio descriptors that compose timbre with two techniques: a linear partial least squares regression (PLSR) approach and a nonlinear artificial neural network model. PLSR is a regression method that uses principal components analysis (PCA) as an integral part and originates from the discipline of chemometrics (Geladi and Kowalski, 1986). However, it has been applied more recently within the field of auditory perception (Rumsey et al., 2005; Kumar et al., 2008; Eerola et al., 2009). PLSR analyzes complex correlational relationships between perceptual measures as dependent variables and arrays of acoustical or psychoacoustical variables (hereafter referred to as audio descriptors) as independent variables. It deals with

collinearity among independent variables by capturing what is common among them in the principal components. It thus carries out simultaneous data reduction and maximization of covariance between the descriptors and the predicted data, preserving correlational patterns between them. Supervised feedforward artificial neural networks with back propagation (i.e., multilayer perceptrons; Rumelhart et al., 1986; Haykin, 2008) are useful connectionist models that act as nonlinear regression functions for emotion prediction based on audio descriptors. The architecture of the feedforward networks is simple with one hidden layer providing the necessary level of nonlinearity.

METHODS

Affect Ratings

The experimental design isolated timbre as an independent variable, similar to Experiment 1 presented in Eerola et al. (2012), and allowed us to examine register, attack, and playing technique as factors contributing to timbre. Modifications, such as an added valence measure and increased range of pitch register, facilitated a comparison between emotional valence and preference scales as well as examining how changes in register contribute to changes in timbral components that influence emotion ratings.

Participants

Forty participants (24 females) were between 18 and 35 years of age ($M = 23$, $SD = 4.4$). Twenty participants reported formal musical training ranging from 7 to 25 years of practice ($M = 16$, $SD = 5.3$), and 14 reported formal training with multiple instruments. The remaining 20 participants reported no musical training at a collegiate level and no more than 1 year of formal music training during childhood. These two groups will be referred to as musicians and nonmusicians, respectively. The difference in age between the two groups was not significant, $t_{(38)} = -0.42$, $p = 0.68$.

Stimuli

One hundred and thirty seven recorded instrument sounds were chosen from the Vienna Symphonic Library (Vienna Symphonic Library GmbH, 2011). The recorded samples consisted of sounds played by orchestral instruments from four instrument families: brass, woodwinds, strings, and pitched percussion. Audio signals were sampled at 44.1 kHz with 16-bit amplitude resolution. The stimuli were edited to have a fixed duration of 500 ms with a raised-cosine ramp applied to fade them out over the final 50 ms. The attack of each sound was unaltered. The brass stimuli varied by an attack parameter, having a weak, normal or strong attack, as labeled in VSL. The percussion stimuli varied by mallet material, using a felt, wood or metal mallet. Pitch class was kept constant at D#, and the dynamic level was *forte*, as labeled in VSL. Samples were chosen from the entire range of the instruments with stimuli ranging from D#1 to D#7 (A4 has a fundamental frequency of 440 Hz). Most instruments cannot successfully play from D#1 to D#7, so stimuli were only taken from appropriate and playable registers for each instrument. Although, some instruments can play outside of that range, there were not enough samples to create useful, balanced groups. Furthermore, various techniques, such as flutter-tonguing for brass and woodwinds and vibrato and pizzicato for strings, were also included. A detailed list of the stimuli is provided in Table S1.

Procedure

All participants passed a pure-tone audiometric test using a MAICO MA 39 (MAICO Diagnostic GmbH, Berlin, Germany) audiometer at octave-spaced frequencies from 125 Hz to 8 kHz and were required to have thresholds at or below 20 dB HL in order to proceed to the experiment (Martin and Champlin, 2000; ISO, 2004).

The interface was created in TouchOSC (Hexler.net, 2011) and consisted of a *play* button, six clearly labeled 9-point, analogical-categorical scales (Weber, 1991), and a *next* button. The *next* button was not activated until all six ratings were completed; pressing this button would reset the display to the original position and play the next sound. All 137 stimuli were presented in a randomized order for each participant and each sound could be played as many times as desired, although this information was not recorded. Participants completed six ratings per sound on the 9-point scales. The first four ratings measured perceived emotion and reflected affect dimensions from the three-dimensional model of affect (Schimmack and Grob, 2000) with an additional measure of negative to positive valence. The scales were labeled at the left and right ends with the following pairs: negative/positive (valence), displeasure/pleasure (valence), tired/awake (energy arousal), and tense/relaxed (tension arousal). The participants were also reminded that a rating of 5 would equate to a neutral rating. These four scales were labeled in blue on the iPad interface. The last two ratings measured participants' preference for and familiarity with each sound. These scales were labeled with the pairs dislike/like and unfamiliar/familiar, respectively. These two scales provided a felt rating of personal preference and familiarity and were labeled in purple to differentiate them from the perceived affect ratings. An example of the interface is displayed in **Figure 1**. Participants were given the following specific instructions: "For the first four scales, you will be rating the degree to which the sound expresses a feeling (NOT how it makes you feel). The last two ratings are how you feel about the sound." Participants completed the task within an hour and were compensated for their time.

Apparatus

Participants completed the experiment individually inside an IAC model 1203 sound-isolation booth (IAC Acoustics, Bronx, NY). The sound samples were played from a Macintosh G5 computer (Apple Computer, Inc., Cupertino, CA), amplified with a Grace Design m904 monitor system (Grace Digital Audio, San Diego, CA), and heard over circumaural Sennheiser HD280 Pro headphones (Sennheiser Electronic GmbH, Wedemark, Germany). The participants were not allowed to adjust the volume. Sound levels were measured with a Brüel and Kjær Type 2205 sound-level meter (A-weighting) connected to a Type 4152 artificial ear (Brüel and Kjær, Nærum, Denmark) to which the headphones were coupled. Stimuli ranged between 59.8 and 77.5 dB SPL ($M = 65.3$, $SD = 5.4$). The participants completed

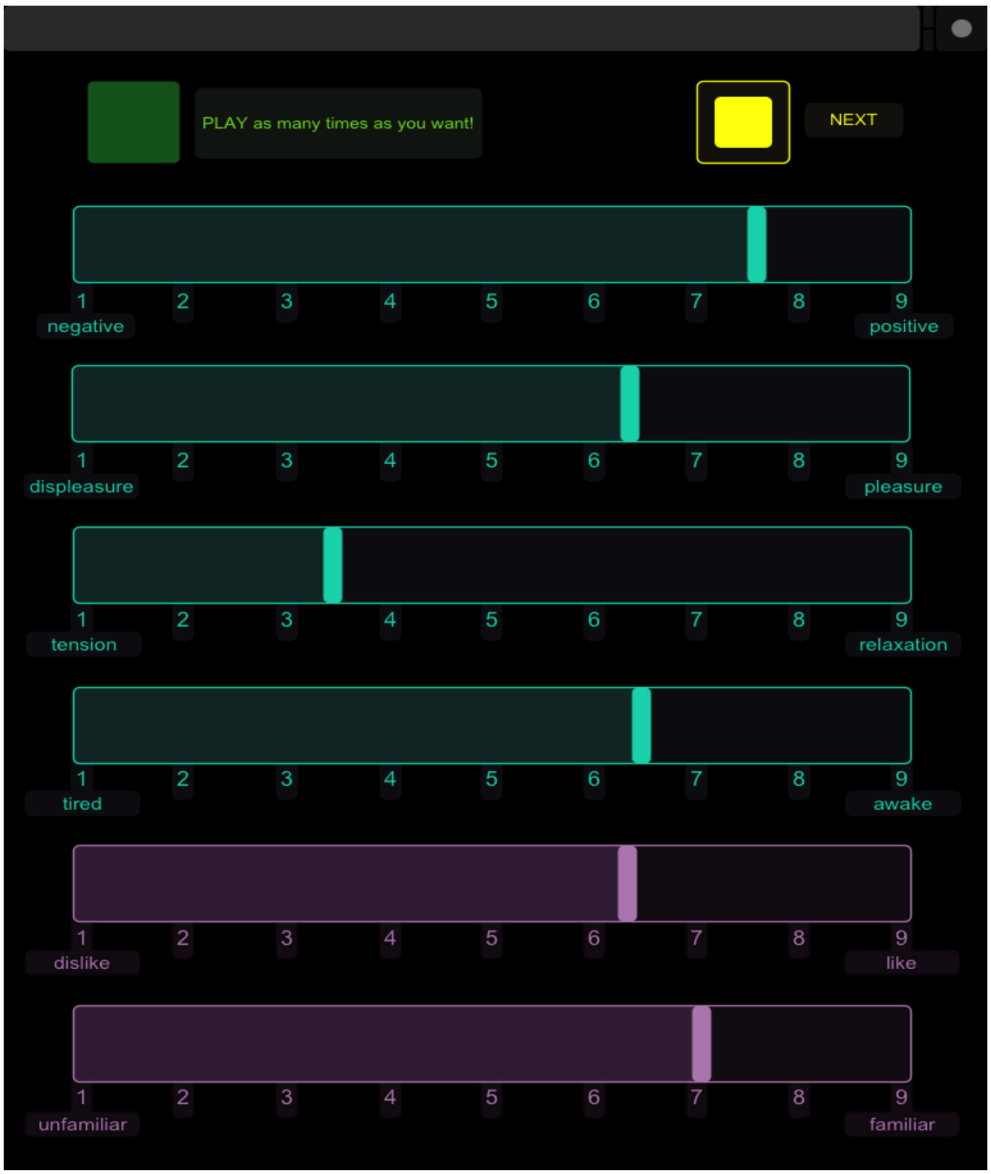


FIGURE 1 | Screenshot of the experimental interface on the iPad.

the experiment on an iPad interface (Apple Computer, Inc., Cupertino, CA). The iPad communicated via OpenSoundControl (Center for New Music and Audio Technologies, Berkley, CA) messages over a wireless network with a Max/MSP version 5.1.9 (Cycling '74, San Francisco, CA) patch run on the Macintosh computer. The Max/MSP patch was designed to randomize and play the stimuli as well as to record and output the ratings.

Control Experiment

A control experiment was completed after the original experiment to validate the original interface. The main purpose was to confirm, with a correlation analysis between the control ratings and the original ratings, that no bias resulted from the order and orientation of the rating scales, which remained in a fixed position for every trial and every participant in the original experiment.

Participants

Twenty participants (12 females) were between 18 and 42 years of age ($M = 25$, $SD = 6.5$). Ten participants reported formal musical training ranging from 13 to 19 years of practice ($M = 16$, $SD = 2.5$), and seven of those reported formal training with multiple instruments. The remaining 10 participants reported no musical training at a collegiate level and no more than a year of formal music training during childhood. The difference in age between the two groups was not significant, $t_{(18)} = 0.69$, $p = 0.50$.

Stimuli

Forty stimuli were selected from the original 137 samples to create a group that was representative of the entire set. Therefore, the control stimuli were brass, woodwind, string, and percussion samples ranging from D#1 to D#7 with weak, normal, and strong attacks. The relevant samples are marked with asterisks in Table S1.

Procedure

The instructions and procedure were identical to the original experiment. However, participants were randomly given one of four different interfaces. The interfaces included the same *play* and *next* buttons as the original, but the order of the six scales was changed as well as the orientation (i.e., the end labels) of some of the scales. However, the blue "perceived" scales and the purple "felt" scales were always grouped together to avoid confusion between perceived and felt ratings.

RESULTS

Consistency and Correlation Analyses

We conducted initial reliability analyses and correlations among the scales. All scales had good internal consistency (Cronbach's α for 40 participants = 0.93 for positive/negative, 0.91 for pleasure/displeasure, 0.92 for relaxed/tense, 0.90 for awake/tired, 0.97 for like/dislike, and 0.99 for familiar/unfamiliar). Subsequently, the ratings were averaged across participants for the correlation analysis, so each of the 137 sounds had one measure for each of the six rating scales.

Table 1 displays the correlations between scales within and between the main and control experiments as well as the correlations between the main and control experiments for the 40 sounds common to both studies. Correlations between dependent variables in the main experiment include all 137 sounds, but the correlations within the control experiment and between control and main experiments are only based on the 40 sounds common to both experiments. As all scales in the experiment were very strongly correlated with the designated control, $r_{(38)} \geq 0.89$, $p < 0.001$, the original interface was confirmed to be valid and reliable. Further analysis is completed on data from the main experiment only.

In the main experiment, the valence scales labeled negative/positive and displeasure/pleasure had a very strong Pearson's correlation of $r_{(135)} = 0.97$, $p < 0.001$. Therefore, in the following analyses, the valence measure will only refer to the negative/positive scale, and the displeasure/pleasure scale will not be analyzed further. The tension-arousal and energy-arousal ratings were only very weakly correlated. There was a strong positive correlation between preference and valence ratings, and a strong positive correlation between preference and tension ratings. However, there was only a very weak positive correlation between preference and energy ratings. Valence was moderately negatively correlated with tension-arousal ratings and strongly correlated with energy-arousal ratings. Familiarity was weakly to moderately correlated with valence, tension arousal, and energy arousal and strongly correlated with preference.

Linear Mixed Model Analyses

Further statistical analyses employed a linear mixed model method (West et al., 2006), which performs a regression-like analysis while controlling for random variance caused by differences in factors such as participant and stimulus. Because each participant rated all stimuli, the model included crossed random effects for participant and stimulus (Baayen et al., 2008). Specifically, a maximal random effects structure was implemented due to the confirmatory hypothesis nature of the analyses and to reduce Type I errors, i.e., false positives (Barr et al., 2013). Analyses were completed with the R software environment v3.0.2 (www.r-project.org) using the lmer function from the lme4 package (Bates et al., 2014), the Anova function from the Companion to Applied Regression (car) package (Fox and Weisberg, 2011), and the lsmeans package for polynomial contrasts (Lenth, 2013). Welch's unequal variance *t*-test is used to test the significance of the polynomial contrasts.

A linear mixed model analysis was completed for each of the three perceived affect ratings (valence, tension arousal, energy

TABLE 1 | Person's correlation coefficients among ratings of perceived valence, tension arousal, energy arousal, preference, and familiarity for sounds common to both the main and control experiments.

		Main					Control			
		Valen	Tens	Ener	Pref	Famil	Valen	Tens	Ener	Pref
Main	Tension	0.46**								
	Energy	0.68**	−0.29**							
	Preference	0.72**	0.75**	0.19						
	Familiarity	0.56**	0.38**	0.31*	0.66**					
Control	Valence	0.89**	−0.65**	0.44	0.81**	0.60**				
	Tension	−0.28	0.89**	0.47	−0.67**	−0.31	0.51*			
	Energy	0.56**	0.37	0.92**	0.01	0.18	0.33	−0.57**		
	Preference	0.57**	−0.72**	0.02	0.89**	0.67**	0.80**	0.69**	−0.06	
	Familiarity	0.46	−0.37	0.20	0.57**	0.90**	0.56**	0.31	0.14	0.65**

*df = 135 for comparisons among Main variables and df = 38 between Main and Control and among Control variables. Bonferroni-corrected *p < 0.05, **p < 0.01.*

TABLE 2 | Linear mixed effects model type III wald *F*-Tests for ratings of perceived valence, tension arousal, energy arousal, preference, and familiarity.

	df	*F*	*p*	*df*	*F*	*p*
	Valence ($R^2 = 0.53$)			**Tension arousal ($R^2 = 0.43$)**		
Intercept	1, 124.4	145.63	< 0.001	1, 121.70	121.96	< 0.001
Training (T)	1, 136.88	0.43	0.512	1, 132.04	1.56	0.214
Family (F)	3, 120.66	7.30	< 0.001	3, 121.11	4.04	0.009
Register (R)	6, 122.01	7.98	< 0.001	6, 120.75	4.15	< 0.001
T × F	3, 104.62	0.24	0.871	3, 95.27	1.38	0.253
T × R	6, 69.38	3.71	0.003	6, 56.34	2.27	0.050
F × R	16, 111.00	2.41	0.004	16, 111.00	2.41	0.004
T × F × R	16, 111.00	1.04	0.417	16, 111.00	2.13	0.011
	Energy arousal ($R^2 = 0.50$)			**Preference ($R^2 = 0.51$)**		
Intercept	1, 124.02	381.45	< 0.001	1, 135.08	120.15	< 0.001
T	1, 137.48	0.05	0.819	1, 96.02	1.97	0.164
F	3, 118.78	1.67	0.178	3, 125.81	10.19	< 0.001
R	6, 112.83	13.27	< 0.001	6, 120.90	1.65	0.139
T × F	3, 91.45	1.44	0.239	3, 94.77	1.10	0.353
T × R	6, 53.83	2.10	0.068	6, 58.39	3.89	0.002
F × R	16, 111.00	3.09	< 0.001	16, 111.00	1.44	0.135
T × F × R	16, 111.00	2.30	0.006	16, 111.00	1.99	0.020
	Familiarity ($R^2 = 0.59$)					
Intercept	1, 149.78	112.47	< 0.001			
T	1, 70.46	5.89	0.018			
F	3, 131.80	6.33	< 0.001			
R	5, 120.36	0.82	0.540			
T × F	3, 81.11	2.09	0.108			
T × R	5, 69.51	0.58	0.716			
F × R	16, 111.00	1.85	0.039			
T × F × R	16, 111.00	1.62	0.084			

N = 5480. All predictors are sum-coded factor variables. The following random effects were included: (a) random intercepts for Participant and Sounds, (b) random slopes for Family and Register (within Participants) and Training (within Sounds).

arousal), as well as for the preference and familiarity ratings. Fixed factors examined in these models included instrument family and pitch register of the sounds and musical training of the participants. Attack and playing technique parameters were not included in these initial analyses in order to simplify this model, and because they are not always comparable across instrument families, e.g., flutter tonguing is confined to wind instruments. However, those factors were included in the models for the individual instrument families. Because all participants rated all sounds, a crossed random effects design was implemented and the maximal random effects structure thus included random intercepts for participant with random slopes for family and register, and random intercepts for the stimuli with random slopes for training.

Instrument Family and Pitch Register

Type III Wald *F*-test results from the five models are displayed in **Table 2**. Musical training alone was only a significant predictor of familiarity ratings, although the interaction of training and register was a significant predictor for valence, tension-arousal, and preference ratings. Family was a significant predictor for all ratings except energy-arousal. Register alone and the interaction between register and family both significantly predicted the perceived affect ratings, but not the preference and familiarity ratings. Register was especially influential for energy-arousal ratings. The three-way interaction between training, family, and register was significant for tension-arousal, energy-arousal, and preference ratings.

Figures 2–6 display plots of predicted means for each rating scale across register for each family and training group. Polynomial contrasts on valence, tension arousal and energy arousal were computed over octaves 2–6 (in which all instrument families are present) with the lsmeans package in R separately for musicians and nonmusicians (see Table S2). For Valence ratings (**Figure 2**), register was highly significant and globally presents a concave (inverted U-shaped) increasing form with a peak around octave 5 or 6. The polynomial contrasts reveal significant linear increasing and concave quadratic trends for

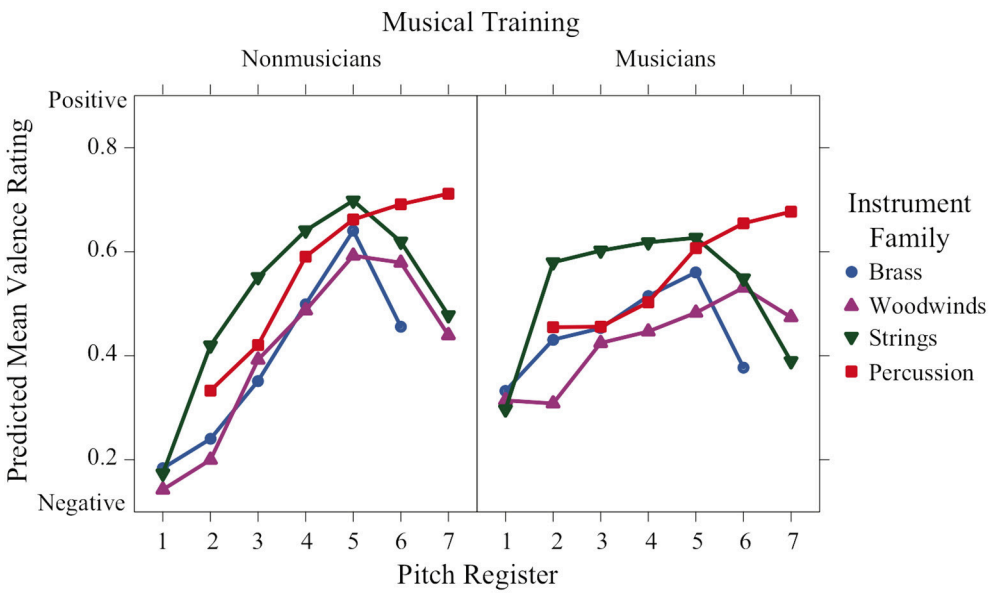


FIGURE 2 | Means of perceived valence ratings across pitch register for each instrument family and each musical training group.

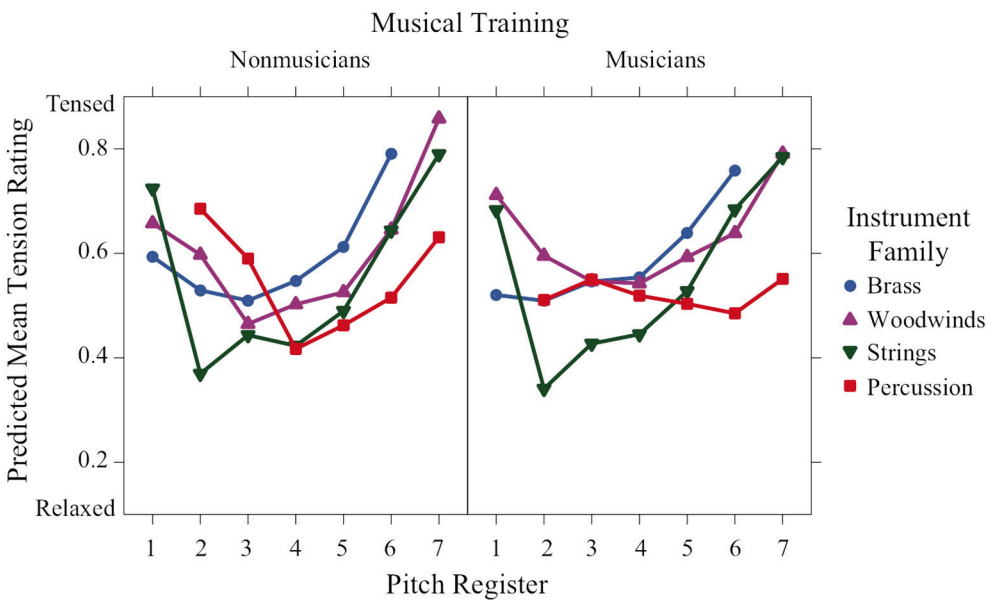


FIGURE 3 | Predicted means of perceived tension-arousal ratings across pitch register for each instrument family and training group.

brass, woodwinds, and strings for nonmusicians and for brass and strings for musicians. Woodwinds present an increasing linear trend for musicians as do percussion for both participant groups. The lack of quadratic trend in these latter cases leads to significant interactions between register and family and register and training. There was a main effect of family—strings > percussion > brass > woodwinds. There were valence peaks in the middle-high register with the exceptions of percussion in the higher registers for both groups and woodwinds in the highest octave for musicians as indicated by training × register and family × register interactions.

Tension-arousal ratings (**Figure 3**) were highly significant for register and followed a convex increasing form, with most families peaking at the lowest and highest octaves. There was a significant training × family × register interaction indicating different patterns across the two groups. The convex increasing trend was apparent for all families in the nonmusician training group, except that the percussion were convex decreasing.

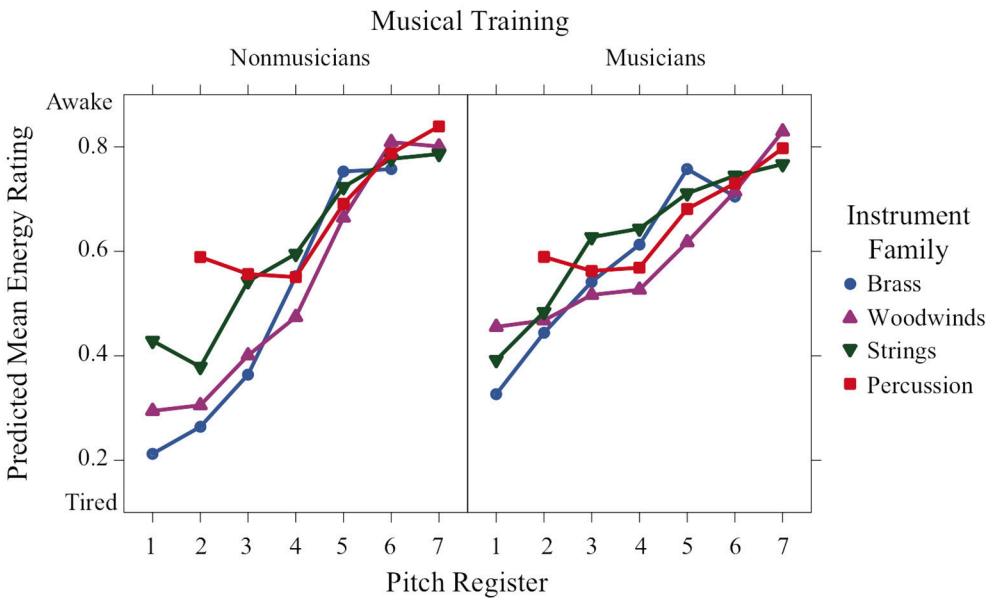


FIGURE 4 | Predicted means of perceived energy-arousal ratings across pitch register for each instrument family and training group.

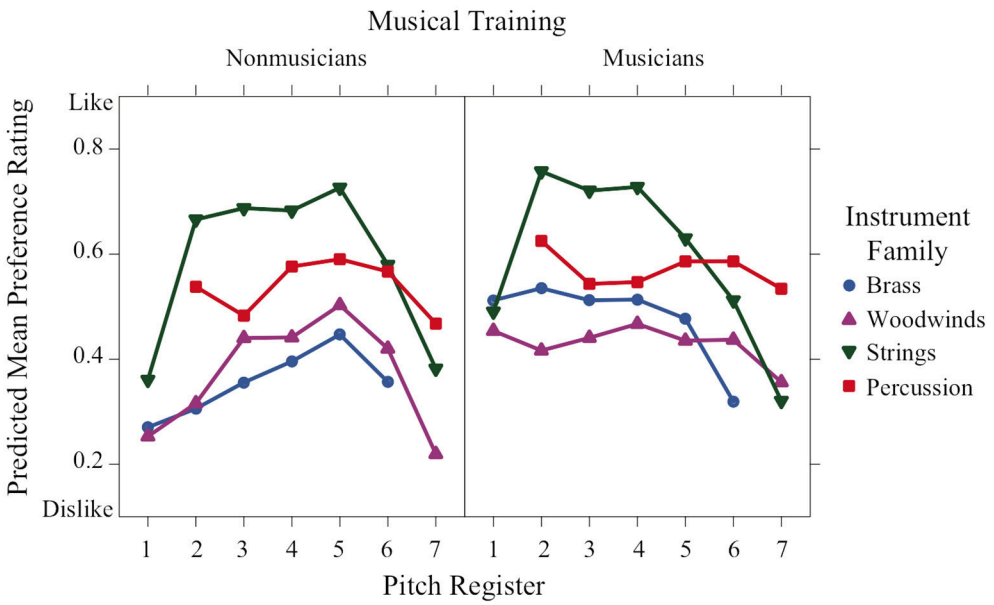


FIGURE 5 | Predicted means of preference ratings across pitch register for each instrument family and training group.

Musicians' ratings were similar for brass, woodwinds and strings, but ratings for the percussion family remained relatively neutral across all the registers as indicated by a lack of either linear or quadratic trend.

Register was a highly significant predictor for energy-arousal ratings (**Figure 4**), and a strong linear trend is visible across registers, with lower registers perceived as more tired and high registers perceived as more awake. Additionally, the register × family interaction was significant. This can be seen, specifically in the percussion family ratings in the second octave (the lowest octave for percussion sounds in this experiment), which were higher than the ratings of the other families in this octave, Welch's unequal variance $t > 5.02$, $p < 0.0001$. Furthermore, the significant training × family × register interaction can be seen when comparing the differences in energy-arousal ratings between the families in the low registers: in the nonmusician group, the families are more spread out in the first three octaves, whereas the ratings of the musician group are more similar across families, even in the low registers.

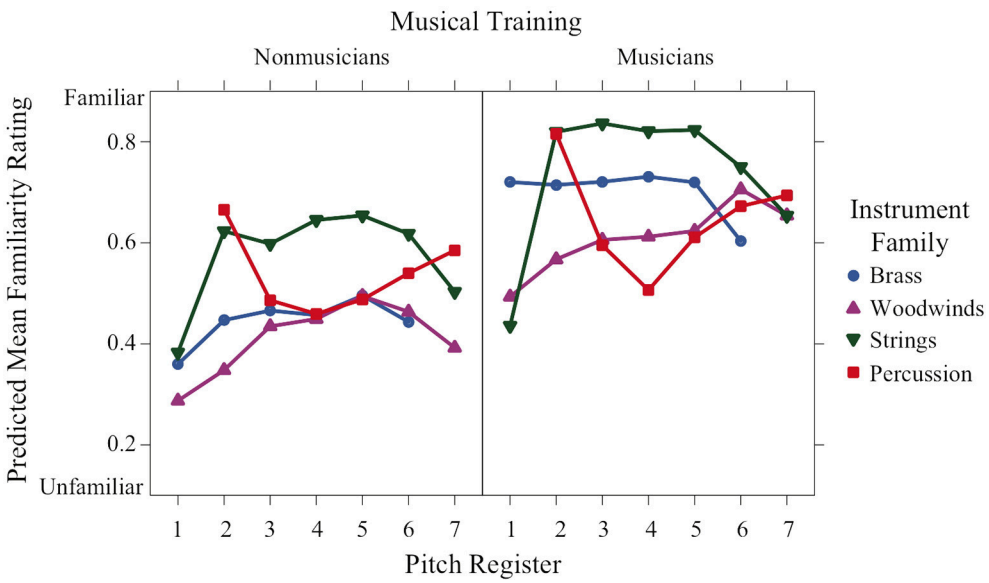


FIGURE 6 | Predicted means of familiarity ratings across pitch register for each instrument family and training group.

Similarities among instrument families were slightly less apparent in the graphs of preference and familiarity ratings compared to the perceived affect ratings. Instrument family was a significant predictor of preference ratings (**Figure 5**), and string sounds in mid-register octaves 2–5 were the most preferred by both musicians and nonmusicians, Welch's $t \geq 2.57$, $df \geq 110.63$, Bonferroni-corrected $p < 0.0021$, with three exceptions (strings vs. percussion for nonmusicians in octaves 2 and 4, and for musicians in octave 5). In line with a significant training × family × register interaction, the musicians' preference ratings for brass and woodwind sounds were relatively neutral at lower and mid-register pitches, then decreased in higher registers, whereas nonmusicians preference ratings for brass and woodwind sounds were low for low and high registers, but increased to a neutral rating around octave 5. This pattern in the nonmusicians' ratings is similar to that found in the perceived valence ratings.

Familiarity ratings varied significantly as a function of both training and family (**Figure 6**). Not surprisingly, they were significantly higher for the musician group than the nonmusician group. There was also a significant family × register interaction depicted by the higher ratings of string sounds in octaves 2–6 compared to the string sounds in octaves 1 and 7. This trend is inversed for percussion sounds, where the highest familiarity ratings occurred in the lowest and highest octaves (2 and 7, respectively).

Attack Strength and Playing Technique

To examine attack strength and playing technique as possible factors contributing to affect ratings, the dataset was separated by instrument family, and 20 linear mixed models were created: one for each of the five ratings for each of the four instrument families. All of the models included fixed factors of training and register. Attack strength was included in the brass and percussion models, and technique was included in the brass, woodwind and string models. This separation was necessary because of the lack of different attack strengths in the woodwind and string samples and the lack of different playing techniques for the percussion samples. As with the full models, a maximal random-effects structure was specified for each instrument family model. The brass familiarity, string preference and familiarity, and percussion familiarity models did not converge. Reducing the random effects structure by removing register as a random slope allowed the models to converge, although this removal increases likelihood of Type I errors. That being said, neither attack strength nor technique was a significant predictor for any individual instrument family model.

Summary

Valence ratings have a nonlinear concave relation to register, with more positive valence in the middle registers, apart from percussion for which valence progresses from negative to positive from the lowest to the highest register. The effect of register depends on both musical training and instrument family. Nonmusicians gave more negative ratings in the lowest registers compared to musicians. Globally strings have the most positive valence followed by percussion, brass, and woodwinds in order of increasingly negative valence.

Tension-arousal ratings have a nonlinear convex relation to register. Again, the effect of register depends on training and instrument family. Percussion sounds for musicians seem to be unaffected by register and remain at a middle level of tension arousal. Globally, brass receive the highest excitation ratings followed by woodwinds, percussion and strings, which were rated as calmest.

Energy arousal ratings increase monotonically with register, but again they depend on training and instrument family. The

ratings are quite similar across families for both groups of participants in octaves 4–7 for nonmusicians and across all octaves for musicians. In the lower octaves, the ratings are spread out for nonmusicians with percussion being rated as most awake followed by strings and then by woodwinds and brass, which are rated similarly.

Preference ratings have a nonlinear concave relation to register that depends on family and training and resemble valence ratings in their form. Familiarity ratings are higher for musicians than nonmusicians. They are concave with respect to register for strings, convex for percussion, and unaffected by register for brass and woodwinds.

Acoustic Descriptors

Due to its multidimensionality, it is necessary to account for multiple acoustic properties when examining timbre (McAdams et al., 1995). There are numerous audio descriptors derivable from the sound signal that can be categorized as spectral, temporal, or spectrotemporal properties of a sound. The following analysis investigates the relationship between the quantitative descriptors and the perceived affect ratings. The tool we use is the Timbre Toolbox (Peeters et al., 2011) as recently updated, corrected, and validated by Kazazis et al. (2016).

The Timbre Toolbox (Peeters et al., 2011) calculates temporal descriptors, such as attack time, spectral descriptors, such as spectral centroid, and spectrotemporal descriptors, such as spectral variation over time, in Matlab (The MathWorks Inc., Natick, MA). There are three stages of computation. First, the input representations of the signal are computed. The Timbre Toolbox has several input representations. The ones we used here included the temporal energy envelope and a Short-Term Fast-Fourier Transform (STFT) with a frequency scale transformed to a physiological scale related to the distribution of frequencies along the basilar membrane in the inner ear as modeled by a scale (ERB-rate) derived from the Equivalent Rectangular Bandwidth (Moore and Glasberg, 1983). To calculate the temporal energy envelope of a given audio signal, the amplitude of the analytic signal, i.e., the signal with no negative-frequency components (Smith, 2007), is given by the Hilbert transform of the audio signal. The amplitude of the analytic signal is then low-pass filtered with a third-order Butterworth filter with a cutoff frequency of 5 Hz, resulting in the temporal energy envelope input representation.

In the second stage of computation, scalar and time-series descriptors are extracted from different input representations. To estimate the attack portion of the signal, the "weakest-effort method" (Peeters, 2004) is applied so that thresholds to detect the start and end time of the attack are not fixed but determined as a proportion of the maximum of the signal's energy envelope. Log-attack time, attack slope, and temporal centroid are calculated from the temporal energy envelope input representation. Log-attack time is the log_{10} of the duration (in seconds) of the attack portion of the signal, and attack slope is the averaged temporal slope of the energy envelope during the attack portion of the signal. Additionally, the temporal centroid is a measure of the center of gravity of the energy envelope of the signal.

Each of the spectral descriptors is calculated from the ERB-transformed STFT representation with Hamming time window of 23.2 ms with a hop size of 5.8 ms, thereby giving a time series for each descriptor. As described by Peeters et al. (2011), spectral centroid is a measure of the center of mass of the spectrum and is perceptually related to the "brightness" of the sound. Spectral spread refers to the standard deviation of the spectrum around the spectral mean value and spectral skewness refers to the degree of asymmetry of the spectrum around the mean. Spectral kurtosis examines the flatness of the distribution around the mean value of the spectrum and can indicate a flat, normal, or peaky distribution. Spectral slope is a linear regression over the spectral amplitude values. Spectral decrease is the average of the set of spectral slopes between the fundamental frequency and the frequency of the kth harmonic. Spectral rolloff refers to the frequency below which 95% of the signal energy is contained. Spectral variation is a measure of the change in the spectral shape over time, quantified as one minus the normalized correlation between the spectra of successive time frames. Spectral flatness captures the noisiness of the signal and varies between completely "tonal" in the sense of being composed of clear, isolated frequency components and completely noisy. Spectral crest measures the degree of emergence of the most intense frequency component above the average amplitudes of the whole spectrum.

Finally, the third stage of computation considers the median and interquartile range (IQR) values of time-series descriptors to represent both central tendency and variability, respectively (Peeters et al., 2011). Time-series descriptors include spectral centroid, spread, skewness, kurtosis, slope, decrease, rolloff, variation, flatness, and crest. Adding the three temporal descriptors gives the 23 descriptors listed in **Table 3**.

Partial Least-Squares Regression

We completed a PLSR to examine the relation of the audio descriptors to the set of affect ratings. PLSR couples multiple linear regression with principle components analysis. Furthermore, we applied a five-fold cross-validation model to each PLSR in which the n cases are divided into five subsets, and the model is trained on four subsets and then predicts the remaining subset. The subsets are then rotated so that the training and prediction steps are applied to all combinations of the subsets. In addition to calculating R^2 as an evaluation of the model fitness, cross-validation also allows for the calculation of predictive relevance Q^2, the squared cross-validation prediction error summed across the five-folds (Wold et al., 2001).

We used the 23 Timbre Toolbox descriptors described above. The median values of spectral time series provide spectral information, and IQR measures as spectrotemporal information represent the variability of the descriptor over time. The PLSR and a subsequent correlation analysis were both completed in Matlab (The MathWorks Inc., Natick, MA).

An initial analysis of collinearity among descriptors across the complete sound set was performed. The 137 values for each descriptor were correlated with those of every other descriptor, and a hierarchical cluster analysis with average linkage was performed on the correlation matrix. The resulting dendrogram is shown in **Figure 7**. Several pairs of descriptors join at very

TABLE 3 | Definition of acoustic descriptors from the timbre toolbox.

	Acoustic descriptor	Definition	Derivative values
Spectral	Centroid (log)	Center of gravity of the spectrum	Med*, IQR*
	Spread	Standard deviation of the spectrum around the mean	Med, IQR*
	Skewness	Asymmetry of the spectrum around the mean	Med*, IQR*
	Kurtosis	Flatness of spectrum around the mean	Med, IQR
	Slope	Linear regression over the spectral amplitude values	Med, IQR
	Decrease	Average of slopes between F0 and 2nd to *k*th harmonic	Med*, IQR*
	Rolloff	Frequency below which 95% of the signal energy is contained	Med, IQR*
	Variation	Variation of the spectrum over time	Med*, IQR*
	Flatness	Ratio of the geometric and arithmetic means of the spectrum	Med*, IQR*
	Crest	Ratio of the spectral maximum to the arithmetic spectral mean	Med*, IQR*
Temporal	Attack time (log)	Duration of the attack portion of the sound	*
	Attack slope	Rate of change of energy over time in the attack portion	*
	Centroid	Center of gravity of the energy envelope	*

*For time-varying spectral descriptors, both the median (Med) and interquartile ranges (IQR) are computed over the duration of the sound, so each of these descriptors produces two measures. *Indicates the 17 descriptors included in partial least-squares regression and neural network analyses (see text).*

low levels indicating high collinearity. A PCA was conducted on the whole set of descriptor values for the 137 sounds. The PCA resulted in errors when including all 23 timbral descriptors plus the nominal pitch of the sounds, because the correlation matrix was not positive definite. Based on the hierarchical cluster analysis, six descriptors that were highly correlated with others were removed: Spectral Slope median and IQR, Spectral Spread median, Spectral Rolloff median, Spectral Kurtosis median, and IQR. These descriptors are highly correlated ($r > 0.905$) with Spectral Centroid median for Spectral Slope, Spread, and Rolloff medians, with Spectral Centroid IQR for Spectral Slope IQR, and with Spectral Skewness median and IQR for Spectral Kurtosis median and IQR, respectively. When these six were removed, the PCA gave a strong Kaiser-Mayer-Olkin index of 0.691, and the removal didn't much affect the total variance explained by the PCA (reduction by 2.2% of the variance explained) or its dimensionality (five components in both cases). Removing pitch as a factor reduced the explained variance by <1%, so it was not included in subsequent analyses either.

The PLSR was thus completed with the group of 17 measures (independent variables) shown with asterisks in **Table 3**. It was conducted for each of three dependent variables: mean ratings across participants of valence, tension arousal, and energy arousal for each of the 137 stimuli. Based on a threshold eigenvalue of 1, the procedure selected three principal components (PC) for valence and energy arousal, and four components for tension arousal. The upper and lower benchmarks of the model are measured by R^2 (explanatory power) and Q^2 (predictive power). We also computed the root mean squared error (MSE) between the mean ratings and the PLSR estimates. These values are displayed in **Table 4**. The valence and energy arousal ratings are better modeled than the tension arousal ratings; although all three models have low RMSE. The loadings of each descriptor on each PC are listed in **Table 5**. They can be interpreted as vector coordinates of the 17 predictors in the three- or four-dimensional spaces of the PCs or as their contributions to each PC.

Valence

Increases in PC1 (58% of the variance explained) are primarily associated with increasing spectral centroid and spectral crest medians, decreasing spectral decrease median, increasing spectral crest variability (IQR), increasing attack slope, and decreasing log attack time and temporal centroid. PC2 (6%) loadings show a positive effect of spectral skewness median and negative effects of spectral centroid and spectral flatness medians. PC3 (3%) is more positive with decreasing spectral variation median and with decreasing log attack time.

Tension Arousal

PC1 (39%) increases with increasing spectral centroid median and decreasing spectral skewness and spectral decrease medians, with additional negative contribution of spectral decrease IQR. PC2 (9%) increases with increasing spectral variation median and IQR. PC3 (5%) increases with increasing attack slope and with decreasing log attack time and temporal centroid. PC4 (3%) increases with decreasing spectral decrease IQR.

Energy Arousal

PC1 (72%) increases with increasing spectral centroid and crest medians, with decreasing spectral skewness and spectral decrease medians, with increasing spectral crest IQR, and with decreasing spectral decrease IQR and temporal centroid. PC2 (2%) increases with decreasing spectral flatness median and decreasing spectral slope IQR. PC3 (2%) increases with increasing variability in spectral flatness and increasing log attack time.

Globally, medians of all the time-varying spectral parameters play a stronger role than do measures of their variability or the temporal parameters, although the latter two groups make a significant contribution. The acoustic underpinnings of emotion portrayal by musical instrument sounds thus seem to result from a complex interplay of spectral, temporal, and spectrotemporal factors.

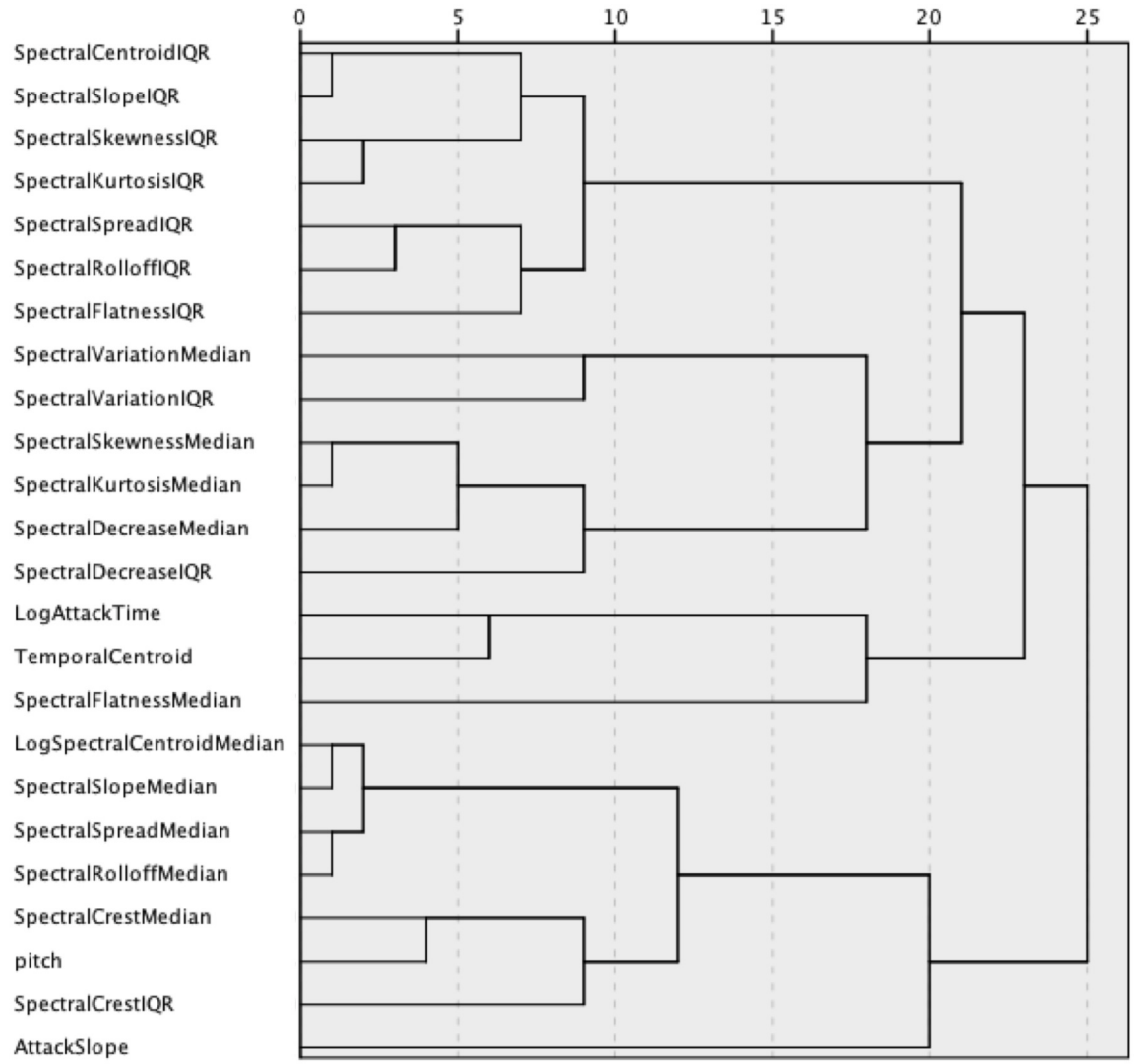


FIGURE 7 | Dendrogram of the hierarchical cluster analysis of descriptor values across the entire stimulus set.

TABLE 4 | R^2, Q^2, and RMSE results for the PLSR models predicting perceived valence, tension arousal, and energy arousal.

	Valence	Tension arousal	Energy arousal
R^2	0.6617	0.5605	0.7535
Q^2	0.6032	0.4406	0.6867
RMSE	0.0827	0.0841	0.0772

Neural Network Model

Previous research has shown that non-linear methods can be particularly useful in situations where linear methods are insufficient to model the relationship between dependent and independent variables. Artificial neural networks may be used as a non-linear regression method (Coutinho and Cangelosi, 2011; Russo et al., 2013; Vempala and Russo, 2013) to predict valence, tension arousal, and energy arousal ratings using timbre descriptors.

We used supervised feedforward networks with back propagation (i.e., multilayer perceptrons) for this purpose (Bishop, 1996; Haykin, 2008; Rumelhart et al., 1986). We built three types of prediction networks—one for valence, one for tension arousal, and one for energy arousal in Matlab. Each type of network consisted of one input layer with 17 units corresponding to the descriptors with asterisks in **Table 3**, one hidden layer with 3 units, and one output unit corresponding to the mean valence, tension arousal or energy arousal value of participants for that stimulus. Separate networks were trained with valence, tension arousal, or energy arousal as the output unit.

Our training paradigm involved five-fold cross-validation to avoid over fitting the network to any specific partitioning of the training and test sets. To enable cross-validation, we partitioned

TABLE 5 | Loadings of each audio descriptor on the Principal Components (PC) for each emotion dimension in the PLSR analysis.

	Valence			Tension Arousal				Energy Arousal		
	PC1	PC2	PC3	PC1	PC2	PC3	PC4	PC1	PC2	PC3
Spec centroid Med	6.18	**-8.53**	1.55	**10.39**	−3.95	1.81	−0.73	**9.21**	−5.05	0.89
Spec centroid IQR	3.25	−4.45	5.35	2.43	−1.67	−2.07	2.83	3.51	−6.72	3.90
Spec spread IQR	1.46	−2.92	4.27	−0.43	1.38	−2.36	−0.30	0.97	−3.63	5.55
Spec skewness Med	−5.88	**8.74**	−1.89	**−10.29**	3.98	−1.07	0.89	**−8.91**	5.92	−0.64
Spec skewness IQR	0.67	−0.41	4.96	−2.29	0.41	−2.28	3.76	−0.49	−3.50	5.20
Spec decrease Med	**−8.10**	4.88	−1.12	**−8.04**	5.88	−3.77	1.60	**−10.03**	0.86	−0.73
Spec decrease IQR	−4.38	3.96	−1.83	−6.82	4.41	−2.93	**−6.32**	−6.21	3.24	−1.61
Spec rolloff IQR	1.20	−1.39	4.61	−1.59	0.50	−2.34	2.26	0.30	−3.21	5.87
Spec variation Med	−2.18	−4.58	**−6.63**	0.99	**8.79**	4.41	−3.74	−1.40	2.32	2.38
Spec variation IQR	−4.81	−5.62	−3.20	1.68	**9.33**	−0.30	−2.94	−3.53	−1.65	2.26
Spec flatness Med	−2.10	−6.56	2.56	4.67	0.23	−5.97	−1.65	−0.07	**−9.26**	−1.43
Spec flatness IQR	1.87	−3.27	3.87	0.92	1.77	−1.54	−2.20	1.87	−1.55	**6.58**
Spec crest Med	**8.20**	−0.22	−0.05	3.89	−5.53	4.97	−3.79	**8.65**	5.09	0.75
Spec crest IQR	6.81	−1.46	1.50	2.38	−3.54	1.85	−5.53	6.87	2.47	1.77
Log attack time	−7.47	−3.19	6.05	4.49	4.06	−7.34	2.65	−4.85	−3.15	**6.56**
Attack slope	7.99	3.02	−5.67	−4.36	−3.72	**7.71**	−3.36	5.35	3.93	−5.19
Temporal centroid	**−8.38**	−1.59	2.56	2.65	3.56	−6.34	4.28	−6.29	−3.41	1.54
Partial R^2	0.58	0.06	0.03	0.39	0.09	0.05	0.03	0.72	0.02	0.02

Loadings greater that 8.0 (or the highest value for a given PC if none are above 8) are shown in bold to highlight the primary contributors discussed in the text.

135 of the 137 stimuli in our dataset into five equal sets of 27. For each fold, we tested the network on the 27 stimuli within that fold along with the remaining two unused stimuli, after training the network on the four additional folds (i.e., 108 stimuli; see Table S1).

As is common in neural net modeling, all input descriptors were range-normalized in the interval [0, 1] to allow the network to maximize performance by capturing similarities and differences within and across descriptors for all examples of the training set. Connection weights from the input layer to the hidden layer and from the hidden layer to the output unit were initialized to random values between –0.05 and 0.05, allowing for optimal adjustment of hidden units during training. Outputs at the hidden layer were computed using sigmoid functions. The sigmoid or logistic sigmoid function is commonly used in multilayer perceptrons because it has desirable properties. It transforms the data nonlinearly while limiting the range of values between 0 and 1, thus acting as a useful squashing function. For each training stimulus, the squared error between the network's predicted output and the mean participant rating was computed. Changes to connection weights over successive epochs were computed using back propagation of errors with gradient descent and were then stored. After completion of each epoch (i.e., 108 training stimuli), connection weights were updated with the sum of the stored weight changes.

While having more hidden units helps the network converge earlier, and reduces the MSE, it also results in the network becoming overfitted to the training set, thus reducing the network's generalizability. Hence, after initial simulations starting with 7 hidden units, we progressively reduced the number of hidden units down to 3 units, upon noticing that the network's performance was still high for all five-folds of cross-validation. Each network was tested on the set of 29 stimuli. We computed the RMSE as a measure of performance.

TABLE 6 | Performance of neural networks modeling valence, tension arousal, and energy arousal.

Network	RMSE		
	Valence	Tension arousal	Energy arousal
1	0.0800	0.0814	0.0703
2	0.0801	0.0813	0.0672
3	0.0809	0.0803	0.0672
4	0.0816	0.0853	0.0687
5	0.0825	0.0822	0.0680
Mean	0.0810	0.0821	0.0683

The valence and energy arousal networks were trained for 700 epochs, but the tension network took longer to converge, requiring 1000 epochs. All networks successfully converged to a MSE of <0.008. Cross-validation performance for each type of output unit is reported for each of the five-folds along with the mean performance in **Table 6**.

To get a sense of which of the timbre predictors were important for each dependent variable, we used a method developed by Milne (1995). This method computes the size and sign of each feature's contribution to the output by taking into account the connection weights from that feature to the hidden layer, and from the hidden layer to the output unit. We computed feature contributions and averaged them across the

TABLE 7 | Primary timbre descriptor contributions to each type of neural net output unit.

Feature	Valence	Tension arousal	Energy arousal
Spectral centroid median	–	+8.6	+8.3
Spectral spread IQR	–	−7.0	–
Spectral decrease median	−9.5	–	−9.6
Spectral rolloff IQR	–	+6.7	–
Spectral variation median	−10.6	+15.9	–
Spectral variation IQR	−11.0	–	−12.1
Spectral flatness median	–	−7.2	–
Spectral crest median	+10.9	–	+7.7
Spectral crest IQR	–	–	+15.2
Attack slope	+7.9	−7.1	–
Temporal centroid	−15.8	–	−13.9

The values indicate the mean % contribution and the sign of the relation between the model prediction and the audio descriptor.

five networks used for cross-validation. The mean contribution proportions of the top six timbre features for perceived valence, tension arousal, and energy arousal are reported as percentages in **Table 7**, along with their signs. A negative sign indicates that increases in the value of the feature are associated with decreases in the emotion dimension. Different combinations of spectral, temporal, and spectrotemporal descriptors contribute to the neural network modeling of valence, tension arousal, and energy arousal. Although, at some level all audio descriptors make some contribution, the primary contributors differ across the rating scales, suggesting acoustic independence among them.

DISCUSSION

We explored the perceived emotional qualities of 137 isolated tones played by standard western orchestral instruments across their entire pitch ranges and using different playing techniques. One novel aspect of this study on the role of timbre in perceived emotion is that both instrument family and pitch register were varied. It is important to recognize that register affects the timbre of notes produced by each instrument in the sense that several spectral measures are strongly or very strongly correlated with pitch octave [in decreasing order: spectral crest, $r_{(135)} = 0.826$; spectral decrease, $r_{(135)} = -0.816$; spectral centroid, $r_{(135)} = 0.690$; and spectral skewness, $r_{(135)} = -0.637$]. Therefore, timbre varies with register, but not directly as a function of fundamental frequency. The aim was to determine the acoustic properties related to timbre that contribute to ratings by musician and nonmusician listeners on continuous scales of the emotional qualities valence (on both positive vs. negative and pleasure vs. displeasure scales), tension arousal and energy arousal. Listeners also rated preference for and familiarity with each sound. We first discuss the rating data as a function of pitch register, instrument family, and the musical training of the listeners. We then discuss the two approaches to modeling the data with linear PLSR and nonlinear neural nets.

Listener Ratings (Ground Truth)

The two valence scales were very strongly correlated [$r_{(135)} = 0.97$], and so subsequent analyses were limited to the positive/negative scale. It is worth recalling that using musical excerpts of varying, but not systematically controlled instrumentation, Bigand et al. (2005) found no correlation between pleasantness and positive-negative valence. So other musical properties may distinguish these two dimensions of musical experience.

Linear mixed effects models that take into account variation due to participants and stimulus items revealed strong interactions of the factors pitch register, instrument family, and musical training for all rating scales. Valence and preference had nonlinear concave relations to register indicating that maximally positive valence and preference corresponded to middle registers. The exception to this pattern for valence was the percussion family, which had a monotonic increasing relation to register. The families in order from negative to positive valence were woodwinds, brass, percussion, and strings. Tension arousal had a nonlinear convex relation to register except for the percussion family for musicians, which had a medium tension level across registers. The families in order of decreasing tension were brass, woodwinds, percussion, and strings. Energy arousal had a monotonic relation to register, with only small differences among the families in the lower registers for nonmusicians. Familiarity ratings were higher for musicians than nonmusicians and were concave with respect to register for strings, convex for percussion and unaffected by register for brass and woodwinds. These results can be compared to those of Eerola et al. (2013) who manipulated several musical parameters on musical phrases, including timbral brightness (flute, horn, trumpet in order of increasing spectral centroid) and pitch height (from F3 to B5 in 6-semitone steps, which corresponds to the middle 2.5 octaves of our 6-octave range). These authors found linear contributions of both timbre and register to ratings of "scary" (increasing with brightness, decreasing with register), "sad" (decreasing with brightness and register), and "peaceful" (decreasing with brightness, increasing with register), but not of "happy." They also found slight quadratic contributions of register to ratings of "scary" (convex) and "peaceful" (concave), but not "happy" or "sad." It is difficult to compare directly these two sets of results, one being in a dimensional framework and the other in a categorical framework, but they both emphasize the complex mapping of emotion onto these musical parameters. One does note, however, that a stronger nonlinearity appears with a wider range of pitch heights.

Musical training interacted with register and instrument family for all three emotion dimension ratings, and preference ratings as well. It only interacted with family for familiarity ratings. Regarding the valence ratings, musicians tended to perceive low-register sounds as less negative than nonmusicians. For tension arousal ratings, nonmusicians had convex curves as a function of register for all families, whereas for musicians only woodwinds and strings had this form; percussion were unaffected by register, and tension increased monotonically with register for brass. Energy arousal ratings were globally less affected by instrument family and musical training, with the

notable exception of results in lower registers where differences in perceived energy arousal between families were found for nonmusicians. As expected, familiarity ratings were higher in the musician group and varied across instrument family. Musicians are more familiar with sounds in extreme registers than nonmusicians, and this familiarity could potentially play a role in the perceived affect ratings. These differences between musical training groups differ from the finding of Filipic et al. (2010) who found no such difference with short musical clips.

In our experiment, there was a moderately strong positive correlation between the perceived valence ratings and preference, but the negative correlation between perceived tension-arousal ratings and preference was slightly stronger. Although, participants typically preferred more positive, less tense timbres, this finding demonstrates that there is not a clear one-to-one relationship between positive valence or tension and listener's preference. Furthermore, pitch register significantly influenced both perceived valence and tension-arousal ratings so that mid-register sounds were rated as more positive and more relaxed than sounds of an extreme high or low register.

We confirmed that listeners can consistently rate the perceived affect of individual sounds from different musical instruments across their pitch registers with short sounds (500 ms). These results are in accordance with those of Eerola et al.'s (2012) Experiment 1 and other studies utilizing short musical samples (Peretz et al., 1998; Bigand et al., 2005; Filipic et al., 2010) in which the participants were able to rate perceived affect in 1-s or 500-ms instrumental music samples with great consistency.

There were a few key differences between our results and those of Ilie and Thompson (2006), on the one hand, and those from Eerola et al.'s (2012) Experiment 1 on the other. First, the tension-arousal and energy-arousal ratings were only weakly, although significantly, correlated in the present study and Ilie and Thompson's, whereas they were strongly correlated in Eerola et al.'s study. As the energy-arousal dimension had a monotonic relation to register, listeners seem to have used primarily spectral cues when making energy ratings and incorporated additional acoustic information when making tension ratings (see discussion of audio descriptors below). Furthermore, valence and preference ratings in this experiment were moderately correlated, whereas they were strongly correlated in Eerola et al.'s study. Listeners did not necessarily prefer sounds with the highest perceived valence. We therefore concur with Ilie and Thompson in emphasizing the importance of differentiating these measures in a larger stimulus context. One crucial difference is the use of a single pitch in Eerola et al. compared to the whole range of registers for each instrument in the present study. This would mean that the sounds from some instruments in their study would be in their extreme high or low registers. Pitch octave in the present study was strongly correlated with valence, $r_{(135)} = 0.624$, $p < 0.0001$, weakly correlated with tension arousal, $r_{(135)} = 0.242$, $p = 0.004$, and very strongly correlated with energy arousal, $r_{(135)} = 0.849$, $p < 0.0001$. So the differential effect of pitch register on the two arousal scales seems to further distinguish them in the current study. Furthermore, the linear mixed model analysis showed that the energy-arousal ratings were strongly influenced by pitch register, and unlike the tension-arousal ratings, were not significantly influenced by instrument family. This finding is a significant contribution to affect and timbre research because it shows that the two arousal dimensions are distinctly perceivable in timbre and not interchangeable, as they are influenced by different factors.

Linear and Nonlinear Modeling of the Acoustic Basis for Perceived Emotion Dimensions

To analyze the contribution to the emotion ratings of acoustic properties related to timbre, we examined 23 acoustic signal parameters taken from the Timbre Toolbox (Peeters et al., 2011), spanning spectral, temporal, and spectrotemporal audio descriptors. Initial hierarchical clustering and principal components analyses suggested reducing these to 17 descriptors due to high collinearity. It is interesting to note that the pitch height descriptor did not make a significant contribution as it was highly collinear with several spectral descriptors, underlining again the fact that timbre and pitch covary strongly in many acoustic musical instruments. A PLSR with these 17 descriptors as predictors of each of the three emotion dimensions allowed us to reduce the dimensionality to three principal components for valence and energy arousal and to four principal components for tension arousal. Additionally, a nonlinear neural network multilevel perceptron model (NN) was programmed with 17 inputs represented by the audio descriptors, three hidden units, and a single output separately modeling the three mean emotion dimension ratings. In both cases, a five-fold cross-validation method was used to estimate the reliability of the models. Several measures compare ground truth values (means across participants for a given emotional dimension) to predicted values (**Table 8**). Model fitness (R^2) is computed on items in the training set (the 4 groups excluding the test set, see Table S1) for each fold and then averaged across the five-folds. The model's predictive power (Q^2) is computed on the five training sets collectively taken across the five-folds. The prediction error (RMSE) is computed on both training and test sets for each fold and is then averaged across the five-folds. The percent improvement of the NN model over the PLSR model is shown in **Table 8**, in which a positive sign indicates higher values for NN. One notes that a much better fit is obtained for all three emotion dimensions with the NN models than with the PLSR models (32–78% improvement), as they are all near perfect prediction. The predictive power across training sets is more equivalent for the two techniques, but it still shows more than 10% improvement for the two arousal dimensions with the NN model. The prediction error is roughly equivalent in the two models for valence and tension arousal, and although better for energy arousal with the PLSR model, the NN model still shows 12% improvement (lower error) for this emotion dimension. The nonlinear approach would thus seem to have modest modeling advantages over the linear approach.

Table 9 presents the primary audio descriptors that contribute to each emotion dimension model for each technique. The ranks of the six descriptors providing the highest loadings for PLSR or highest percent contribution for NN are shown. Descriptors that make major contributions to both models are highlighted

TABLE 8 | Comparison of model fitness, predictive power, and prediction error for PLSR and neural network models.

Method	R^2			Q^2			RMSE		
	Valence	Tension	Energy	Valence	Tension	Energy	Valence	Tension	Energy
PLSR	0.6617	0.5605	0.7535	0.6032	0.4406	0.6867	0.0827	0.0841	0.0772
NN-MLP	0.9971	0.9963	0.9983	0.6117	0.4870	0.7658	0.0810	0.0821	0.0683
Percentage of improvement (%)	51	78	32	1	11	12	−2	−2	−12

TABLE 9 | Ranks of primary audio descriptors contributing to PLSR and NN models.

Audio Descriptor	Type	Valence		Tension Arousal		Energy Arousal	
		PLSR	NN	PLSR	NN	PLSR	NN
Spectral centroid median	Spectral	2	–	1	2	3	5
Spectral centroid IQR	Spectrotemporal	–	–	–	–	6	–
Spectral spread IQR	Spectrotemporal	–	–	–	5	–	–
Spectral skewness median	Spectral	1	–	2	–	4	–
Spectral skewness IQR	Spectrotemporal	–	–	–	–	–	–
Spectral decrease median	Spectral	5	5	5	–	1	4
Spectral decrease IQR	Spectrotemporal	–	–	–	–	–	–
Spectral rollof IQR	Spectrotemporal	–	–	–	6	–	–
Spectral variation median	Spectrotemporal	–	4	4	1	–	–
Spectral variation IQR	Spectrotemporal	–	2	3	–	–	3
Spectral flatness median	Spectral	–	–	–	3	2	–
Spectral flatness IQR	Spectrotemporal	–	–	–	–	–	–
Spectral crest median	Spectral	4	3	–	–	5	6
Spectral crest IQR	Spectrotemporal	–	–	–	–	–	1
Log attack time	Temporal	–	–	–	–	–	–
Attack slope	Temporal	6	6	6	4	–	–
Temporal centroid	Temporal	3	1	–	–	–	2

PLSR is in green and NN in orange. Descriptors that make major contributions to both models are highlighted with darker colors.

with darker colors. Note firstly that different combinations of audio descriptors make major contributions to the three emotion dimensions, suggesting that they are carried by distinct acoustic properties. Valence is primarily carried by spectral and temporal properties. It is more positive with lower spectral slopes (more high-frequency energy), a greater emergence of strong partials, and an amplitude envelope with a sharper attack and earlier decay. To the contrary, Eerola et al. (2012) found more positive valence ratings with sustained sounds having more low-frequency energy, and Ilie and Thompson (2006) found more positive valence for lower-register sounds. Tension arousal ratings are driven by all three types of descriptors. Higher tension is carried by brighter sounds, more spectral variation and more gentle attacks. This result is coherent with Ilie and Thompson's finding that increase pitch height is associated with increase tension arousal. Energy arousal seems primarily spectral in nature, and greater energy is associated with brighter sounds with higher spectral centroids and slower decrease of the spectral slope, as well as with a greater degree of spectral emergence. The spectral aspect echoes Eerola et al.'s result showing this emotion dimension to be associated with more dominant high-frequency components, although those authors also found sharper attacks to be associated with higher ratings of energy arousal. Ilie and Thompson found no effect of pitch height (and concomitantly, spectral distribution) on energy arousal. The factors that distinguish these three studies is that ours covers a much wider range of pitches, thus augmenting the role played by pitch height and its concomitant timbral attributes related to spectral properties.

CONCLUSION

This study examined timbre and its complex covariation with pitch as musical elements capable of conveying emotion information. It highlights the fact that changes in pitch are accompanied by significant changes in timbral properties as quantified by timbral audio descriptors. It also demonstrates the fact that different intrinsic emotional qualities of musical instrument sounds are carried by different, but overlapping, sets of acoustic dimensions, suggesting that it is their complex combination that specifies emotional tone. This work provides a foundation for work on the acoustic underpinnings of perceived emotion in musical sound that could stimulate additional work

in music informatics by providing tools for including timbre in content-based approaches to automatic identification of mood in music (Kim et al., 2010). Future research should apply these results to increasingly ecological studies to validate the relationship between timbre, pitch, and perceived affect in a musical context and examine how that relationship interacts with additional relationships between perceived affect and other musical variables such as dynamics, tempo, harmony and texture. But even on its own, this work provides a rough map of how sounds produced by musical instruments in given registers relate to perceived emotional tone, suggesting basic acoustic characteristics upon which composers capitalize in sculpting musical experience.

ETHICS STATEMENT

All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was certified for ethics compliance by the McGill University Research Ethics Board II.

AUTHOR CONTRIBUTIONS

SM and CD conceived the experiments and interpreted the analyses of behavioral data. CD conducted the experiments and performed the analyses of behavioral data. SM, CD, and NV conducted acoustics analyses and modeling and contributed to the writing of the paper.

FUNDING

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2015-05028) and a Canada Research Chair (950-223484) awarded to SM.

ACKNOWLEDGMENTS

Portions of this work were presented at the 13th International Conference on Music Perception and Cognition, Seoul, August 2014. The authors would like to thank Bennett K. Smith for technical advice and help with preparation of the stimuli.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2017.00153/full#supplementary-material

REFERENCES

Adler, S. (2002). *Study of Orchestration, 3rd Edn.* New York, NY: W. W. Norton.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear Mixed-Effects Models Using Eigen and S4.* R Package Version 1.0–5. Available online at: http://CRAN.R-project.org/package=lme4

Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., and Dacquet, A. (2005). Multidimensional scaling of emotional responses to music: the effect of musical expertise and of the duration of the excerpts. *Cogn. Emot.* 19, 1113–1139. doi: 10.1080/02699930500204250

Bishop, C. M. (1996). *Neural Networks for Pattern Recognition*. Oxford, UK: Oxford University Press.

Coutinho, E., and Cangelosi, A. (2011). Musical emotions: predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements. *Emotion* 11, 921–937. doi: 10.1037/a0024700

Eerola, T., Ferrer, R., and Alluri, V. (2012). Timbre and affect dimensions: evidence from affect and similarity ratings and acoustic correlates of isolated instrument sounds. *Music Percept.* 30, 49–70. doi: 10.1525/mp.2012.30.1.49

Eerola, T., Friberg, A., and Bresin, R. (2013). Emotional expression in music: contribution, linearity, and additivity of primary musical cues. *Front. Psychol.* 4:487. doi: 10.3389/fpsyg.2013.00487

Eerola, T., Lartillot, O., and Toiviainen, P. (2009). "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)* (Kobe), 621–626.

Filipic, S., Tillmann, B., and Bigand, E. (2010). Judging familiarity and emotion from very brief musical excerpts. *Psychon. Bull. Rev.* 17, 335–341. doi: 10.3758/PBR.17.3.335

Fox, J., and Weisberg, S. (2011). *An R Companion to Applied Regression 2nd Edn*. Thousand Oaks, CA: Sage.

Frick, R. W. (1985). Communicating emotion: the role of prosodic features. *Psychol. Bull.* 97, 412–429. doi: 10.1037/0033-2909.97.3.412

Gabrielsson, A., and Lindström, E. (2010). "The influence of musical structure on emotional expression," in *Handbook of Music and Emotion: Theory, Research, Application*, eds P. N. Juslin and J. A. Sloboda (New York, NY: Oxford University Press), 367–400.

Geladi, P., and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Anal. Chim. Acta* 185, 1–17. doi: 10.1016/0003-2670(86)80028-9

Hailstone, J. C., Omar, R., Henley, S. M., Frost, C., Kenward, M. G., and Warren, J. D. (2009). It's not what you play, it's how you play it: timbre affects perception of emotion in music. *Q. J. Exp. Psychol.* 62, 2141–2155. doi: 10.1080/17470210902765957

Haykin, S. (2008). *Neural Networks and Learning Machines*. Upper Saddle River, NJ: Prentice Hall.

Hexler.net (2011). *TouchOSC: Modular OSC and MIDI Control Surface for iPhone, iPod Touch, and iPad.* Available online at: http://hexler.net/

Holmes, P. A. (2011). An exploration of musical communication through expressive use of timbre: the performer's perspective. *Psychol. Music* 40, 301–323. doi: 10.1177/0305735610388898

Huron, D. (2008). A comparison of average pitch height and interval size in major- and minor-key themes: evidence consistent with affect-related pitch prosody. *Empir. Musicol. Rev.* 3, 59–63. Available online at: http://hdl.handle.net/1811/31940

Huron, D., Anderson, N., and Shanahan, D. (2014). You can't play a sad song on the banjo: acoustic factors in the judgment of instrument capacity to convey sadness. *Empir. Musicol. Rev.* 9, 29–41. doi: 10.18061/emr.v9i1.4085

Huron, D., Kinney, D., and Precoda, K. (2006). Influence of pitch height on the perception of submissiveness and threat in musical passages. *Empir. Musicol. Rev.* 1, 170–177. Available online at: http://hdl.handle.net/1811/24068

Ilie, G., and Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Percept.* 23, 319–329. doi: 10.1525/mp.2006.23.4.319

ISO (2004). *Acoustics – Reference Zero for the Calibration of Audiometric Equipment – Part 8: Reference Equivalent Threshold Sound Pressure Levels for Pure Tones and Circumaural Earphones (ISO 389–8)*. Technical report, International Organization for Standardization, Geneva.

Juslin, P. N., and Laukka, P. (2003). Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol. Bull.* 129:770. doi: 10.1037/0033-2909.129.5.770

Juslin, P. N., and Laukka, P. (2004). Expression, perception, and induction of musical emotions: a review and a questionnaire study of everyday listening. *J. New Music Res.* 33, 217–238. doi: 10.1080/0929821042000317813

Juslin, P. N., and Västfjäll, D. (2008). Emotional responses to music: the need to consider underlying mechanisms. *Behav. Brain Sci.* 31, 559–575. doi: 10.1017/S0140525X08005293

Kazazis, S., Esterer, N., Depalle, P., and McAdams, S. (2016). "Testing the robustness of the Timbre Toolbox and the MIRtoolbox," in *Proceedings of the 14th International Conference for Music Perception and Cognition,* ed T. Zanto (San Francisco, CA: University of California), 24.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scot, J., et al. (2010). "Music emotion recognition: a state of the art review," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)* (Utrecht), 255–266.

Krumhansl, C. L. (2002). Music: a link between cognition and emotion. *Curr. Dir. Psychol. Sci.* 11, 45–50. doi: 10.1111/1467-8721.00165

Kumar, S., Forster, H. M., Bailey, P., and Griffiths, T. D. (2008). Mapping unpleasantness of sounds to their auditory representation. *J. Acoust. Soc. Am.* 124, 3810–3817. doi: 10.1121/1.3006380

Lartillot, O., and Toiviainen, P. (2007). "A Matlab toolbox for musical feature extraction from audio," in *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx)* (Bordeaux), 237–244.

Lenth, R. V. (2013). *lsmeans: Least-Squares Means. R Package Version 2.00-4.* Available online at: https://CRAN.R-project.org/package=lsmeans (Accessed 18 October, 2016).

Marozeau, J., de Cheveigné, A., McAdams, S., and Winsberg, S. (2003). The dependency of timbre on fundamental frequency. *J. Acoust. Soc. Am.* 114, 2946–2957. doi: 10.1121/1.1618239

Martin, F. N., and Champlin, C. A. (2000). Reconsidering the limits of normal hearing. *J. Am. Acad. Audiol.* 11, 64–66.

McAdams, S. (1993). "Recognition of sound sources and events," in *Thinking in Sound: The Cognitive Psychology of Human Audition*, eds S. McAdams and E. Bigand (Oxford, UK: Oxford University Press), 146–198.

McAdams, S., and Goodchild, M. (forthcoming). "Musical structure: sound and timbre," in *Routledge Companion to Music Cognition*, eds R. Ashley and R. Timmers (New York, NY: Routledge).

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. *Psychol. Res.* 58, 177–192. doi: 10.1007/BF00419633

Milne, L. (1995). "Feature selection using neural networks with contribution measures," in *AI'95, Australian Joint Conferences on Artificial Intelligence, Canberra*. Singapore, World Scientific.

Moore, B. C., and Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *J. Acoust. Soc. Am.* 74, 750–753. doi: 10.1121/1.389861

Peeters, G. (2004). *A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project*. Technical Report, IRCAM, Paris, France.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (2011). The timbre toolbox: extracting audio descriptors from musical signals. *J. Acoust. Soc. Am.* 130, 2902–2916. doi: 10.1121/1.3642604

Peretz, I., Gagnon, L., and Bouchard, B. (1998). Music and emotion: perceptual determinants, immediacy, and isolation after brain damage. *Cognition* 68, 111–141. doi: 10.1016/S0010-0277(98)00043-2

Risset, J.-C., and Wessel, D. L. (1999). "Exploration of timbre by analysis and synthesis." in *The Psychology of Music*, ed D. Deutsch (San Diego, CA: Academic Press), 113–169.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Rumsey, F., Zieliński, S., Kassier, R., and Bech, S. (2005). On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. *J. Acoust. Soc. Am.* 118, 968–976. doi: 10.1121/1.1945368

Russo, F. A., Vempala, N. N., and Sandstrom, G. M. (2013). Predicting musically induced emotions from physiological inputs: linear and neural network models. *Front. Psychol.* 4:468. doi: 10.3389/fpsyg.2013.00468

Scherer, K. R., and Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motiv. Emot.* 1, 331–346. doi: 10.1007/BF00992539

Schimmack, U., and Grob, A. (2000). Dimensional models of core affect: a quantitative comparison by means of structural equation modeling. *Eur. J. Pers.* 14, 325–345. doi: 10.1002/1099-0984(200007/08)14:4<325::AID-PER380>3.0.CO;2-I

Schimmack, U., and Reisenzein, R. (2002). Experiencing activation: energetic arousal and tense arousal are not mixtures of valence and activation. *Emotion* 2, 412–417. doi: 10.1037/1528-3542.2.4.412

Schubert, E. (1999). Measuring emotion continuously: validity and reliability of the two-dimensional emotion-space. *Aust. J. Psychol.* 51, 154–165. doi: 10.1080/00049539908255353

Sloboda, J. A., and O'Neill, S. A. (2001). "Emotions in everyday listening to music," in *Music and Emotion: Theory and Research*, eds P. N. Juslin and J. A. Sloboda (Oxford, UK: Oxford University Press), 415–430.

Smith, J. O. (2007). *Mathematics of the Discrete Fourier Transform (DFT): With Audio Applications.* Stanford, CA: W3K Publishing.

Vempala, N. N., and Russo, F. A. (2013). "Exploring cognitivist and emotivist positions of musical emotion using neural network models," in *Proceedings of the 12th International Conference on Cognitive Modeling*, eds R. L. West and T. C. Stewart (Ottawa, ON), 257–262.

Vienna Symphonic Library GmbH (2011). *Vienna Symphonic Library*. Available online at: http://vsl.co.at

Weber, R. (1991). "The continuous loudness judgement of temporally variable sounds with an "analog" category procedure," in *Fifth Oldenburg Symposium on Psychological Acoustics*, eds A. Schick, J. Hellbruck, and R. Weber (Oldenburg: BIS), 267–294.

West, B. T., Welch, K. B., and Galecki, A. T. (2006). *Linear Mixed Models: A Practical Guide Using Statistical Software*. Boca Raton, FL: Chapman & Hall/CRC Press.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometr. Intell. Lab. Syst.* 58, 109–130. doi: 10.1016/S0169-7439(01)00155-1

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 05 January 2018
doi: 10.3389/fpsyg.2017.02239

Modeling Music Emotion Judgments Using Machine Learning Methods

Naresh N. Vempala[1] *and Frank A. Russo*[1,2]*

[1] SMART Lab, Department of Psychology, Ryerson University, Toronto, ON, Canada, [2] Toronto Rehabilitation Institute, Toronto, ON, Canada

Emotion judgments and five channels of physiological data were obtained from 60 participants listening to 60 music excerpts. Various machine learning (ML) methods were used to model the emotion judgments inclusive of neural networks, linear regression, and random forests. Input for models of perceived emotion consisted of audio features extracted from the music recordings. Input for models of felt emotion consisted of physiological features extracted from the physiological recordings. Models were trained and interpreted with consideration of the classic debate in music emotion between cognitivists and emotivists. Our models supported a hybrid position wherein emotion judgments were influenced by a combination of perceived and felt emotions. In comparing the different ML approaches that were used for modeling, we conclude that neural networks were optimal, yielding models that were flexible as well as interpretable. Inspection of a committee machine, encompassing an ensemble of networks, revealed that arousal judgments were predominantly influenced by felt emotion, whereas valence judgments were predominantly influenced by perceived emotion.

Keywords: music cognition, music emotion, physiological responses, computational modeling, neural networks, machine learning, random forests

OPEN ACCESS

Edited by:
Massimiliano Palmiero, University of L'Aquila, Italy

Reviewed by:
Wen Li, Florida State University, United States
Daniel Mullensiefen, Goldsmiths, University of London, United Kingdom

***Correspondence:**
Frank A. Russo
russo@ryerson.ca

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *07 February 2017*
Accepted: *11 December 2017*
Published: *05 January 2018*

Citation:
Vempala NN and Russo FA (2018) Modeling Music Emotion Judgments Using Machine Learning Methods. Front. Psychol. 8:2239. doi: 10.3389/fpsyg.2017.02239

INTRODUCTION

The classic philosophical debate on music emotion pits a "cognitivist" view of music emotion against an "emotivist" view (see e.g., Kivy, 1989). The cognitivist view recognizes music as expressing emotion without inducing it in the listener (Konečni, 2008). The emotivist view supposes that music achieves its emotional ends by inducing genuine emotion in the listener. That is to say that the listener not only perceives but also feels the emotion expressed by the music. These feelings may give rise to or be the consequence of physiological responses. Meyer (1956) concedes that while music may on occasion induce a genuine emotional response in the listener, the accompanying physiological responses are likely too undifferentiated to be meaningful.

The debate is far from reconciled, and has been further complicated by the observation that emotion that is perceived in music can in some instances be distinct from emotion that is felt [Gabrielsson, 2002; see Schubert (2014) for a review]. Moreover, Juslin and Västfjäll (2008) argue convincingly that there are likely multiple mechanisms that give rise to felt emotion, ranging from brainstem reflexes to evaluative conditioning. Nonetheless, numerous studies have documented interpretable physiological responses elicited during music listening (Krumhansl, 1997; Nyklicek et al., 1997; Rainville et al., 2006; Lundqvist et al., 2009). Scherer and Zentner (2001) have characterized the cognitivist and emotivist views as complementary, arguing that a fulsome view of music emotion needs to consider both perspectives and the factors that give rise to their dominance.

To the best of our knowledge, the emotivist–cognitivist debate has not been considered from a computational perspective. In the current study, we obtained judgments about emotion conveyed by the music as well as physiological responses. To minimize biasing judgments in favor of one view of music emotion, we told participants that we were interested in judgments of emotion for each excerpt without being explicit regarding "perceived" or "felt" emotion. Judgments were made using a two-dimensional model of emotion encompassing valence and arousal (VA; Russell, 1980). Valence was defined as the hedonic dimension of emotion, ranging from pleasant to unpleasant. Arousal was defined as the mobilization of energy, ranging from calm to excited. In contrast with the discrete view of emotions that argues independent processes for distinct emotions (e.g., Ekman, 1992, 1999), the dimensional approach proposes that all affective states may be characterized on the basis of underlying dimensions of emotion. This approach is in widespread use in music cognition research (e.g., Schubert, 1999; Gomez and Danuser, 2004; Witvliet and Vrana, 2007), and has been found to be particularly effective in characterizing emotionally ambiguous stimuli (Eerola and Vuoskoski, 2011).

We assumed that if the cognitivist position were true, we should be able to model emotion judgments on the basis of deep and surface-level features obtained from the music. Likewise, we assumed that if the emotivist positions were true, we should be able to model emotion judgments on the basis of physiological responses. Another possibility that we considered is that emotion judgments are the result of a meta-level cognitive decision-making process that combines output from a perception module and a feeling module (**Figure 1**). In this scenario, the perception module would take its input from features drawn from the music and the feeling module would take its input from features drawn from physiology. While we acknowledge that this account of emotion judgments is skeletal and reliant on some crude assumptions, it provides a framework to guide our modeling exercise.

We had two main objectives in this study. The first was to develop computational models of emotion judgments. We begin by modeling cognitivist and emotivist positions separately using multilayer perceptrons. We then extend these models to reflect a hybrid position in which both expert networks are considered. We refer to this hybrid, meta-level cognitive framework, as a committee machine[1]. Previous studies have modeled emotion recognition (a) exclusively using audio features [see Kim et al. (2010), for an extensive review; Coutinho and Cangelosi, 2009], (b) exclusively using physiological features (Kim and André, 2008), and (c) using a combination of audio and physiological features in a common network (Coutinho and Cangelosi, 2010). However, none of these studies have modeled emotion recognition as a combination of felt and perceived emotion using a meta-level framework.

[1]A preliminary version of the committee machine described here was reported in Vempala and Russo (2013). Although this prior work was informed by the same theoretical framework, the computational model was based on only 12 excerpts of classical music. Given this small number of excerpts and the lack of genre diversity, the generalizability of the model was extremely limited.

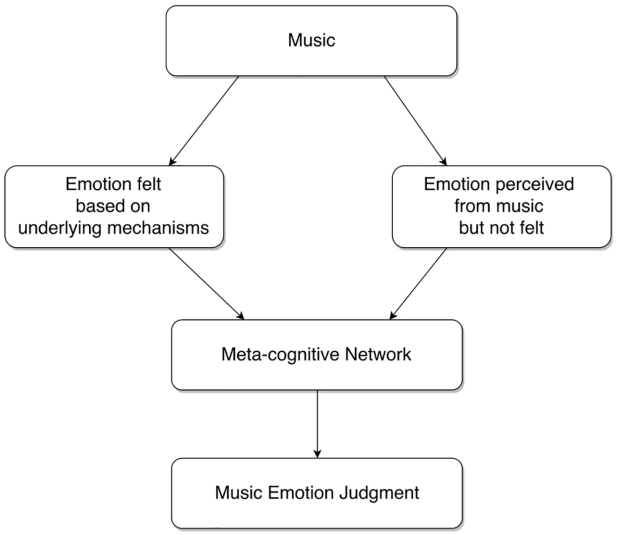


FIGURE 1 | A meta-cognitive network of emotion judgment combining perceived and felt emotion.

Our second objective in this study was more methodological in nature. With the current advent of machine learning (ML), availability and accessibility of ML toolkits, application of ML methods has become more viable for researchers interested in music cognition. While this accessibility to ML methods has opened up new avenues for research, the justification for using specific ML methods is often unclear. In this study, we compared the success of our committee machine with two other ML approaches with the intent of highlighting the relative merits of the different approaches.

MATERIALS AND METHODS

Participants

Our experiment was designed such that it required obtaining emotion judgments and physiological response data from 60 participants. On the basis of previous physiological studies involving testing sessions lasting more than 1 h we were expecting several sources of data loss (e.g., electrodes recording facial muscle activity losing contact due to perspiration). Therefore, we recruited more than the necessary number of participants on an ongoing basis, 110 in total, through our departmental participant pool, until we obtained a complete data set from 60 participants. On average, the final 60 participants (40 females, 15 males, 5 undeclared) were 22.9 years of age ($SD = 7.2$) with 4.0 years of music training ($SD = 3.9$).

Stimuli and Apparatus

Our stimuli consisted of 60 excerpts of high-quality MIDI music drawn from across four genres – Blues, Metal, Pop, and R&B (15 excerpts per genre). Each excerpt spanned approximately 32 bars in duration. We chose to use MIDI music because of the broad range of meta-level information that may be precisely extracted,

consisting of both musical features (e.g., pitch and tempo) and event-related features (e.g., velocity and event onset times), which we plan to use in a separate project.

All 60 excerpts, listed in Appendix 1, were selected such that audio renderings of these MIDI files were representative of their respective genres, and were reasonably consistent with the original versions released commercially. We used MIRtoolbox (Lartillot and Toiviainen, 2007; Lartillot et al., 2008) to extract 12 low-level acoustic to mid-level musical features. These features captured information corresponding to rhythm, timbre, dynamics, pitch, and tonality, and were used in several previous studies on music and emotion (MacDorman et al., 2007; Mion and de Poli, 2008; Laurier et al., 2009; Eerola and Vuoskoski, 2011). The 12 features – *rms*, *lowenergy*, *eventdensity*, *tempo*, *pulseclarity*, *centroid*, *spread*, *rolloff*, *brightness*, *irregularity*, *inharmonicity*, and *mode* – were obtained for each bar of each excerpt (technical descriptions are available in Lartillot, 2014).

Participants used their dominant hand for providing continuous emotion ratings, while their non-dominant hand was connected to the Biopac MP150 data acquisition system for measurement of physiological responses[2]. The five channels of physiological data included heart rate (HR), respiration rate (Resp), skin conductance level (SCL), and facial muscle activity from zygomaticus major (Zyg) and corrugator supercilii (Corr). HR was collected by attaching a photoplethysmogram transducer, using a Velcro strap, to the distal phalange of the middle finger of the non-dominant hand. The transducer was connected to a PPG100C amplifier which measured capillary expansion through an infrared sensor. Resp was measured using a TSD201 respiration belt tightened around the abdomen and attached to an RSP100C amplifier that recorded changes in abdominal circumference. SCL was measured by attaching two TSD203 Ag–AgCl electrodes to the distal phalanges of the index and ring fingers of the non-dominant hand using Velcro straps, connected to a GSR100C amplifier. Facial muscle activity was measured by placing two electrodes over Zyg and two electrodes over corrugator supercilii muscle regions, separated by 25 mm and attached to an EMG100C amplifier.

Physiological data were subjected to feature analysis in order to extract features that have previously been associated with the VA dimensions of emotion. Physiological correlates of valence include Zyg and Corr activity (e.g., Witvliet and Vrana, 2007; Lundqvist et al., 2009; Russo and Liskovoi, 2014). Physiological correlates of arousal include autonomic measures such as HR, respiration, and galvanic skin response (e.g., Iwanaga et al., 1996; Krumhansl, 1997; Baumgartner et al., 2005; Etzel et al., 2006; Sandstrom and Russo, 2010; Russo and Liskovoi, 2014).

Experimental Design and Data Collection

Our experiment was designed such that (a) each participant listened to 12 of the 60 excerpts (i.e., three from each of four genres) and (b) each excerpt was heard by 12 unique participants. Participants received a listening order that was independently randomized to minimize the influence of presentation order. Each excerpt was preceded by 30 s of white noise and followed by 50 s of silence. The root-mean-square (RMS) level of white noise was equalized with the mean RMS level across all 60 excerpts. White noise was used as our baseline for physiological measurements on the basis of previous studies suggesting the appropriate use of RMS-matched white noise as an emotionally neutral baseline for isolating the effects of emotion on physiology (Nyklicek et al., 1997; Sokhadze, 2007; Sandstrom and Russo, 2010).

Each participant heard a stimulus file with 12 excerpts in randomized order, white noise, and silence in the following sequence:

$$\text{WN} \rightarrow \text{Ex} \rightarrow \text{S} \rightarrow \text{WN} \rightarrow \text{Ex} \rightarrow \text{S} \ldots$$

Here, WN indicates white noise, Ex indicates excerpt, and S indicates silence. During the silence phase, participants provided familiarity and preference ratings on the excerpt they heard last. All excerpts were presented at approximately 75 dB SPL over Sennheiser HD 580 Precision Headphones. We used the EMuJoy Software (Nagel et al., 2007) for continuous data collection of emotion ratings on the two-dimensional axes of VA (Russell, 1980).

The experimenter provided participants with a description of the two-dimensional model prior to data collection. It was explained that the *x*-axis conveyed emotion ranging from negative to positive (i.e., valence) and the *y*-axis conveyed emotion ranging from calm to excited (i.e., arousal). Participants were asked to continuously rate each excerpt on a two-dimensional grid while listening. Before commencing, listeners familiarized themselves with the EMuJoy interface while listening to two test excerpts that were not included in the formal experiment. After completion of data collection from all 60 participants, mean VA ratings were computed for each participant, for 32 bar-length segments and for the entire excerpt (i.e., the data were averaged per track for each participant). These values were then averaged across the 12 participants to obtain a mean emotion rating profile for that excerpt. This procedure was repeated for all 60 excerpts.

Data Preparation

Similar to emotion ratings, audio features were extracted for each bar of the excerpt and then aggregated.

Filtering and baseline subtraction for physiological data were performed using FeatureFinder (Andrews et al., 2014), a free Matlab toolbox for physiological signal analysis. The following high-pass (HP) and/or low-pass (LP) filters were applied to raw physiological data: HR (LP = 4 Hz; HP = 0.5 Hz), Resp (LP = 1 Hz; HP = 0.05 Hz), GSR (LP = 10 Hz; HP = 0.5 Hz), Zyg, and Corr (LP = 500 Hz; HP = 5 Hz). Features were obtained for each excerpt and baseline corrected by subtracting the equivalent feature obtained in the final 20 s of 30 s white noise that preceded the excerpt. Similar to audio features and their corresponding emotion ratings, physiological features were

[2]The current study utilizes mean responses (emotion judgments and physiological responses); continuous ratings will be modeled in a separate study.

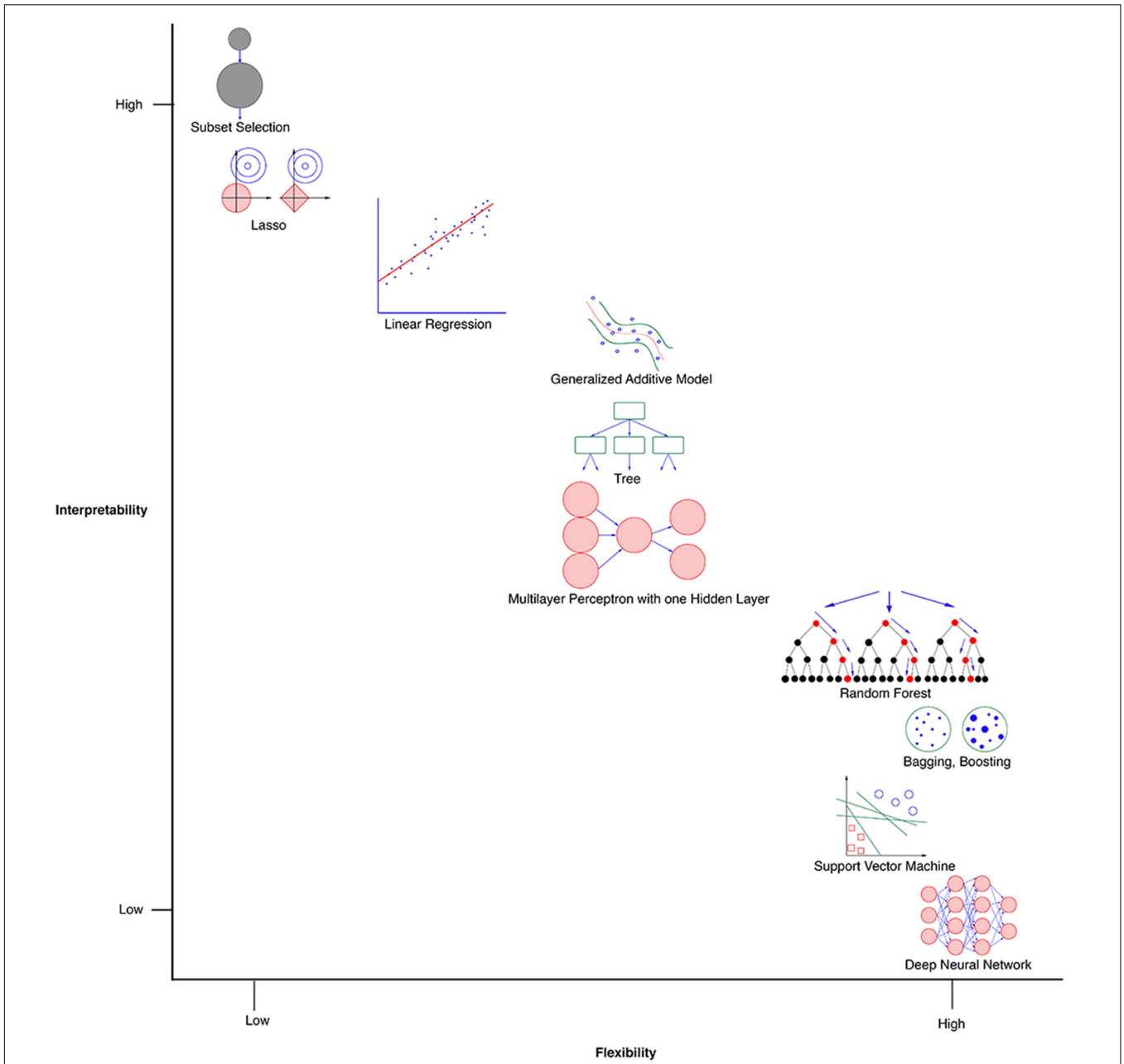


FIGURE 2 | A conceptual visualization of ML methods with respect to interpretability and flexibility, adapted from James et al. (2013).

computed for each bar of the excerpt and then averaged for its entire duration.

Machine Learning Models

There exists a multitude of ML methods for both classification and regression. **Figure 2** provides a conceptual visualization that plots flexibility of methods against interpretability of methods. Since our problem involves predicting emotion ratings as opposed to identifying emotion classes, it is a regression problem. There is no single perfectly suited method for a regression problem. In general, models that are developed with methods that are flexible tend to be powerful in terms of fitting the training data (Hastie et al., 2009; James et al., 2013). However, the ability to interpret the salience of features tends to be better in models that have been developed using methods with less flexibility.

Another related issue is that while flexible models can outperform simpler models as regards to reducing training error, they tend to overfit the regression function to the training set. Hence, the performance of models on a given test set can vary dramatically, making their predictions less generalizable. There are two typical ways of addressing this generalizability

problem. Option 1 involves starting with methods with low flexibility and then moving toward methods with more flexibility until arriving at a model with good performance and generalizability. Option 2 involves starting with a flexible method that improves the likelihood of arriving at a model with good performance, and then moving toward a simpler method that performs relatively well (Kuhn and Johnson, 2013). We chose to adopt a hybrid approach, starting with a method that typically yields intermediate flexibility (i.e., artificial neural networks), and then progressing to methods with lower or higher flexibility – linear regression and random forests (RFs), respectively.

Feature Reduction

When dealing with a high-dimensional dataset, feature reduction by PCA or other means is typically an important step, reducing the storage and computational space while increasing interpretability. In our case, since we were dealing with only 12 audio features, our intention was on the removal of confounding variables. These 12 features serve as independent variables used by our models for predicting the dependent variable – valence or arousal. Although a feature may be strongly correlated with the dependent variable when assessed in isolation, its correlation with the dependent variable may be suppressed when assessed in a model involving numerous features that share common variance. Hence, we computed a correlation matrix of all 12 features. We used a threshold of $r = |0.8|$ to remove features that were strongly correlated with each other. Among the four features – *spectral centroid*, *spectral spread*, *rolloff*, and *brightness*, our results (**Figure 3**) showed that *spectral centroid* was strongly correlated with all three features – *spectral spread*, *rolloff*, and *brightness* ($r > |0.8|$, $p < 0.001$) whereas *spectral spread* and *rolloff* were correlated only with two of the remaining three features. *Brightness* was strongly correlated only with spectral centroid. As a result, we chose to remove *spectral centroid* and *rolloff* from our set of features. We also computed a correlation matrix of the five physiological features for all 60 excerpts, with the same threshold of $r = |0.8|$ for feature removal. None of the features were strongly correlated with each other. Hence, all five features were retained in our models.

Initial Analyses

As a first step in our exploration of the data, we checked to see how well the independent variables accounted for the dependent variables, by examining correlations between the features and the mean VA ratings for the 60 excerpts. We examined correlations for the audio features and physiological features separately since they were being used for separate prediction models.

We observed positive correlations with arousal ratings for *eventdensity*, $r(58) = 0.48$, $p < 0.005$ and brightness, $r(58) = 0.27$, $p < 0.05$. We observed negative correlations with valence ratings for *eventdensity* ($r(58) = -0.33$, $p < 0.05$), *spectral centroid* ($r(58) = -0.3$, $p < 0.05$), and *brightness* ($r(58) = -0.34$, $p < 0.05$). Among the five physiological features, there were no significant correlations with arousal ratings but several with valence. In particular, we observed a negative correlation with valence ratings for Corr, $r(58) = -0.26$, $p < 0.05$, and a positive correlation with valence ratings for Resp, although the latter only reached marginal significance, $r(58) = -0.24$, $p = 0.06$.

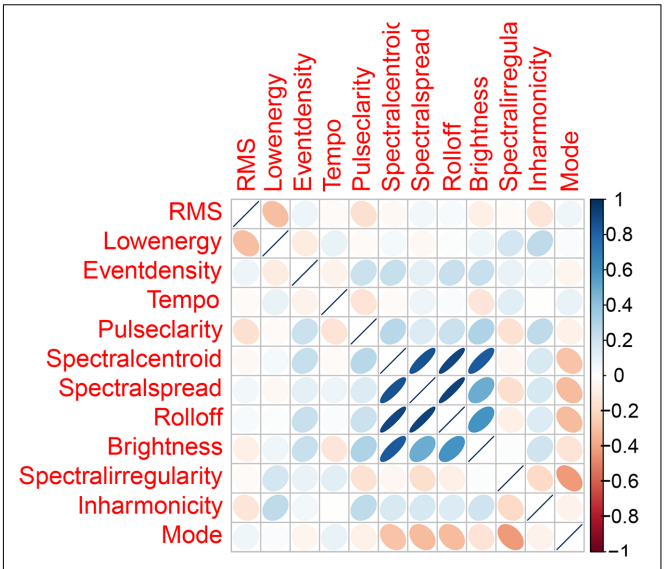


FIGURE 3 | Correlations of the 12 audio features.

Artificial Neural Networks

Our objective in modeling was not to merely provide a prediction method for emotion judgments, but to also provide a theoretical explanation for music emotion judgments. Multilayer perceptrons (i.e., a type of artificial neural network) (Rumelhart et al., 1986; Haykin, 2008) have been known to serve as useful connectionist models for exploring theories in cognitive science (see McClelland and Rumelhart, 1989; Vempala, 2014). Our previous work (Vempala and Russo, 2012, 2013; Russo et al., 2013) has shown that multilayer perceptrons with a single hidden layer can lead to nonlinear regression functions for emotion prediction with good explanatory power. Importantly, these models also lend themselves to interpretation.

We implemented three different types of artificial neural network ensembles for predicting emotion judgments of listeners – one that used only audio features from music to model emotion perceived by a listener (perception model), another that used only physiological responses as features to model emotion felt by a listener (feeling model), and a hybrid ensemble that combined outputs from both these network ensembles (hybrid model), henceforth referred to as a committee machine. All the networks were implemented in Matlab. For all three models (i.e., perception model, feeling model, and hybrid model), the dependent variables were the same – VA. The independent variables were audio features for the perception model, and physiological features for the feeling model. Since the hybrid model was a meta-level network that combined outputs from both these models, its independent variables were both audio and physiological features.

We built two networks with audio features as input – one for predicting valence and one for predicting arousal. Each network was a supervised, feedforward network that consisted of 10 input

units (i.e., one unit for each feature), one hidden layer, and one output unit for either valence or arousal. One important consideration in the use of neural networks is the propensity to overfit to training data, leading to underperformance when exposed to new data. To make our neural networks more robust, we adopted the following training and testing procedure.

Dataset preparation for training and testing

For testing a neural network's performance, the dataset is usually split into a training set consisting of approximately 70–90% of the data, and a test set consisting of 10–30% of the data, respectively. Some decrease in the network's performance is expected from the training set to the test set. Poor performance on the test set indicates that the network has either not fully converged while training (i.e., has been under-trained) or has been over-trained. Hence, the network is retrained accordingly. While this is a widely accepted method for validating performance, problems tend to arise because of idiosyncrasies associated with partitioning. In general, some partitions will lead to overfitting while other partitions will lead to underfitting.

To mitigate problems associated with partitioning the dataset, we used *k*-fold cross-validation. Here, the dataset is split into *k* equal-sized partitions called folds. *k* is typically 5 or 10. This allows us to use each of the *k*-folds as a test set with the remaining *k*−1-folds as the training set. The procedure is repeated *k* times. Performance results on all *k*-folds are then averaged. We used fivefold cross-validation, which enabled us to come up with five different trained networks. We separated our dataset of 60 excerpts such that 44 were used for training the models and the remaining 16 were used for testing the models. Forty of the 44 excerpts were partitioned into fivefolds for cross-validation. So, each fold consisted of eight excerpts with two excerpts from each of the four genres. Each of the five networks was trained on 36 excerpts – 32 from the remaining fourfolds along with the additional four excerpts that were not used for cross-validation.

Network architecture

For methodological reasons, we used separate networks for predicting VA. This architectural decision enabled us to train networks individually without letting convergence for one dependent variable affect the other. It also allowed us to examine feature salience separately for VA.

The networks had to predict VA ratings based on 10 audio features and/or five physiological features (**Figures 4**, **5**). As such, the training set for each of the networks predicting valence consisted of 36 input vectors and 36 corresponding output values for valence, representing the 36 excerpts. Likewise, the training set for each of the networks predicting arousal consisted of 36 input vectors and 36 corresponding outputs for arousal. For the perception networks, each input vector had 10 values, one for each feature. For the feeling networks, each input vector had five values, one for each physiological feature, collapsed across participants. The corresponding outputs with VA values were again collapsed across participants. To maximize network learning (within and across channels), all of the audio and physiological inputs were scaled to a value between 0 and 1

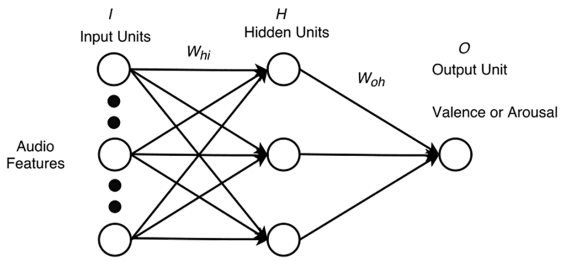


FIGURE 4 | Perception network with 10 features, 3 hidden units, and 1 output.

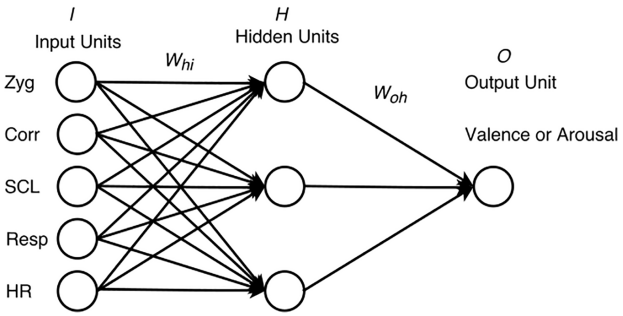


FIGURE 5 | Feeling network with five features, three hidden units, and one output.

(Bishop, 1995) for each feature. VA values for all excerpts were obtained on a scale ranging from −1 to 1. To make these values compatible across the networks, we scaled them to a range between 0 and 1. We chose to reduce the number of hidden units to a number that offered us a flexible non-linear solution while minimizing the likelihood of overfitting. To do so, we used an iterative process of trial and error where we started with the number of hidden units equal to the number of input units, then reduced this number by one at each step, while checking to see if the network's performance remained consistent. Following this process, we decided to keep the number of hidden units to 3. Thus, the network architecture consisted of either 10 input units (one for each audio feature) or five input units (one for each physiological feature), a single hidden layer with three units, and one output unit (either for valence or for arousal).

The following procedure was used to train the network:

(1) Connection weights W_{hi} (input units to hidden units) and W_{oh} (hidden units to output units) were initialized to random numbers between −0.05 and 0.05. Input vectors were fed to the network from the training set in a randomized order. Inputs were multiplied with the connection weights W_{hi}, and summed at each hidden unit.
(2) Hidden unit values were obtained by passing the summed value at each hidden unit through a sigmoid function. These values were multiplied with the connection weights W_{oh}, summed at each output unit, and passed through a sigmoid function to arrive at the final output value between 0 and 1.

(3) Squared errors between the network's output and the mean valence or arousal rating were computed. The backpropagation algorithm using gradient descent was applied and changes in connection weights were stored. At the end of the entire epoch, connection weights were updated with the sum of all stored weight changes.

The perception networks were trained for approximately 2000–3000 epochs by repeating step (2) to reduce the mean-squared error to less than 0.045. The feeling networks took longer to train than the perception networks, and required approximately 15,000–30,000 epochs of training in order to reduce the mean-squared error to less than 0.045. The learning rate parameter was set to 0.1.

After training, each network was tested on its fold and the root mean-squared error (RMSE) was computed. RMSE values for the audio and physiological networks are shown in **Tables 1**, **2**, respectively. The mean and standard errors for perception and feeling networks, for VA, indicate that both types of networks were more-or-less similar in their averaged performance across the fivefolds.

Performance of perception and feeling networks

After completing network training, we tested the trained networks on the remaining 16 excerpts. We used all five perception networks together as an ensemble and averaged their outputs to give the final output for each test excerpt, for VA. We used the same procedure to compute outputs from the feeling networks. For valence, RMSE values for the perception network ensemble and the feeling network ensemble were 0.27 and 0.34, respectively, suggesting that the perception networks performed better than the feeling networks in predicting valence. For arousal, RMSE values for the perception network ensemble and the feeling network ensemble were 0.24 and 0.23, respectively, suggesting that both networks performed similarly.

TABLE 1 | RMSE values of the five perception networks.

Fold	Valence RMSE	Arousal RMSE
1	0.27	0.18
2	0.21	0.34
3	0.16	0.33
4	0.26	0.14
5	0.16	0.15
Mean	0.21	0.23
SE	0.03	0.05

SE indicates standard error.

TABLE 2 | RMSE values of the five feeling networks.

Fold	Valence RMSE	Arousal RMSE
1	0.26	0.25
2	0.24	0.33
3	0.19	0.29
4	0.23	0.24
5	0.24	0.35
Mean	0.23	0.29
SE	0.01	0.02

SE indicates standard error.

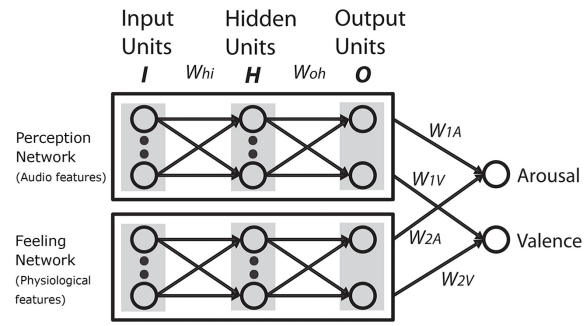


FIGURE 6 | Committee machine – a hybrid network combining results from the perception and feeling network ensembles.

Committee machine

Our next step was to build a model under the assumption that (a) listeners make separate emotion assessments based on what they perceive from the music and what they feel when listening, and (b) their final appraisal of emotion is based on a weighted judgment that takes contributions from both sources into account. This led us to implement our final hybrid model – a committee machine (Haykin, 2008). The committee machine is a meta-level network, as shown in **Figure 6**, which combines outputs from each individual ensemble to arrive at its final output.

First, we implemented a basic committee machine, which merely averaged the outputs from both network ensembles. Specifically, when predicting either the valence or the arousal of an excerpt, outputs from the perception network ensemble and the feeling network ensemble were combined with equal weight contributions of 0.5. RMSE values for the committee machine with ensemble averaged weights (CMEA) were 0.28 for valence and 0.21 for arousal. These results indicate that for valence, the basic committee machine performed about as well as the perception networks and better than the feeling networks. However, for arousal, the basic committee machine performed better than both perception and feeling networks.

Next, we implemented a committee machine that was consistent with our hybrid framework where weights from each of these network ensembles contributed to the final emotion judgment in a way that illustrated meta-level decisions based on emotion conveyed by perception and feeling. To obtain an optimal linear combination of the weights (Hashem, 1997) from each of these individual network ensembles, we performed multiple linear regression such that outputs from these individual ensembles were used as independent variables and mean VA ratings were used as dependent variables. Linear regression was performed on the entire set of 60 excerpts. The models for VA are provided in Equations (1) and (2), respectively.

$$y_{\mathrm{V}} = 0.757x_{1\mathrm{V}} + 0.164x_{2\mathrm{V}} + 0.056 \quad (1)$$

$$y_{\mathrm{A}} = 0.813x_{1\mathrm{A}} + 0.968x_{2\mathrm{A}} - 0.396 \quad (2)$$

Here, y_V and y_A refer to the VA outputs of the committee machine on a scale from 0 to 1. x_{1V} and x_{1A} refer to the VA outputs from the perception network ensemble on a scale from 0 to 1. Likewise, x_{2V} and x_{2A} refer to the VA outputs from the feeling network ensemble on a scale from 0 to 1.

Based on Equation (1), for valence, the meta-level network applies a weight of 0.757 to the perception ensemble output, 1.164 to the feeling ensemble output, and has a bias unit of weight 0.056. Likewise, for arousal, based on Equation (2) the meta-level network applies a weight of 0.813 to the perception ensemble output, 0.968 to the feeling network output, and has a bias unit of weight −0.396. To understand the salience of each individual network's contribution to the overall prediction, we computed their proportion contributions while ignoring the intercepts. For valence, the weight contributions were 82.2% from the perception ensemble and 17.8% from the feeling ensemble. For arousal, the weight contributions were 45.6% from the perception ensemble and 54.4% from the feeling ensemble. As expected, this committee machine (CMLR) performed better than the individual ensembles, and the CMEA, with RMSE values of 0.26 for valence, and 0.2 for arousal.

Linear Regression

Although neural networks helped us from the perspective of cognitive modeling, we wanted to ensure from the perspective of ML that neural networks were not too powerful for our needs. Perhaps a simpler and more interpretable approach could predict VA ratings just as well. To mitigate the possibility of overfitting and to allow for a consistent comparison between models obtained from different ML methods, we again used fivefold cross-validation with the same 44 excerpts that were used for our neural networks. We performed stepwise forward regression to examine which of the 10 audio features were strongly correlated with the VA ratings. The stepwise criteria in our regression models included variables which increased probability of F by at least 0.05, and excluded variables which decreased probability of F by less than 0.1. This led to four derived regression models that predicted valence, and five derived regression models that predicted arousal, using audio features.

For valence, the first model accounted for 17.9% of the variance in ratings, $F(1,34) = 7.39$, $p < 0.05$. The model contained only brightness as its predictor variable ($p < 0.05$). The second model accounted for 18.3% of the variance in ratings, $F(1,34) = 7.63$, $p < 0.01$. Again, the model contained only brightness as its predictor variable ($p < 0.01$). The third model accounted for 25.4% of the variance in ratings, $F(2,33) = 5.62$, $p < 0.01$. The model contained brightness and lowenergy as its predictor variables ($p < 0.01$, $p < 0.05$, respectively). The fourth model accounted for 36.9% of the variance in ratings, $F(3,32) = 6.24$, $p < 0.01$. The model contained brightness, lowenergy, and mode as its predictor variables ($p < 0.01$, $p < 0.05$, and $p < 0.05$, respectively).

For arousal, the first model accounted for 33.4% of the variance in ratings, $F(1,34) = 17.02$, $p < 0.001$. The model contained only eventdensity as its predictor variable ($p < 0.001$). The second model accounted for 43.9% of the variance in ratings, $F(1,34) = 26.6$, $p < 0.001$. The model contained only eventdensity as its predictor variable ($p < 0.001$). The third model accounted for 39.0% of the variance in ratings, $F(2,33) = 10.6$, $p < 0.001$. The model contained eventdensity and mode as its predictor variables ($p < 0.01$ and $p < 0.05$, respectively). The fourth model accounted for 23.6% of the variance in ratings, $F(1,34) = 10.5$, $p < 0.01$. The model contained only eventdensity as its predictor variable ($p < 0.01$). The fifth model accounted for 28.5% of the variance in ratings, $F(1,34) = 13.6$, $p < 0.01$. Again, the model contained only eventdensity as its predictor variable ($p < 0.01$).

We performed stepwise forward regression with the same criteria as before, using the five physiological features as our predictors. This led to three derived regression models that predicted valence. No significant model emerged for arousal.

For valence, the first model accounted for 12.8% of the variance in ratings, $F(1,34) = 5.01$, $p < 0.05$. The model contained only Corr as its predictor variable ($p < 0.05$). The second model accounted for 14.4% of the variance in ratings, $F(1,34) = 5.72$, $p < 0.05$. The model contained only Corr as its predictor variable ($p < 0.05$). The third model accounted for 24.9% of the variance in ratings, $F(1,34) = 11.28$, $p < 0.01$. Again, the model contained only Corr as its predictor variable ($p < 0.01$).

We tested these linear regression models on the 16 excerpts, which the networks had previously not been exposed to. We used all four perception models for valence and all five perception models for arousal as ensembles by averaging their outputs to give the final output for each test excerpt. We used the same procedure for averaging outputs from the three feeling models for valence. For valence, RMSE values for the perception ensemble and the feeling ensemble were 0.25 and 0.66, respectively, clearly showing that the perception ensemble performed much better than the feeling ensemble in predicting valence. For arousal, a comparison between perception and feeling ensembles could not be made since no significant model emerged using physiological features. RMSE for the perception ensemble was 0.23. These results indicate that with audio features, a linear model was sufficient to achieve prediction performance similar to a more flexible model such as a neural network; however, with physiology features, a flexible, nonlinear ML model was necessary to capture the predictive capacity of the independent variables.

Random Forests

Our next step was to see if an approach to modeling with greater flexibility than neural networks could lead to better performance. To reiterate, we were interested in whether a different ML model could offer better prediction, ignoring its suitability as a cognitive computational model. We used RFs (Hastie et al., 2009; James et al., 2013) for this purpose, and implemented them using the *caret* (Kuhn et al., 2016) and *mlbench* (Leisch and Dimitriadou, 2010) packages in R. Random forests create an ensemble of decision trees. Features from the available list are randomly

selected with replacement to first construct individual decision trees using the training data. After several such decision trees are constructed, whenever a new sample is fed to the random forest, predictions are made by these trees. The mean of all predictions is used as the bagged final prediction of the random forest. So, RFs, by nature, are an ensemble method, and are therefore useful for reducing error due to overfitting. An additional aspect of RFs is that they repeatedly take bootstrapped samples from the training data, with replacement, to construct decision trees. This process also helps in reducing error due to overfitting. As such, splitting the data using *k*-fold cross-validation is considered to be unnecessary.

Again, we trained separate random forest models for VA using audio features and physiology features and tested these trained models on the 16 test excerpts. For valence, RMSE values for the perception model and the feeling model were 0.25 and 0.28, respectively, displaying the same pattern as before, with perception features enabling better performance than feeling features. For arousal, RMSE values for the perception model and the feeling model were 0.2 and 0.26, respectively, suggesting that the perception model had an advantage.

As seen in **Table 3**, the Random Forest models obtained using audio features or physiological features were comparable in performance to the committee machine derived using an ensemble of neural networks.

TABLE 3 | Summary of all ML model results.

	Machine learning methods					
	Neural networks		Multiple linear regression		Random forests	
	Valence	Arousal	Valence	Arousal	Valence	Arousal
Audio features (perception models)	Five trained models from fivefold cross-validation (44 excerpts)	Five trained models from fivefold cross-validation (44 excerpts)	Four trained models from fivefold cross-validation (44 excerpts)	Five trained models from fivefold cross-validation (44 excerpts)		
Ensemble performance (16 excerpts) RMSE	0.27	0.24	0.25	0.23	0.25	0.20
Physiology features (feeling models)	Five trained models from fivefold cross-validation (44 excerpts)	Five trained models from fivefold cross-validation (44 excerpts)	Three trained models from fivefold cross-validation (44 excerpts)	No model		
Ensemble performance (16 excerpts) RMSE	0.34	0.23	0.66	No model	0.28	0.26
Committee Machine – CMLR (16 excerpts) RMSE	0.26	0.20				

DISCUSSION

In this study, we revisited the classic debate on music and emotion involving the cognitivists and the emotivists. We approached the debate from a computational modeling perspective by using neural networks (multilayer perceptrons). We modeled emotion judgments from the cognitivist perspective using deep and surface-level audio features obtained from the music alone. Likewise, we modeled emotion judgments from the emotivists perspective using features that relate to felt emotion (i.e., physiological responses). Both networks performed similarly for arousal. However, for valence, the perception networks (i.e., cognitivist) performed better than the feeling networks (i.e., emotivist).

We also proposed another possibility that emotion judgments can be modeled as a meta-level cognitive decision-making process that combines output from a perception module and a feeling module (**Figure 1**) – a hybrid of the cognitivist and emotivist positions. In this scenario, a perception module takes its input from features drawn from the music, while a feeling module takes its input from features drawn from listener physiology. We modeled this possibility using a committee machine that combined VA contributions from two separate network ensembles – a perception network ensemble and a feeling network ensemble. The committee machine performed better than the individual ensembles.

The committee machine enabled us to understand the contribution of each individual network ensemble. For valence, the weight contributions were 82.2% from the perception ensemble and 17.8% from the feeling ensemble. For arousal, the

weight contributions were 45.6% from the perception ensemble and 54.4% from the feeling ensemble. From a theoretical perspective, these findings suggest that felt emotion is more salient in arousal judgments and that perceived emotion is more salient in valence judgments. Given that the feeling ensemble consists of physiological features, and contributed more toward arousal than the perception ensemble, these findings also support the current view in the field about the tight correspondence between physiological features and the arousal dimension of emotion.

We also assessed the validity of our ML method (i.e., neural networks) used for building the committee machine, by comparing it with two other ML methods – multiple linear regression and RFs. To keep comparisons between ML methods consistent, we used the same partitioning of data for training and testing with fivefold cross-validation. This comparison allowed us to ensure that we found the right balance between feature interpretability and model flexibility with neural networks. Multiple linear regression while being less flexible than neural networks as a regression method afforded us the ability to interpret features better. However, this approach revealed its own limitations associated with lack of flexibility. We found that linear methods were not sufficient for deriving a robust, generalizable regression function, using physiological features. When physiological features were used individually as predictors, they were not able to yield a regression model with significant predictors. We refer to these cases as "no model," indicating that none of the features satisfied the inclusion criteria as predictors in a regression model. However, when the features were used in combination with each other as a nonlinear regression function within neural networks, they performed as well or better than audio features in predicting arousal. We chose RFs as our third method, since they are a highly flexible ML method offering various benefits (i.e., building decision trees through binary recursion, repeated subsampling of features and training data to create variance, and ensemble averaging of trees to avoid overfitting). Despite these advantages, the RF approach did not lead to models with greater explanatory power than that which was obtained using neural networks.

There are several important limitations to this work. First, it is important to acknowledge that we cannot fully isolate features that reflect felt emotion as distinct from those that reflect perceived emotion. In all likelihood, the perception of emotion influences the feeling of emotion, independent of the way in which these two networks eventually combine at the level of cognition. Future work should attempt to reconcile this important detail. As we noted at the outset, the models considered here are skeletal and built upon some rather crude assumptions. Second, we have no way of assessing the quality of the features that we provided to the models. The audio features considered as input in the perception models may or may not have been a subset of the full profile of features that were actually processed by listeners. Similarly, although the physiological features we extracted are clearly associated with felt emotion, they do not likely represent the full profile of neurobiological features underlying felt emotion. Accordingly, the power of all of the networks considered here should be considered as bounded by the decisions that were made regarding inputs. Finally, our modeling attempts were handicapped by the size of our dataset. We noticed correlations between some of the physiological features and arousal in some of the genres considered. However, the size of these correlations was reduced when the entire dataset was modeled. Since each genre was limited to 15 excerpts, models derived at the genre level should be interpreted with caution due to concerns about generalizability.

ETHICS STATEMENT

All participants gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Ryerson Research Ethics Board (REB# 2012-343).

AUTHOR CONTRIBUTIONS

NV was responsible for design, data collection, analysis, modeling, and writing. FR was responsible for design, analysis, and writing.

FUNDING

This research was supported by Natural Sciences and Engineering Research Council of Canada Collaborative Research and Development grants awarded to FR and co-sponsored by WaveDNA, Inc. (CRDPJ 430443-12 and CRDPJ 470378-14).

ACKNOWLEDGMENTS

The authors thank Glen Kappel for assistance with gathering and curating the MIDI excerpts. They also thank James McGrath and Salma Shaikh for assistance with data collection, and Gabriel Nespoli for assistance with signal processing for physiological data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.02239/full#supplementary-material

REFERENCES

Andrews, A. J., Nespoli, G., and Russo, F. A. (2014). *FeatureFinder (Version 2.5)*. Available at: http://www.featurefinder.ca/

Baumgartner, T., Esslen, M., and Jäncke, L. (2005). From perception to emotion experience: Emotions evoked by pictures and classical music. *Int. J. Psychophysiol.* 60, 34–43. doi: 10.1016/j.ijpsycho.2005.04.007

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. New York, NY: Oxford University Press.

Coutinho, E., and Cangelosi, A. (2009). The use of spatio-temporal connectionist models in psychological studies of musical emotions. *Music Percept.* 29, 359–375. doi: 10.1525/mp.2009.27.1.1

Coutinho, E., and Cangelosi, A. (2010). A neural network model for the prediction of musical emotions. *Adv. Cogn. Syst.* 71, 333. doi: 10.1049/pbce071e_ch12

Eerola, T., and Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychol. Music* 39, 18–49. doi: 10.1177/0305735610362821

Ekman, P. (1992). Are there basic emotions. *Psychol. Rev.* 99, 550–553. doi: 10.1037/0033-295X.99.3.550

Ekman, P. (1999). "Basic emotions," in *Handbook of Cognition and Emotion*, eds T. Dalgleish and M. J. Power (New York, NY: John Wiley & Sons), 45–60.

Etzel, J. A., Johnsen, E. L., Dickerson, J., Tranel, D., and Adolphs, R. (2006). Cardiovascular and respiratory responses during musical mood induction. *Int. J. Psychophysiol.* 61, 57–69. doi: 10.1016/j.ijpsycho.2005.10.025

Gabrielsson, A. (2002). Emotion perceived and emotion felt: same or different? *Music. Sci.* 5, 123–147.

Gomez, P., and Danuser, B. (2004). Affective and physiological responses to environmental noises and music. *Int. J. Psychophysiol.* 53, 91–103. doi: 10.1016/j.ijpsycho.2004.02.002

Hashem, S. (1997). Optimal linear combinations of neural networks. *Neural Netw.* 10, 599–614. doi: 10.1016/S0893-6080(96)00098-6

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd Edn. New York, NY: Springer.

Haykin, S. (2008). *Neural Networks and Learning Machines*, 3rd Edn. Upper Saddle River, NJ: Prentice Hall.

Iwanaga, M., Ikeda, M., and Iwaki, T. (1996). The effects of repetitive exposure to music on subjective and physiological responses. *J. Music Ther.* 33, 219–230. doi: 10.1093/jmt/33.3.219

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. New York, NY: Springer.

Juslin, P. N., and Västfjäll, D. (2008). Emotional responses to music: the need to consider underlying mechanisms. *Behav. Brain Sci.* 31, 559–621. doi: 10.1017/S0140525X08005293

Kim, J., and André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 2067–2083. doi: 10.1109/TPAMI.2008.26

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., et al. (2010). "Music emotion recognition: a state of the art review," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, 255–266.

Kivy, P. (1989). *Sound Sentiment: An Essay on the Musical Emotions, Including the Complete Text of the Corded Shell*. Philadelphia, PA: Temple University Press.

Konečni, V. J. (2008). Does music induce emotion? A theoretical and methodological analysis. *Psychol. Aesthet. Creat. Arts* 2, 115–129. doi: 10.1037/1931-3896.2.2.115

Krumhansl, C. (1997). An exploratory study of musical emotions and psychophysiology. *Can. J. Exp. Psychol.* 51, 336–352. doi: 10.1037/1196-1961.51.4.336

Kuhn, M., and Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY: Springer.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., et al. (2016). *Caret: Classification and Regression Training, R package version 6.0–64*. Available at: https://github.com/topepo/caret/

Lartillot, O. (2014). *MIRtoolbox 1.6.1 User's Manual*. Technical report, Aalborg: Aalborg University.

Lartillot, O., and Toiviainen, P. (2007). "A Matlab toolbox for musical feature extraction from audio," in *Proceedings of the 10th International Conference on Digital Audio Effects*, Bordeaux, 237–244.

Lartillot, O., Toiviainen, P., and Eerola, T. (2008). "A matlab toolbox for music information retrieval," in *Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization*, eds C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker (Berlin: Springer), 261–268.

Laurier, C., Lartillot, O., Eerola, T., and Toiviainen, P. (2009). "Exploring relationships between audio features and emotion in music," in *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM 2009)*, Jyväskylä.

Leisch, F., and Dimitriadou, E. (2010). *Mlbench: Machine Learning Benchmark*. R package version 2.1–1.

Lundqvist, L., Carlsson, F., Hilmersson, P., and Juslin, P. N. (2009). Emotional responses to music: experience, expression, and physiology. *Psychol. Music* 37, 61–90. doi: 10.1177/0305735607086048

MacDorman, K. F., Ough, S., and Ho, C. C. (2007). Automatic emotion prediction of song excerpts: index construction, algorithm design, and empirical comparison. *J. New Music Res.* 36, 281–299. doi: 10.1080/09298210801927846

McClelland, J. L., and Rumelhart, D. E. (1989). *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises*. Cambridge, MA: MIT press.

Meyer, L. (1956). *Emotion and Meaning in Music*. Chicago, IL: University of Chicago Press.

Mion, L., and de Poli, G. (2008). Score-independent audio features for description of music expression. *IEEE Trans. Audio Speech Lang. Process.* 16, 458–466. doi: 10.1109/TASL.2007.913743

Nagel, F., Kopiez, R., Grewe, O., and Altenmüller, E. (2007). EMuJoy: software for continuous measurement of perceived emotions in music. *Behav. Res. Methods* 39, 283–290. doi: 10.3758/BF03193159

Nyklicek, I., Thayer, J. F., and Van Doornen, L. J. P. (1997). Cardiorespiratory differentiation of musically-induced emotions. *J. Psychophysiol.* 11, 304–321.

Rainville, P., Bechara, A., Naqvi, N., and Damasio, A. R. (2006). Basic emotions are associated with distinct patterns of cardiorespiratory activity. *Int. J. Psychophysiol.* 61, 5–18. doi: 10.1016/j.ijpsycho.2005.10.024

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 1161–1178. doi: 10.1037/h0077714

Russo, F. A., and Liskovoi, L. (2014). "Physiological responses," in *Music in the Social and Behavioral Sciences: An Encyclopedia*, ed. W. F. Thompson (London: SAGE Publications), 862–865.

Russo, F. A., Vempala, N. N., and Sandstrom, G. M. (2013). Predicting musically induced emotions from physiological inputs: linear and neural network models. *Front. Psychol.* 4:168. doi: 10.3389/fpsyg.2013.00468

Sandstrom, G. M., and Russo, F. A. (2010). Music hath charms: the effects of valence and arousal on the regulation of stress. *Music Med.* 2, 137–143. doi: 10.1177/1943862110371486

Scherer, K. R., and Zentner, M. R. (2001). "Emotional effects of music: production rules," in *Series in Affective Science. Music and Emotion: Theory and Research*, eds P. N. Juslin and J. A. Sloboda (New York, NY: Oxford University Press).

Schubert, E. (1999). *Measurement and Time-Series Analysis of Emotion in Music*. Ph.D. thesis, University of New South Wales, Sydney, NSW.

Schubert, E. (2014). Emotion felt by the listener and expressed by the music: literature review and theoretical perspectives. *Front. Psychol.* 4:837. doi: 10.3389/fpsyg.2013.00837

Sokhadze, T. (2007). Effects of music on the recovery of autonomic and electrocortical activity after stress induced by aversive visual stimuli. *Appl. Psychophysiol. Biofeedback* 32, 31–50. doi: 10.1007/s10484-007-9033-y

Vempala, N. N. (2014). "Neural network models," in *Music in the Social and Behavioral Sciences: An Encyclopedia*, ed. W. F. Thompson (London: SAGE Publications), 805–807.

Vempala, N. N., and Russo, F. A. (2012). "Predicting emotion from music audio features using neural networks," in *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR)*, London.

Vempala, N. N., and Russo, F. A. (2013). "Exploring cognitivist and emotivist positions of musical emotion using neural network models," in *Proceedings of the 12th International Conference on Cognitive Modeling (ICCM)*, Ottawa, ON.

Witvliet, C. V., and Vrana, S. R. (2007). Play it again Sam: repeated exposure to emotionally evocative music polarises liking and smiling responses, and influences other affective reports, facial EMG, and heart rate. *Cogn. Emot.* 21, 3–25. doi: 10.1080/02699930601000672

Conflict of Interest Statement: The research was co-sponsored by WaveDNA, an industry partner. Although the manuscript presents no opportunity for commercial promotion (there was no use or evaluation of commercial products), it is possible that some version of the computational models described here will be integrated into future releases of WaveDNA's commercial software.

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 27 April 2017
doi: 10.3389/fpsyg.2017.00537

Impaired Maintenance of Interpersonal Synchronization in Musical Improvisations of Patients with Borderline Personality Disorder

Katrien Foubert[1]*, *Tom Collins*[2,3] and *Jos De Backer*[1]

[1] Music Therapy, Department of Music, LUCA School of Arts, Association KULeuven, Leuven, Belgium, [2] Department of Psychology, Lehigh University, Bethlehem, PA, USA, [3] Music Artificial Intelligence Algorithms, Inc., Davis, CA, USA

Borderline personality disorder (BPD) is a serious and complex mental disorder with a lifetime prevalence of 5.9%, characterized by pervasive difficulties with emotion regulation, impulse control, and instability in interpersonal relationships and self-image. Impairments in interpersonal functioning have always been a prominent characteristic of BPD, indicating a need for research to identify the specific interpersonal processes that are problematic for diagnosed individuals. Previous research has concentrated on self-report questionnaires, unidirectional tests, and experimental paradigms wherein the exchange of social signals between individuals was not the focus. We propose joint musical improvisation as an alternative method to investigate interpersonal processes. Using a novel, carefully planned, ABA′ accompaniment paradigm, and taking into account the possible influences of mood, psychotropic medication, general attachment, and musical sophistication, we recorded piano improvisations of 16 BPD patients and 12 matched healthy controls. We hypothesized that the insecure attachment system associated with BPD would be activated in the joint improvisation and manifest in measures of timing behavior. Results indicated that a logistic regression model, built on differences in timing deviations, predicted diagnosis with 82% success. More specifically, over the course of the improvisation B section (freer improvisation), controls' timing deviations decreased (temporal synchrony became more precise) whereas that of the patients with BPD did not, confirming our hypothesis. These findings are in accordance with previous research, where BPD is characterized by difficulties in attachment relationships such as maintaining strong attachment with others, but it is novel to find empirical evidence of such issues in joint musical improvisation. We suggest further longitudinal research within the field of music therapy, to study how recovery of these timing habits are related to attachment experiences and interpersonal functioning in general.

Keywords: interpersonal synchronization, musical improvisation, interpersonal functioning, borderline personality disorder, music information retrieval, music therapy, attachment, timing

OPEN ACCESS

Edited by:
Frank A. Russo,
Ryerson University, Canada

Reviewed by:
Susanne Metzner,
University of Augsburg, Germany
Simone Dalla Bella,
University of Montpellier 1, France

***Correspondence:**
Katrien Foubert
katrien.foubert@luca-arts.be

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *26 August 2016*
Accepted: *23 March 2017*
Published: *27 April 2017*

Citation:
Foubert K, Collins T and De Backer J (2017) Impaired Maintenance of Interpersonal Synchronization in Musical Improvisations of Patients with Borderline Personality Disorder. Front. Psychol. 8:537. doi: 10.3389/fpsyg.2017.00537

INTRODUCTION

Borderline personality disorder (BPD) is a serious and complex mental disorder characterized by pervasive difficulties with emotion regulation, impulse control, and instability in interpersonal relationships and self-image (Skodol et al., 2002). The lifetime prevalence of BPD is 5.9% (Grant et al., 2008). Since its earliest descriptions in the literature, impairment in interpersonal functioning has been a prominent characteristic of people with BPD, both from theoretical and diagnostic standpoints (Stern, 1938; Kernberg, 1967).

Despite the long history and that all current evidence-based treatments of BPD include strategies to improve interpersonal functioning, there remains a serious need to elucidate the specific interpersonal processes that are problematic for individuals diagnosed with BPD (Hill et al., 2011).

There are different methods that have been used for assessing interpersonal functioning in BPD individuals. Interpersonal functioning is traditionally measured by self-report questionnaires and interviews (Sinnaeve et al., 2015). Researchers have recently used other methods, such as experimental paradigms, behavioral observations, ecological momentary assessment, neuroscience based and psychophysiological tasks, with the aim to assess and characterize better interpersonal difficulties (see review Lazarus et al., 2014). However, most studies in BPD use unidirectional tests, such as concerning facial emotions expressed in pictures (Roepke et al., 2013; Lowyck et al., 2016). A disadvantage of both self-report questionnaires and current experimental paradigms is that the "hallmark of social interaction, the circular exchange of social signals between two or more individuals" (Roepke et al., 2013, p. 9) is not the focus of study. In this paper, we propose accompanied musical improvisation as an alternative method to investigate interpersonal processes associated with BPD. The embodied context of the musical interaction makes it possible to study the automatic, preconscious behavior within complex interpersonal interactions, which constitutes a lacuna in unidirectional tests.

Musical improvisation is frequently used in case studies to study interpersonal processes in music therapy with BPD patients (De Backer and Sutton, 2014). Clinical research in music therapy has a long tradition of qualitative research, based on detailed video and audio analyses of cases (Wheeler and Kenny, 2005; Lee and McFerran, 2015). The various methods and approaches that have been developed to study musical improvisations require many cycles of subjective listening and reflection in order to describe, analyze and interpret the therapeutic significance of the music (Bonde, 2005; Wosch and Wigram, 2007). Case study research from music therapy describes difficulties in musical interaction within the BPD population (Kupski, 2007; Knoche, 2009; Odell-Miller, 2011; Plitt, 2012; Hannibal, 2014; Strehlow and Lindner, 2016). Strehlow and Lindner (2016) described and categorized different interpersonal interaction dimensions of a music therapy process with BPD patients on the basis of an intensive case study ($n = 20$). Based on subjective analysis of music therapy video recordings, they identified 10 interaction patterns reflecting typical BPD themes such as splitting phenomena, trauma genesis, aggression and mentalization, and regulation of proximity and distance. One of the contributions of our study is to provide more objective, empirical evidence of the playing habits, and interpersonal behavior of BPD patients. For this, we will be using Music Information Retrieval (MIR) variables to quantify the playing habits and interpersonal behavior in musical improvisation with BPD individuals. To our knowledge, there is no existing research on the actual playing and interactions (i.e., interpersonal musical behaviors) of patients with BPD in music therapy.

Attachment Theory Predicts Impairments in Temporal IPS in BPD Individuals

Previous experimental research on musical improvisation has focused on individual performers (e.g., Keller et al., 2011; Norgaard, 2011, 2014). More recently, researchers have emphasized the interaction in joint improvisation as an ecologically valid domain to investigate interpersonal processes, and spontaneous coordinated behavior such as interpersonal synchronization (IPS) in particular (Keller et al., 2014; Walton et al., 2015). In a musical joint improvisation, the playing behavior emerges within a context of social collaboration, and without musical scores. Joint musical improvisation is a complex interaction to study, but Jeung and Herpertz (2014) stress the importance of socially complex stimuli to study interpersonal processes in patients with BPD.

Fundamental to the interactions involved in joint musical improvisation are affective and temporal IPS (Iyer, 2004; De Backer and Foubert, 2011; Hennig, 2014). Affective IPS in musical improvisations consists of shared moments that are important in changing the relationship and moving it to a deeper level of intersubjectivity within a therapeutic process. There have been a number of studies concerning affective IPS, addressing synchronicity (De Backer, 2008), meaningful moments (Amir, 1996), significant moments (Trondalen, 2006), affect attunement (Trondalen and Skårderud, 2007), and inter-affective synchronization (Schumacher and Calvet, 2007).

In this study, we will focus on temporal IPS. Temporal IPS entails the capacity to plan and execute specific actions at precise times, in relation to other performers. People can synchronize spontaneously, such as when people start to walk unintentionally in the same gait cadence. Other forms of temporal IPS can be intentional, for instance when dancers attune their movements to those of a partner. Temporal synchronization in a joint action is generally assessed based on measurements of "asynchronies" or timing deviations between people (Mills et al., 2015).

Experimental research in the normal healthy population demonstrates a strong relationship between the quality of temporal IPS in (musical) joint action and experiences related to social cohesion (Marsh et al., 2009), cooperation (Anshel and Kipper, 1988; Wiltermuth and Heath, 2009), bonding and attachment (Hove and Risen, 2009; Wheatley et al., 2012). As for the BPD population, individuals appear to cooperate less in an experimentally manipulated interpersonal context than do controls (Lazarus et al., 2014). Further, BPD individuals are likely to have more difficulties in repair of

relationship ruptures than controls (King-Casas et al., 2008). Ruptures in cooperation seem to be associated with diminished trust in the interacting partner (Seres et al., 2009; Unoka et al., 2009). Finally, oxytocin, a neuropeptide known to enhance cooperation and prosocial behavior for instance in musical joint action (e.g., Grape et al., 2002), may have paradoxical effects for BPD individuals. For example, a study of Bartz et al. (2010) showed that intranasal administration of oxytocin did not have its normal trust facilitating effects in response to a hypothetical partner cooperation in BPD individuals.

From a theoretical viewpoint, BPD is typically characterized by disturbed attachment (Agrawal et al., 2004; Gunderson and Lyons-Ruth, 2008; Beeney et al., 2016). According to attachment theory (Bowlby, 1988), the quality of relationships, such as measured by child-caregiver IPS, results in the development of mental representations, including beliefs about the self, expectations about interpersonal relationships and their quality, all of which act as prototypes or attachment patterns (e.g., secure/insecure) in later adult social interactions (Fraley, 2002; Shaver and Mikulincer, 2005; Scott et al., 2009; Lindsey and Caldera, 2014). This attachment theory is supported by research suggesting that the quality of child-caregiver IPS is critical to the emergence of other socio-cognitive and socio-affective abilities (Crandell et al., 2003; Feldman, 2007a; Newman et al., 2007; Feldman, 2007b, 2012; Gratier, 2009; Hobson et al., 2009; Guedeney et al., 2011; Kiel et al., 2011; Kleinspehn-Ammerlahn et al., 2011; Dumas et al., 2014).

Empirical research shows that BPD patients have difficulties in maintaining close relationships, and attachment relationships in particular (e.g., romantic partner, Melges and Swartz, 1989; Levy, 2005; Gunderson and Lyons-Ruth, 2008; Choi-Kain et al., 2009; Fonagy and Luyten, 2009; Beckes and Coan, 2011; Levy et al., 2015; Beeney et al., 2016). Difficulties in attachment relationships are characterized by oscillations between opposing fears of abandonment and dependency, between neediness and angry withdrawal (Melges and Swartz, 1989). This leads to unstable relationships and difficulties in maintaining strong attachments with others (Bodner et al., 2011). For example, a recent study by Lazarus and Cheavens (2016) found that women with BPD reported more relationship ruptures within the previous month compared to healthy control women.

Based on attachment theory and associated empirical research, we predict that in our study involving an accompanied musical improvisation, the (insecure) attachment system will be activated in BPD patients, and this will affect temporal IPS between therapist and BPD patients. More specifically, we predict:

(1) poorer temporal IPS, represented by higher overall timing deviations, for BPD patients compared to normal controls;
(2) more oscillations (e.g., more variability) in timing deviations between therapist and BPD patients compared to normal controls;
(3) problems in maintaining and improving IPS between therapist and BPD individuals in the course of the joint improvisation compared to normal controls.

Impulsivity Traits Predict Differences in Temporal IPS in BPD Individuals

Additionally, from the perspective of BPD pathology, we assume that impulsivity, a core feature of BPD, will influence temporal IPS in a joint musical improvisation. Impulsivity is one of the 9 diagnostic criteria in the Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV; American Psychiatric Association, 1994). In the literature, BPD is often described and conceptualized as a disorder characterized by high levels of impulsivity (Silk, 2000; Depue and Lenzenweger, 2001; Widiger and Costa, 2002; Scott et al., 2009). These findings motivate a further prediction for our own study, that (4) BPD patients will play in a more impulsive manner than normal controls. In other words, we predict that BPD individuals will be less inhibited in their playing than normal controls and adapting to the therapist's playing more readily.

THE PRESENT STUDY

In this study, we propose using MIR variables for investigating how aspects of a participant's piano playing vary across an accompanied improvisation. Generally, the field of MIR is concerned with the extraction of meaningful information from musical content (Peeters, 2013). Relevant existing work includes research on performance style analysis (Dannenberg et al., 1997; Widmer, 2002; Widmer and Goebl, 2004; Stamatatos and Widmer, 2005; Cheng and Chew, 2008; Chew, 2012), temporal coordination between performers (e.g., Loehr and Palmer, 2011; Keller et al., 2014; Washburn et al., 2014), and one improvisation study involving people with mental retardation (Luck et al., 2006). Luck et al. (2006) found significant associations between musical behavior and diagnosis level—that "most of the features that predicted level of mental retardation related to temporal aspects of the clients' improvisations" (p. 43). We use MIR variables to measure the presence and development of temporal IPS between accompanist and participant, and to measure the presence of rhythmic motifs/patterns in participants' playing. The temporal IPS variables overlap with those used in previous work (e.g., Widmer and Goebl, 2004; Luck et al., 2006; Loehr and Palmer, 2011), although, as we are motivated by different aspects of theory and are therefore investigating different predictions, there is not necessarily a one-to-one correspondence between variable definitions. It is the application of these variables in the context of BPD and joint improvisation that is novel.

Work on performance style analysis and temporal coordination between performers tends to identify temporal, dynamic, and articulatory variable categories. There is a focus on how performers vary in playing the same piece, with less attention paid to *what* notes are played, since this is the same or very similar across performances. Relative to this literature, the variables we calculate include some novel quantifications of *what* is being played—of rhythmic motifs/patterns, based on previous investigations into automatic pattern discovery in music (Collins et al., 2010, 2016; Collins and Meredith, 2013). While our hypotheses are concerned mainly with temporal IPS, it could be that aspects of attachment style and impulsivity manifest not so

much in timing information as in other dimensions of musical organization.

Taking the introduction and this section on MIR variables together, in this study we propose a novel structured piano improvisation paradigm and MIR variables to investigate how aspects of temporal IPS vary across the improvisation. We use logistic regression modeling with these MIR variables as independent variables, to predict whether a given participant is a patient or a control, as well as to address our predictions (1)–(4).

METHODS

Participants

A sample of 16 carefully screened BPD patients and 12 matched normal controls participated in the study. Participants in the BPD group were patients consecutively admitted in the psychiatric hospital UPC KULeuven, Kortenberg (Belgium), who met the following inclusion criteria: (a) a primary diagnosis of BPD according to the structured clinical interview for Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM–IV) Axis II disorders (SCID-II), (b) age between 21 and 60 years, and (c) not participated in music therapy sessions previously. Twenty eight patients were screened for BPD, 20 patients who fulfilled the inclusion criteria were asked to participate in the study, and, subsequently, 16 patients confirmed willingness to do so.

Participants in the matching normal control sample were recruited from the community, based on characteristics from the BPD group. Control participants were pairwise-matched with the BPD sample on gender, age, level of education, level of musical education, and musical principal instrument. Sixteen potential participants were asked and agreed to participate. Four of these participants were excluded because they fulfilled criteria for at least one personality disorder based on the Assessment of DSM-IV personality Disorders (ADP-IV), a self-report questionnaire for personality pathology (see below).

In summary, we collected and carried forward to the analysis data from 16 BPD patients and12 matched controls. The absence of matches for four BPD patients was not of particular concern, because from a methodological standpoint, even if matching has been performed, this does not necessitate a matched analysis (Pearce, 2016).

The study was carried out in accordance with the recommendations of the local ethics committee, UPC KULeuven, and the central ethics committee, UZ KULeuven. All subjects gave written informed consent in accordance with the Declaration of Helsinki. After being provided with the necessary information, all participants (BPD group and matched normal control group) signed informed consent forms and were given an appointment to participate in the improvisation within 4 days. After completing the musical improvisation, participants were asked to fill out the questionnaires as detailed below.

Most of the BPD patients were female (12 female; 3 male; 1 transgender). The mean age was 31 years (sd = 9.41; range 21–51). Three patients completed primary high school, four secondary high school, six higher education (professional bachelor), and three higher education (academic master). Seven patients indicated experience of playing a musical instrument, among which five patients indicated that they had received musical education and two patients described themselves as autodidacts. Two patients had 1 year of musical education; two patients had 2 years of musical education; one patient had 7 years of musical education; one patient had one and a half years autodidactic experience; and one patient had 7 years autodidactic experience. Three patients indicated voice as principal musical instrument, two patients played guitar, one patient piano, and one patient flute.

At the time of the study, 50% of the BPD patients were receiving psychotropic medication. Most patients were using more than one type of psychotropic medication ($n = 5$); only three patients (19%) were using one type of psychotropic medication.

Questionnaires and Measurements

Both BPD patients and normal controls completed (a) a listening test of beat perception, (b) a questionnaire assessing age, gender, educational level, musical principal instrument, musical educational level, music therapy history, motoric restrictions, hearing problems, and sensitivity for sound, (c) self report measures of musical sophistication to analyse the confounding influence of musical experiences (d) self report measures of attachment (see below). Data on depression and current psychotropic medication were only gathered for the BPD group. Our reasoning for testing the influence of depression was because depression has been hypothesized as a possible confounder in interpersonal functioning in BPD (Fonagy and Bateman, 2008; Lowyck et al., 2016). Current psychotropic medication was gathered based on the medical records of the BPD patients.

The Goldsmiths Musical Sophistication Index (Gold-MSI)

The Goldsmiths Musical Sophistication Index (Müllensiefen et al., 2014) is a self-report inventory for individual differences in musical sophistication. Because no Dutch translation was available, we made use of a back-translated design (Hambleton, 2005) to provide a Dutch translated version of the test. Gold-MSI is a 38-item self-report questionnaire. A range of musical skills, abilities, and behaviors are measured which are observable in both musicians and non-musicians. The Gold-MSI assesses General Musical Sophistication and includes additional five subscales: Active Engagement, Perceptual Abilities, Musical Training, Singing Abilities, and Emotions.

Beat Alignment Test

The Iversen and Patel's (2008) beat alignment test is a beat perception test that includes 18 short fragments of instrumental music (each excerpt 10–16 s in duration). The 18 excerpts originate from nine musical pieces within three different styles: Rock, jazz, and well-known classical. The tempi of the short excerpts have a range 85–165 BPM. Participants were invited to listen to the excerpts and to respond whether a simultaneous beep track was on or off the beat of the music. Half of the excerpts had beep tracks exactly on the beat of the music, the other excerpts had beep tracks off the beat.

Diagnostic Inventory for Depression (DID)

The DID (Zimmerman et al., 2004) is a 38-item self-report scale. Both severity of depression and symptom frequency are assessed based on DSM-IV criteria. From this study, we used only the nineteen-item severity subscale. The DID has high levels of test-retest reliability, and good convergent and discriminant validity. The DID was only administered in the BPD group.

Structured Clinical Interview for DSM-IV Axis II Disorders (SCID II)

The SCID II interview (First et al., 1997), in a dutch translated version (Weertman et al., 2000), consists of 119 questions assessing the DSM-IV personality disorders (i.e., paranoid, borderline, narcissistic, schizoid, schizotypal, antisocial, histrionic, avoidant, dependent, and obsessive compulsive). We administered a selection of the SCID, namely the questions assessing borderline personality disorder (15 questions). The SCID-II was only registered in the BPD group and was executed by a senior psychologist, trained in the assessment of the SCID-II interview.

Assessment of DSM-IV Personality Disorders (ADP-IV)

The ADP-IV (Schotte et al., 1998) was administered in the BPD group and the control group as a screening tool to detect potential personality pathology. The ADP-IV is a screening tool for personality disorder and includes 94 items in a randomized order, which represent 80 criteria of the 10 DSM-IV personality disorders, as well as two personality disorders listed in the DSM–IV for research purposes (the depressive and passive-aggressive personality disorders), which are represented in additional 14 research criteria. Each item is rated on a seven-point trait scale, from 1 (*totally disagree*) to 7 (*totally agree*). When a person recognizes the presence of a trait and is giving a score of five (*rather agree*) or higher on a trait question, he/she is asked to answer an additional distress question, "Has this characteristic ever caused you or others distress or problems?" His/her additional answer is scored on a three-point scale: 1 (*totally not*), 2 (*somewhat*), 3 (*most certainly*). The ADP-IV provides dimensional and categorical scoring formats. Categorical personality disorder diagnoses are acquired according to the DSM-IV thresholds. In this study we used the categorical scoring format. Control subjects were excluded in this study when they scored above the respective DSM-IV thresholds.

Relationship Structures (ECR-R)

The Relationship Structures questionnaire (Fraley et al., 2011) is a self-report measurement that is designed to assess two fundamental dimensions underlying attachment patterns: Anxiety and avoidance (Fraley et al., 2000). The anxiety dimension assesses the extent to which people have the tendency to worry about attachment-related concerns, such as the availability and responsiveness of an attachment figure. The avoidance dimension assesses the extent to which people have the tendency to depend on others and to be uncomfortable opening up to them. Prototypically secure people tend to score low on both anxiety and avoidance dimensions. BPD patients tend to score high on anxiety dimensions (Levy, 2005; Levy et al., 2015). The measurement has 9 items and is developed with the aim to assess patterns of attachment across several distinct relationships (mother, father, romantic partner, and best friend). Participants were asked to indicate for each item on a seven-point scale the extent to which they agreed or disagreed with the statement (1: *strongly disagree*; 7: *strongly agree*). The same 9 items can be used with the distinct relationships described above. Recently, a new supplementing item set was designed to assess people's general attachment styles (Fraley et al., 2015). The 9 items can be used also to assess only one kind of relationship, which is described as a short 9-item version of the ECR-R. We included one set of 9 items to assess only one relationship style: People's general attachment styles. This was administered both in the BPD group and the normal control group.

Stimuli

We use a novel, structured piano improvisation paradigm distinguishing between two different accompaniment frameworks—a predictable repetitive interaction, and a more dynamic, socially complex interaction. The therapist's accompaniment was designed to be in a three-part ABA′ structure (see **Figure 1**): In part A, the accompanist played a single low note that sounded for one beat before a two-note chord was played and both were sustained for three beats to make up a four-beat pattern (see the staff notation below A in **Figure 1**). The musical term for this type of accompaniment figure is "bourdon," and it was repeated throughout section A at a steady tempo; as implied by the label A', this bourdon pattern returned in the third section of the therapist's accompaniment; the content of section B was somewhat freer, but it generally contained an increase in tempo and dynamic level, as well as a change from Phrygian to Aeolian modes.

The rationale for this accompaniment design is that in the more dynamic B part of the improvisation, the interaction comes to the fore. Our premise is that in part B, the attachment system will be more activated in BPD patients than in either parts A or A'. As such, differences in temporal IPS between patients and controls may well be revealed in the B part of the accompanied improvisation. The most convenient way to determine whether changes in IPS have occurred within part B is to split the music data for this section in two, B1 and B2, and calculate variables for these subparts separately. In experiments on visual working memory (e.g., Brady et al., 2009), it is quite common to establish regularities in stimuli, upon which participants may come to rely in order to improve task performance, before subverting those regularities and measuring participants' sensitivities. Our ABA′ accompaniment structure, where A establishes the regular bourdon and B subverts it, can be seen as a less common and therefore relatively novel musical analog of experimental paradigms that establish and then subvert regularities in order to measure participants' sensitivities.

As mentioned above, we made use of an ABA′ structured piano improvisation, distinguishing between two different accompaniment frameworks. In the next section, we will give more music-theoretic details related to our improvisation design.

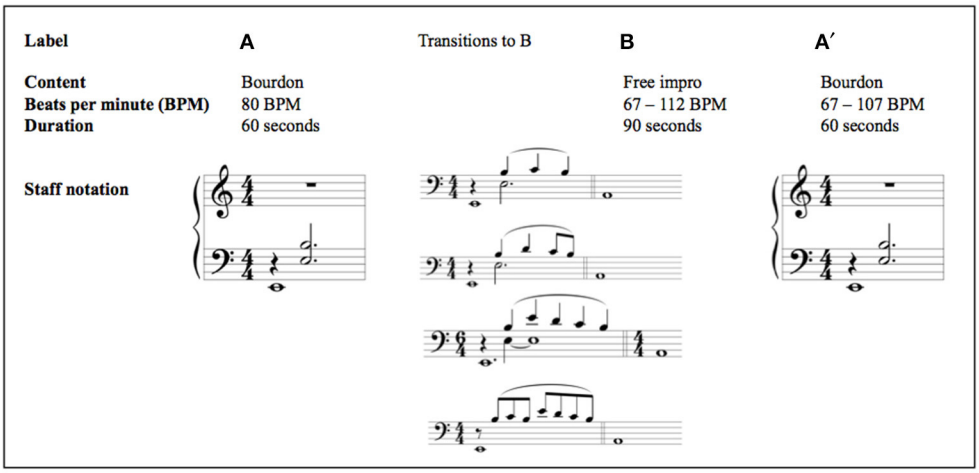


FIGURE 1 | This figure represents the ABA′ structure of the accompaniment design, with characteristics in content, beats per minute (BPM), and duration. To make the transition to part B, the improviser added a short melodic phrase above the bourdon, which initiates the new character and mode. The staff notation excerpts contain the bourdon accompaniment figure, as well as the most frequently played transitions from A to B in this study.

Part A

The A part of the piano improvisation is defined by a repetitive bourdon figure (as presented in **Figure 1**). A bourdon is a sustained or repetitive tonic tone of a scale or mode. When a tone a fifth above the tonic is added (as in **Figure 1**), one speaks of a "fifth bourdon." The technique of sustaining a tone or fifth is originally derived from folk music, wherein melodies were developed over a sustained fifth bourdon. The advantages of using a fifth bourdon are as follows: (1) it provides a technically simple accompaniment with a harmonic basis; (2) this basis is flexible with respect to mode (e.g., an elaboration might be major/minor, modal, or atonal); (3) bourdon offers many possibilities for the development of a melody or a polyphonic elaboration by a participant.

The bourdon can also hinder musical elaboration/improvisation when its use is too open (**Figure 2**, left) or too rigid (**Figure 2**, center). Therefore, the bourdon in this context is articulated with a metric pulse on the first and second beat in common time (**Figure 2**, right). When a meter manifests itself as such, the participant may experience this as a supportive framework for improvisation.

Participants were instructed to play only the white keys of the piano. This constrained the tonal scope of the improvisations somewhat, but still left open the possibility that the participant might emphasize (implicitly or explicitly) one or more modes (e.g., Ionian by emphasizing pitch-class C, Dorian by emphasizing pitch-class D, etc.). Taken as a whole with the therapist's bourdon (which emphasizes pitch-class E), the implied mode may well be E Phrygian. The Phrygian mode, while having a distinctive sound, can be found in a lot of musical cultures (in Japanese scales, Spanish music, jazz, etc.).

Finally, we chose a playing speed of 80 BPM. This was indicated to the therapist via a beep sound (rather, than say, a blinking light), just before the improvisation began. This was done aloud to indicate the speed to the participant also.

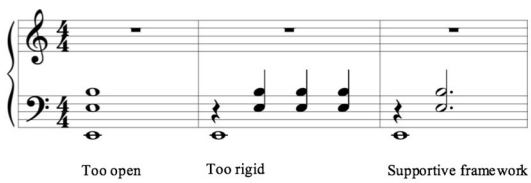


FIGURE 2 | Different options of rhythmic articulation for the bourdon accompaniment.

As measured in adults, this tempo is more toward the lower boundary of speeds to synchronize with an external pulse (Drake et al., 2000). We chose this lower speed because several of the participants had little piano playing or musical experience, so the slow speed gave them the opportunity to explore the instrument without the pressure of a faster tempo. We expected an accelerando (speeding up) in the B part of the improvisation.

Transition and Part B

Part B of the improvisation always starts in A Aeolian, following a brief transition. There was only one exception to this in all our data. This new and clear mode constitutes a substantial change after the repetitive A section with its open fifths character. To make the transition to the B part, the improviser adds a short melodic phrase, above the fifth bourdon, which initiates the new character part B (see **Figure 1**). The B part is characterized by relatively little repetition and freer improvisation. The tonal content remains modal, however. In this section, the therapist was asked to attune and adapt his playing (tempo, timbre, and dynamic) to that of the participant. Generally, we observed an initial increase in tempo and dynamics in this section.

Part A′

The A′ part of the improvisation was a return to the repetitive bourdon figure of part A, the only difference being that generally the tempo began higher (due to coming from the faster B part), and we placed no restriction on it returning to 80 BPM (although sometimes it did).

MIR Variables

In the previous section, we stated the potential utility of MIR variables for investigating how aspects of a participant's piano playing vary across an accompanied improvisation. In the interests of clarity, we defer details of the music data processing and mathematical definitions of all MIR variables to Supplementary Materials. In brief, the music data was beat-tracked by a professional musician/music therapist, and each improvisation was then quantized automatically using the Lisp package MCStylistic and Matlab package PattDisc (Collins, 2011). The purpose of these steps (beat-tracking and quantization) is to map and/or compare each performed note to a start time (called ontime) commensurate with how it would be written in staff notation, as a basis for measuring participants' IPS. The variables we considered are shown in **Table 1** (see Supplementary Materials for full definitions). Below we mention only those that became most relevant in our analyses. To avoid our analyses becoming too exploratory, we employed a common, principled variable selection technique called stepwise selection. As can be seen from **Table 1**, the focus was on variables associated with IPS (seven out of 15), but for the sake of thoroughness we included several from other categories (tempo, rhythmic patterns, and interpersonal imitation) that were either straightforward (Occam's razor) or could be obliquely related to IPS.

(1) MD_m stands for mean metrical deviation. This calculates the average deviation between each note performed by a participant and the underlying eighth-note beat of the therapist to which it is closest. It is indicated by the blue horizontal lines in **Figure 3**. The larger the value of MD_m, the more the participant deviates from the beat over an improvisation section, and the more "out of time" their playing will sound. We use this as one operational definition for the participant's overall asynchronies.

(2) The above variable says nothing about whether the participant tends to play ahead of or behind the beat, or leader-follower behavior. The variable LP, standing for lag proportion, is the proportion of times that a participant's notes are behind the

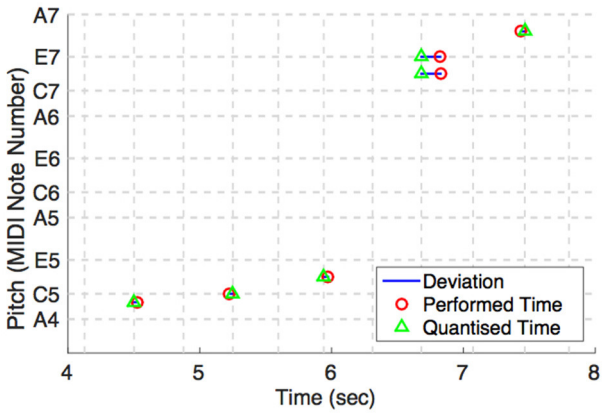


FIGURE 3 | Metrical deviations as a behavioral measure of IPS. Horizontal dashed lines indicate the MIDI note numbers of important pitches in A Aeolian. Vertical dashed lines indicate eighth-note beats. On the eighth-note beat before 7 s, there is a relatively large timing deviation between the participant's and therapist's playing. It can be seen that the participant's performed time of two notes (red circles) lags behind that of the therapist's (green triangles). The blue lines are a visual aid to indicate the deviation from the closest eighth-note beat.

TABLE 1 | Summary of the MIR variables used to quantify aspects of a participant's playing in the accompanied improvisation.

No.	Variable label	Variable name/Description	Relation to theory
1	MD_m	Mean metrical deviation	IPS
2	MD_sd	Standard deviation of metrical deviation	IPS
3	LP	Lag proportion	IPS
4	MDA_m	Mean of metrical deviations that are ahead of the beat	IPS
5	MDA_sd	Standard deviation of metrical deviations that are ahead of the beat	IPS
6	MDB_m	Mean of metrical deviations that are behind the beat	IPS
7	MDB_sd	Standard deviation of metrical deviations that are behind the beat	IPS
8	TMP_m	Mean tempo	Tempo
9	TMP_sd	Standard deviation of tempo	Tempo
10	CR_dur	Compression ratio applied to (ontime, duration)-pairs	Rhythmic patterns
11	TC_o	Translational coefficient applied to ontimes	Rhythmic patterns
12	RS	Rhythmic simplicity	Rhythmic patterns
13	DN	Note density	Interpersonal imitation
14	AI_mu	Mean articulation interaction	Interpersonal imitation
15	AI_min	Minimum articulation interaction	Interpersonal imitation

The final column indicates the aspect of theory to which a particular variable may be most relevant.

beat over the course of an improvisation section. If a participant is always ahead of the beat, then LP = 0, and thus the participant shows more leader behavior; if a participant is always behind the beat, then LP = 1, and thus the participant shows more follower behavior.

(7) The variable MDB_sd, standing for standard deviation of metrical deviations behind the beat, measures the consistency of timing deviations of those notes played late by the participant. If a participant tends to play late (behind the beat) in a consistent manner, then this variable will take a relatively small value; if a participant tends to play late in an erratic manner, then this variable will take a relatively large value.

(10) CR_dur stands for compression ratio applied to (ontime, duration)-pairs. Existing work posits that the more it is possible to compress data, the more structure or patterning the original data contains (Collins et al., 2010, 2016; Collins and Meredith, 2013). The more rhythmic motifs or patterns in a participant's playing, the more their corresponding (ontime, duration)-point set tends to be compressible, and the higher the compression ratio will be.

(12) The variable RS, standing for rhythmic simplicity, measures the prevalence of a participant's most prevalent rhythm. We tally their inter-onset times (the time differences between the notes played), determine their modal (most prevalent) time difference, and define RS as the proportion of all time differences that belong to this mode. If a participant plays only isochronous (evenly spaced) notes (possibly of differing pitches), then their RS = 1; if a participant plays n notes such that the time difference between two consecutive notes is never the same, then RS = $1/n$, i.e., is close to 0.

Variables were calculated from the separate parts of the accompanist's ABA′ structure, with the additional bisection of section B into B1 and B2 (on the basis of the overall duration of part B), to enable investigation of participant sensitivities to the changes in musical content at the beginning of section B.

Apparatus

The piano improvisation was recorded using a Yamaha Disklavier MPX70 piano. Each key was connected with a specially designed optical sensor, and these were connected to a USB MIDI interface (Motu Midi Express 128). Improvisations were recorded with Logic Pro X (Mac system) and exported as MIDI files for subsequent analyses. The MIR variables were calculated in Matlab, and R was used for conducting statistical analyses.

Procedure

Participants were asked to play intuitively and freely on the piano's white keys, without playing well-known songs, but with the aim of exploring joint interaction with the accompanist. They were informed about the ABA′ structure of the improvisation.

The accompanist was a senior registered music therapist, and undertook all the improvisations. The accompanist was blind in the sense that no knowledge about the background of the participants (control or BPD group) was known. The accompanist had 35 years of clinical experience, was experienced in the use of clinical improvisation, and had expertise in the therapeutic musical interventions described by De Backer et al. (2014).

The piano was chosen as an instrument based on a previous study about choices of musical instruments used in individual music therapy sessions with BPD patients (De Backer et al., 2016). In that study, seven Belgian music therapists were asked to fill in questionnaires about the musical instruments chosen by BPD patients in individual sessions over a period of 1 year. Piano was the most frequently used instrument in this population.

The accompanist was sitting on the left side of the piano and played the lower registers of the keyboard. The participant was sitting on the right side, and was playing the upper registers (as shown in **Figure 4**). This setting was based on the concept of the left and right hand position within music therapy (De Backer et al., 2014)—that the therapist can sustain and support (harmonically) the play of the participant. The keyboard had a split point on G4, which enabled (mostly) convenient splitting of the therapist and participant's playing in Logic.

RESULTS

We conducted both musical and statistical analyses of our data[1]. In terms of musical analysis, **Figure 5** (clickable in the online version of the paper) shows transcriptions of some representative excerpts and a plot of how they might be located in a two-dimensional space consisting of temporal synchronization and structural organization. There is clear evidence of

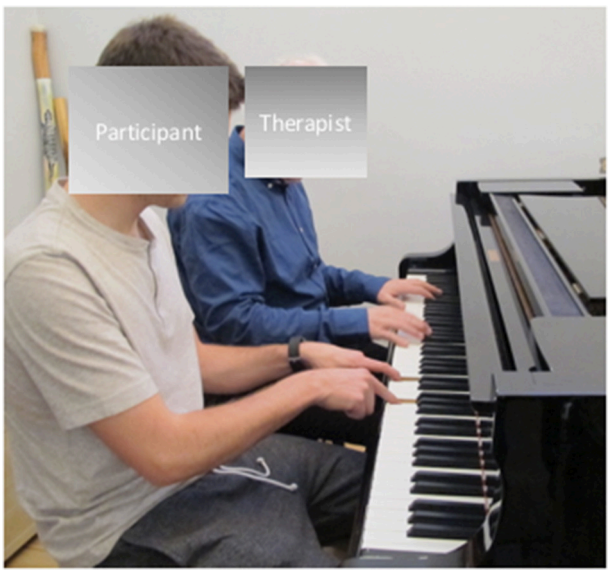


FIGURE 4 | Piano improvisation setting. Participant and therapist play together on one piano. The figure shows the position of the therapist, toward the left side of the keyboard, and the participant, toward the right side of the keyboard (when looking from behind them).

[1]The data, analyses, and plots that underpin the paper are available at http://bit.ly/2bgT77f.

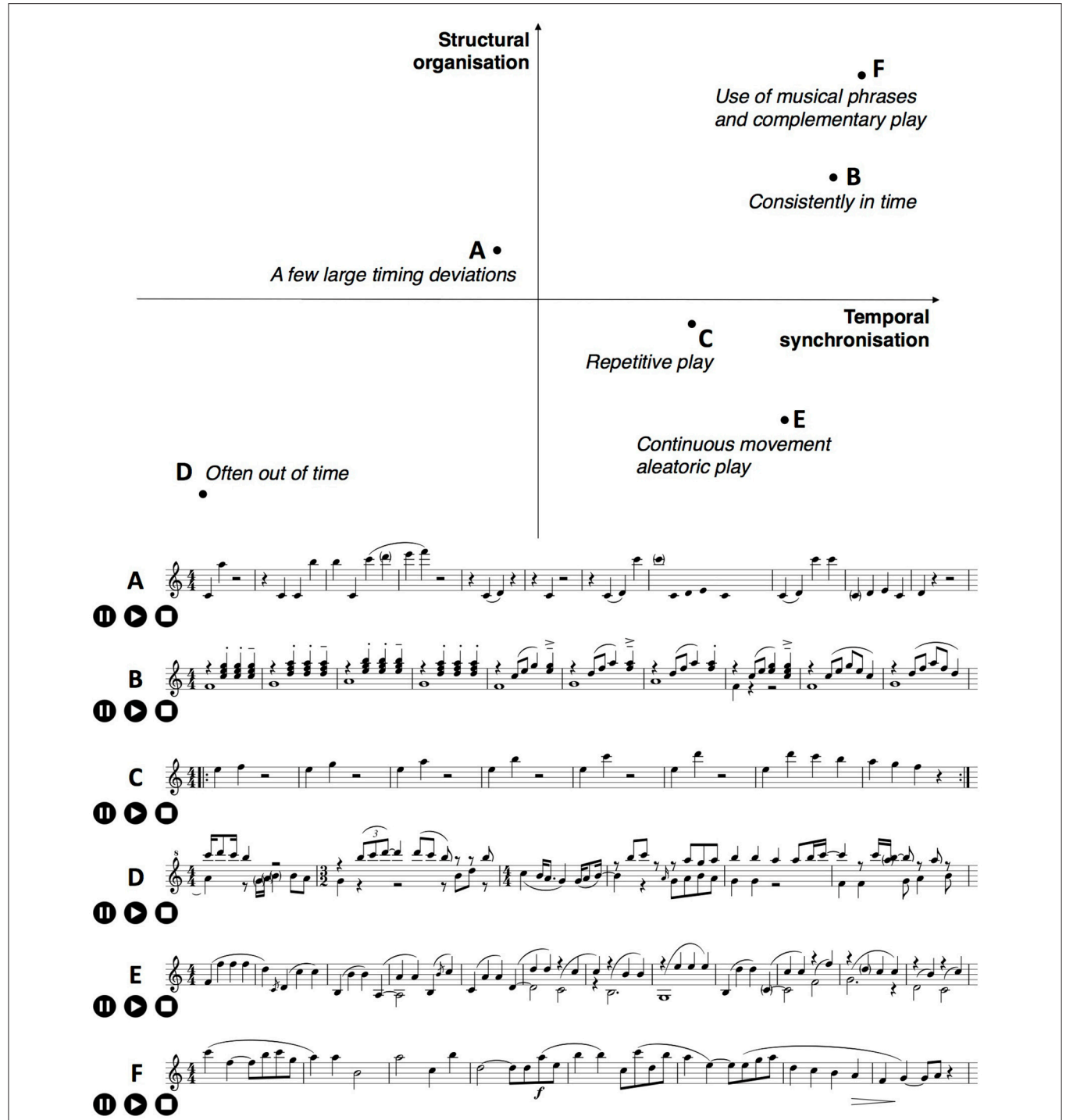


FIGURE 5 | A plot to describe musical characteristics of excerpts from recorded musical improvisations, with structural organization of musical notes on the vertical axis and temporal synchronization on the horizontal. Excerpts **(A–F)** below the plot are transcriptions of the excerpts into staff notation. A clickable version of this plot is available in the online version of this paper.

distinct playing habits, but these transcriptions and plot were made primarily to deepen our knowledge of participants' improvisations, rather than to address any of our four predictions directly.

Statistical Analysis with MIR Variables

In terms of statistical analyses, we conducted logistic regressions on a dependent variable of BPD (patient = 1, control = 0), using independent variables as described in the section "MIR

variables." In other words, we investigated whether it was possible to predict the category of a given participant (BPD or control), based solely on quantifications of their playing habits. Initially, we employed a stepwise selection procedure. Incorporated in the first stage of stepwise selection is a comparative assessment of the discriminative power of each variable in isolation. Supplementary Figure 2 shows the distribution for each variable, split into patients and controls, and that there was only one variable (lag proportion in section B1, or LP_B1) with significant differences between BPD_patients and controls, [$t_{(25.04)} = -2.32$, $p = 0.029$]. As such, LP_B1 is the first variable to enter the model.

While rigorous, stepwise selection and the inclusion of further variables tended to result in overfitting of the data and coefficient blowup.[2] En route to overfitting, however, we identified a parsimonious model that provided strong predictions of BPD or control, as summarized in the following equation and **Table 2A**. The model consists of mean metrical deviation in section B1 (MD_m_B1) and mean metrical deviation in B2 (MD_m_B2):

$$y = 3.35 - 257.44\ \text{MD\_m\_B1} + 210.28\ \text{MD\_m\_B2} + \varepsilon$$

where y is the log odds of having BPD (patient ≈ 1, control ≈ 0) and ε is an error term. Nagelkerke's $R^2 = 0.57$ for this model, and the Hosmer-Lemeshow test indicates that the actual diagnoses (patient or control) are not significantly different from those predicted by the model, $\chi^2(8) = 5.31, p = 0.72$. The signs of the coefficients, −257.44 and +210.28, are opposite, suggesting it is the difference between metrical deviation in sections B1 and B2 that drives prediction of borderline personality disorder. On further inspection, patients' metrical deviations either tended to become bigger in section B2 than in B1, or remain the same, meaning their log odds of having BPD was driven toward 1 in the above formula by the constant term being not much reduced by $-257.44 \times$ MD_m_B1 $+ 210.28 \times$ MD_m_B2. Controls, on the other hand, had smaller metrical deviation in B2 than in B1, meaning their log odds of having BPD was driven toward 0 by $-257.44 \times$ MD_m_B1 being negative and of greater magnitude than $210.28 \times$ MD_m_B2.

A plot of metrical deviation in section B1 is shown in blue in **Figure 6**, metrical deviation in section B2 is shown in red, the difference MD_m_B2−MD_m_B1 is shown in green, and a dashed black line indicates how the difference acts as an effective discriminator between BPD patient and control. All but three patients have a difference above the cut off and all but two

TABLE 2 | Summary of three binary regressions on (A) mean metrical deviation in section B1 (MD_m_B1) and the same in section B2 (MD_m_B2), (B) lag proportion in section B1 (LP_B1), and (C) lag proportion in section B1 (LP_B1) and mean metrical deviation in section B2 (MD_m_B2).

Variable	*B*	*SE B*	*z*-value	*P*
A				
Intercept	3.35	2.10	1.60	0.111
MD_m_B1	−257.44	103.40	−2.49	0.013
MD_m_B2	210.28	95.46	2.20	0.028
Null deviance: 38.24 on 27 degrees of freedom				
Residual deviance: 22.80 on 25 degrees of freedom				
AIC: 28.80				
B				
Intercept	3.55	1.71	2.08	0.038
LP_B1	−7.66	3.83	−2.00	0.045
Null deviance: 38.24 on 27 degrees of freedom				
Residual deviance: 33.16 on 26 degrees of freedom				
AIC: 37.16				
C				
Intercept	5.73	2.47	2.32	0.021
LP_B1	−8.81	4.39	−2.01	0.045
MDB_sd_B1	−37.86	24.06	−1.57	0.116
Null deviance: 38.24 on 27 degrees of freedom				
Residual deviance: 30.13 on 25 degrees of freedom				
AIC: 36.77				

The second column B, contains the coefficient estimate, the third column SE B, contains the standard error of that coefficient, the fourth column contains the z-value and the fifth column the associated p-value. As well as reporting null and residual deviances for each model, Aikaike's information criterion (AIC) is reported also. Models with lower AIC are said to have a better fit to the data, while taking into account the number of constituent variables.

[2]Selection proceeded from the null model according to improvement (reduction) in Aikaike's information criterion (AIC). It was not possible for selection to proceed from a full model because we had more independent variables than data points.

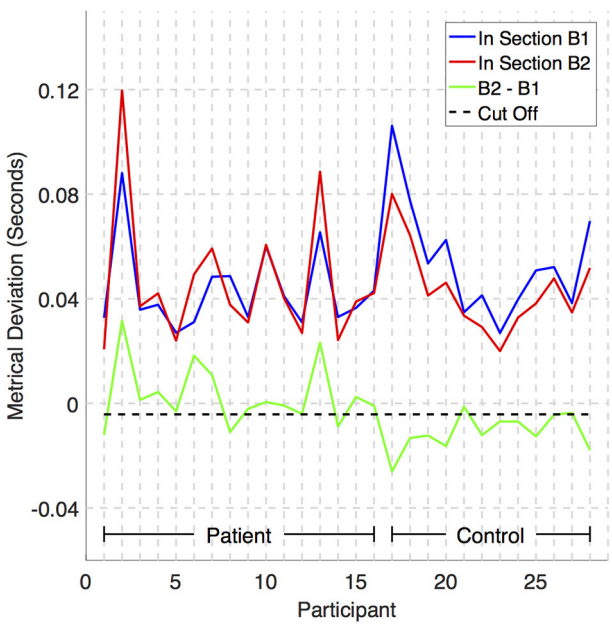


FIGURE 6 | A plot of mean metrical deviations against participant. Mean metrical deviation MD_m is the mean of the absolute deviations between each played note onset and the time of the closest underlying beat. The blue line is MD_m for section B1, the red line is MD_m for section B2, the green line is the difference MD_m_B2−MD_m_B1, and the black horizontal line indicates a cut off that discriminates between most BPD patients and controls.

controls have a difference below. While the difference may seem small (e.g., cut off is ∼4 ms), the pattern in results in **Figure 6** is clear: The blue line is always above the red for controls, but not always so for BPD patients.

Applying a leave-one-out cross-validation procedure with this model, we found a prediction error of 0.18. In other words, this model successfully predicts whether a given participant has or does not have BPD in 23 (= 0.18 × 28) out of 28 cases (82% success). The chance that a baseline (guessing) model succeeds in predicting more than 18 cases is less than 0.05 [$P(B_{n=28,\,p=0.5} > 18) = 0.043$]. Even if we "assist" the baseline model further, by including the knowledge that the proportion p of BPD patients is 16/28, success in predicting more than 20 cases is less than 0.05 [$P(B_{n=28,\,p=16/28} > 20) = 0.040$]. That is, our predictive model for BPD performs significantly better than chance.

As suggested by the first stage of stepwise selection, another model that might provide strong predictions of BPD is based on lag proportion in section B1, LP_B1 (**Table 2B**). Nagelkerke's $R^2 = 0.22$ for this model, and the Hosmer-Lemeshow test indicates that the actual diagnoses are not significantly different from those predicted by the model, $\chi^2(8) = 4.77$, $p = 0.78$. The negative coefficient on LP_B1 suggests that the more a participant lags behind the beat in section B1, the more likely that participant is to be a control. AIC was not as good for the lag variable (= 37.16) as for the model in **Table 2A** (AIC = 28.80), however, and also prediction error on cross-validation was worse (0.24). To investigate whether we might improve the lag proportion model further, we built a third model based on it and variation in playing behind the beat in section B1 (MDB_sd_B1, see **Table 2C**). Nagelkerke's $R^2 = 0.34$ for this model, and the Hosmer-Lemeshow test indicates that the actual diagnoses are not significantly different from those predicted by the model, $\chi^2(8) = 4.45$, $p = 0.81$. This was motivated by seeding a stepwise selection procedure with LP_B1 and including the strongest predicting variable in the next stage, which happened to be MDB_sd_B1. Despite the inclusion of an extra variable, AIC (= 36.77) was not as low as for the metrical deviation model in **Table 2A** (AIC = 28.80). The MDB_sd_B1 variable was not significant in its own right ($p = 0.116$ in **Table 2C**) and prediction error on cross-validation was worse (0.23).

Overall, therefore, we recommend the metrical deviation model as a parsimonious and, according to cross-validation, robust predictor for BPD. As described above, we explored various possibilities in an attempt to find a better model. Now we use the t-test results mentioned briefly at the beginning of this subsection to address questions of significant differences between patients and controls for the metrical deviation (MD_m_B1) and lag (LP_B1) variables: (1) is there a significant difference in MD_m_B1 between BPD patients and controls? According to Welch's two-sample t-test, there is no significant difference [$t_{(19.28)} = -1.49$, $p = 0.153$]. If we restrict the data to matched participants so that we can conduct a (generally more powerful) paired t-test, still there is no significant difference [$t_{(11)} = -1.09$, $p = 0.297$]; (2) is there a significant difference in LP_B1 between BPD_patients and controls? As stated previously, there is a significant difference [$t_{(25.04)} = -2.32$, $p = 0.029$], with controls lagging significantly more in section B1 than do BPD patients. In summary, when the music accompaniment changes markedly in section B1, BPD patients do not play significantly less or more in time than do controls, but controls do tend to lag behind the beat more often than do BPD patients. As a final remark in this results section, we point out that in the second stage of a stepwise selection procedure seeded with LP_B1, there are other interesting variables that could make significant improvements to the model (e.g., rhythmic simplicity in section B or RS_B, compression ratio of ontime-duration pairs in sections A and A' or CR_dur_A, CR_dur_A'). These variables did not contribute as significantly as MDB_sd_B1, however, so we did not explore them further, but they could be investigated by the interested reader via the URL given in the caption of **Figure 5**.

Additional Analysis with the Metrical Deviation Model (MD_m_B1−MD_m_B2)

Based on the significant findings of the metrical deviation model, we calculated a new variable "MD_m_B1−MD_m_B2," which describes the behavior of improvement in IPS during the B part of the improvisation. The interpretation of this measure is that positive scores on this measure imply a trend in improving IPS behavior. A negative score implies a trend in worsening IPS behavior.

We conducted a series of separate analyses to test possible influences of medication, severity of depression, and musical capacities on IPS behavior. To measure musical capacities, we assessed both perceptual and experiential capacities. Finally, we included a psychological measurement of general attachment, to explore if the variance in "MD_m_B1−MD_m_B2" could be at least partially explained by the underlying fundamental two dimensions of attachment: Avoidance and anxiety.

Influence of Musical Sophistication and Beat Perception

General musical sophistication, and the additional five subscales (Active Engagement, Perceptual Abilities, Musical Training, Singing Abilities, and Emotions) did not correlate with "MD_m_B1−MD_m_B2" in BPD patients. In normal controls, no significant correlation was found. Neither was beat perception significantly correlated with "MD_m_B1−MD_m_B2" in BPD patients and normal controls.

Influence of Psychotropic Medication and Mood in BPD Patients

Medication use did not correlate with "MD_m_B1−MD_m_B2" in BPD patients. And also severity of depression did not show a significant correlation.

Influence of General Attachment Style

There was a positive significant correlation between "MD_m_B1−MD_m_B2" and avoidance general attachment style [$r_{(14)} = 0.68$, $p < 0.01$] in BPD patients, and a negative significant correlation between "MD_m_B1−MD_m_B2" and anxious general attachment style [$r_{(14)} = -0.55$, $p < 0.05$].

Avoidance and anxiety dimensions accounted for 52% of the variance in "MD_m_B1−MD_m_B2" in BPD patients [$R^2 =$

0.5229, $F_{(2, 13)} = 7.124$, $p < 0.01$]. There was no correlation between "MD_m_B1−MD_m_B2" and avoidance and anxiety dimension in normal controls. Finally, we want to make clear that we did not Bonferroni-correct the *p*-values for these correlational analysis. If we do, the remaining significant result concerns avoidant general attachment style.

DISCUSSION

Conclusions

With a lifetime prevalence of 5.9% and serious consequences for emotion regulation, impulse control, and interpersonal relationships, BPD is a condition that has been and remains an important subject of research in psychology and neuroscience. Whereas, existing research on BPD—such as self-report questionnaires and unidirectional studies—has not focused on measuring the exchange of social signals between individuals, we attempted to do so via the use of an ABA′ accompanied improvisation paradigm. Impairments in IPS are a known characteristic of BPD, and this paradigm made it possible to measure timing habits in IPS (e.g., the exchange of social signals) over the course of the musical interaction. In the B part of the improvisation (freer improvisation), the intervention from the therapist (accompaniment) invited a greater degree of (social) interaction from the participant. We quantified 15 aspects of each participant's playing across the improvisation sections (A, B, split also into B1and B2, and finally A′), focusing on temporal characteristics that may act as behavioral measures of IPS, as well as some aspects intended to measure impulsivity.

Our main predictions were that there would be: (1) poorer temporal IPS, represented by higher overall timing deviations, for BPD patients compared to normal controls; (2) more oscillations (e.g., more variability) in timing deviations between therapist and BPD patients compared to normal controls; (3) problems in maintaining and improving IPS between therapist and BPD individuals in the course of the joint improvisation compared to normal controls; (4) more impulsivity (less inhibition) in the playing of BPD patients than normal controls. Among our main findings were that: (i) the control group showed significant improvements in IPS over the course of section B2 (variable name MD_m_B2) compared with section B1 (MD_m_B1) of the improvisation, contrary to the BPD group who showed less improvement in IPS over the course of part B (freer improvisation). This finding substantiates prediction 3; (ii) normal controls were significantly more likely to play behind the beat in section B1 (variable name LP_B1) than were BPD individuals, which substantiates prediction 4; (iii) a logistic regression model built on the difference in mean metrical deviation between sections B1 and B2 performed significantly better than chance at categorizing given participants as either having BPD or being a control (82% success rate). So while there was not clear evidence to support predictions 1 and 2 in our findings, we did find evidence to substantiate predictions 3 and 4, as well as a model whose discriminatory power suggests that our behavioral measures of IPS are relevant to the diagnosis of BPD.

Overall Timing Deviations and Oscillations in IPS

Contrary to our prediction 1, results showed that difficulties to synchronize with others represented by strong overall timing deviations (as measured by the variable MD_m) was not related to BPD pathology. Neither was evidence of our prediction 2 found in the results—that BPD individuals would show more oscillations in their playing (as measured by the variable MD_sd), such as being very close in time to the therapist followed by tendencies to withdraw from the therapist. This suggests that differences in overall timing deviations and oscillations in temporal IPS in a joint improvisation are not related to BPD characteristics. Probably these specific timing aspects are more related to other individual characteristics as proposed elsewhere (e.g., Loehr and Palmer, 2011). For instance, Loehr and Palmer (2011) address the correlation between individual tempo profiles of two partners (in piano duet performances) and overall timing deviations in temporal IPS in a joint musical interaction. Their study shows that partners who have a similar "tempo profile" synchronize better. Moreover, well-matched partners are better able to simulate the timing of the other (e.g., action simulation), they adapt better to the timing of the other in the course of the interaction, and there is also more mutual adaptation between the two partners, compared to less well-matched partners (Loehr and Palmer, 2011). These findings were also found in research about movement coordination in joint action (Schmidt and Turvey, 1994; Amazeen et al., 1995; Richardson et al., 2007). We suggest that individual differences in tempo profiles between therapist and patients will influence overall timing deviations in a joint musical improvisation instead of BPD characteristics. Further research may gain insight into the influence of therapist/patient tempo profiles in therapeutic processes.

Maintaining and Improving Temporal IPS

As expected, results showed that BPD patients had difficulties in maintaining and improving temporal IPS during the improvisation compared to normal controls. This was only visible within the B part (in particular, the B1–B2 transition) of the improvisation, where the therapist's playing invited more musical interaction compared to parts A or A', where the therapist was repeating a short, stable pattern. In other words, when the (insecure) attachment system of the patient was activated, difficulties were found in maintaining and improving temporal IPS in musical improvisations with BPD patients.

In addition, we suppose that the underlying cognitive motor skills associated with anticipation (Keller et al., 2007; Pecenka and Keller, 2009; Rankin et al., 2009) and adaptation (Large and Jones, 1999; Repp, 2001, 2011; Large et al., 2002; Repp and Keller, 2008; Loehr et al., 2011; Repp and Su, 2013) are hindered in their ability to regulate and facilitate improvements in temporal IPS when the attachment system is activated in BPD patients.

Taken together, it could be that inner representations of attachment relationships and/or the quality of such relationships are embedded/embodied in cognitive-motor strategies of BPD patients, and that anticipatory mechanisms related to prediction errors are hindered in their capacity to maximize prediction of the future. Brain reward mechanisms are known to regulate

prediction errors. In this sense, our findings seem to support current theories about the relation between alterations in the brain reward system in BPD individuals, attachment and prediction error (Friston, 2005, 2010; Atzil et al., 2011; Fonagy et al., 2011; Brown and Brüne, 2012; Enzi et al., 2013; Herpertz and Bertsch, 2015).

Our findings may have interesting implications in relation to music-therapeutic embodied strategies. If, within the music-therapeutic process, BPD patients can experience repeated experiences of "good enough" temporal IPS, this could lead to implicit repair of maladaptive embodied timing strategies, related to attachment experiences. This might mitigate affectively-oriented interpersonal features in BPD patients, such as intolerance of loneliness, conflicts with dependency, discomfort with care, and fear of abandonment. These suggestions are consistent with research suggesting that attachment patterns could be changed as a result of significant changes in relationships (e.g., Waters et al., 2000). However, we have to be careful about making such predictions, because the findings in our study are based on a cross-sectional experiment and thus are not related to longer and more complicated therapeutic interpersonal processes.

In a recent study (Choi-Kain et al., 2010) an important distinction was made between core affectively-oriented interpersonal features (e.g., attachment fears, intolerance of loneliness) and behavioral interpersonal features (e.g., sadism, entitlement, boundary violations, recurrent breakups, demandingness). In particular it was shown that the core affectively-oriented interpersonal features are more persistent than behavioral interpersonal features. The affectively-oriented symptoms are slower in remission and 15–25% of people with BPD did not show improvement in these symptoms compared to baseline in a 10-year follow-up (Choi-Kain et al., 2010). Our findings promote music therapy as a possible complementary therapy in the current field of evidence-based treatments, especially for treating these affectively oriented interpersonal problems, such as attachment fears, with BPD patients. That said, longitudinal research is necessary to put these hypotheses to the test.

Finally, our findings may further augment the expertise and knowledge of music therapists, offering new tools with which to attune to the timing capacities of the BPD patient, with the aim of making improvements more readily, or where none could be made before.

Impulsivity and Temporal IPS

Results showed that normal controls have the tendency to play their notes significantly more often behind the notes of the therapist than did the BPD patients (so-called lag proportion) in section B1 of the improvisation. This is consistent with our assumption that normal healthy controls are more inhibited in their timing than BPD patients. This is only visible in B1—in the first part of the freer improvisation. When the interaction comes to the fore (part B), BPD patients seems less inhibited, and seem to pursue the more immediate reward of joining the interaction. This is in accordance with previous research about the specificity of impulsivity in BPD patients. In a recent study concerning impulsivity in BPD individuals, a distinction was made between (a) choice or reward-related impulsivity and (b) motor impulsivity (Barker et al., 2015). In particular, the results showed that motor impulsivity was not significantly different between BPD individuals and controls, instead reward-related impulsivity was significantly greater in BPD individuals. Reward-related impulsivity is characterized by choices of small immediate reward, with a focus on the present and with little regard to the future.

Our results suggest that BPD patient have the tendency to pursue the musical interaction more immediately (reward-related impulsivity), relative to healthy controls, who have the tendency to wait longer to join the interaction.

It is plausible that the impulsive playing behavior in BPD patients interferes with our previously described finding about attachment-related impaired maintenance of IPS in joint improvisation. A structural relationship between adult attachment style and impulsivity trait is described in different opposing theoretical models (Scott et al., 2009). The two most common models are: (1) when the insecure attachment system is activated, one cannot rely anymore on secure, adaptive and support-seeking coping. Deficiences in coping strategies may intensify central traits such as impulsivity in BPD patients (Levy et al., 2006). In our study this means that because of the activation of the insecure attachment sytem in the B part of the improvisation, the impulsive behavior in BPD patients is intensified, as seen in B1; (2) an opposing theoretical model contends that the dispositional trait of impulsivity can impede joint interactions, and may contribute to disturbed attachment styles (e.g., Eisenberg et al., 1997, 2000). In our study this means that because of the impulsive behavior in B1, there are difficulties in improving IPS in the course of the improvisation. In this sense, the impulsive behavior impedes the joint interaction in BPD patients. We have to be cautious with possible interpretations, however, because our study was not designed to reveal causal relationships. In either case, our findings are consistent with the existence of a relationship between impulsivity traits and attachment difficulties with regard to impairments in temporal IPS.

Ecological Validity

We would like to stress the ecological validity of this study—that a free musical improvisation approximates social collaboration in the real world more than do experimental studies that make use of methods such as virtual partners (e.g., Repp, 2005; Fairhurst et al., 2013, 2014). It is our premise that the complex and intensive interactions arising from a (freer) musical improvisation are more likely to activate the (insecure) attachment system in BPD patients.

Limitations

This study has some limitations. First our sample is too small to claim any generalizability of our findings. Second, the individuals with BPD in our sample group participated in the context of an inpatient treatment facility, so our results may not be generalizable to other BPD patients. Further research in bigger and other samples is needed. Third, the beat tracking in this study

was done manually, which may cause subjective interferences. Apart from using a beat-tracking algorithm (which may be error prone and so require manual, subjective corrections anyway), we suggest synchronization tasks with computer-generated pacing signals as a possible means of reducing subjectivity. While such tasks miss the human deviations (i.e., variations in timbre and intensity) in the joint improvisation, studies on action simulation have demonstrated that even very reduced stimuli can be experienced as a human product with social meaning (Steinbeis and Koelsch, 2009; Sevdalis and Keller, 2011). Finally, we included only the subscale of the general attachment style in this study. In future research, the assessment of attachment style across several distinct relationships is necessary to gain better insights into attachment-related correlations.

Future Directions

In terms of assessment and diagnosis of BPD, the methods developed in our study could become part of a comparatively lightweight tool to detect possible cases of BPD, in order to reduce the need to administer more onerous questionnaires. With regards treatment of BPD, longitudinal research (such as randomized controlled trials) in music therapy is needed to investigate the extent to which improvement in implicit interpersonal processes of IPS is correlated with the improvement of affectively oriented interpersonal functioning in BPD. In terms of both assessment and treatment, recent improvements in machine learning techniques and musical improvisation invite the possibility that accompaniments for participants could be computer-generated in real time, and thus musical parameters of the accompanist could be controlled more exhaustively, which might lead to more objective measures of (the development of) a participant or patient's musical behavior.

AUTHOR CONTRIBUTIONS

KF and JDB designed and performed experiments; KF analyzed data and wrote the paper; KF and TC developed MIR variables; TC wrote code and ran the analysis; KF and TC prepared the figures; KF, JDB, and TC discussed the results; TC and JDB edited the paper.

FUNDING

Funding source, OPaK (LUCA, School of Arts, KULeuven).

ACKNOWLEDGMENTS

For support in the data collection, we want to thank our colleagues of UPC KULeuven, Kortenberg: Yannic Verhaest, Prof. Benedicte Lowyck, Prof. Rudi Vermote, and Anna Steelandt. For support in the musical analysis, we want to thank our colleagues of LUCA, School of Arts: Martin Valcke en Luk Van Wuytswinkel.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2017.00537/full#supplementary-material

REFERENCES

Agrawal, H. R., Gunderson, J., Holmes, B. M., and Lyons-Ruth, K. (2004). Attachment studies with borderline patients: a review. *Harv. Rev. Psychiatry* 12, 94–104. doi: 10.1080/10673220490447218

Amazeen, P. G., Schmidt, R. C., and Turvey, M. T. (1995). Frequency detuning of the phase entrainment dynamics of visually coupled rhythmic movements. *Biol. Cybern.* 72, 511–518. doi: 10.1007/BF00199893

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders*. Washington, DC.

Amir, D. (1996). "Experiencing music therapy. Meaningful moments in the music therapy process," in *Qualitative Music Therapy Research: Beginning Dialogues*, eds M. Langenberg, K. Aigen, and J. Frommer (Gilsum, NH: Barcelona Publishers), 109–30.

Anshel, A., and Kipper, D. A. (1988). The influence of group singing on trust and cooperation. *J. Music Ther.* 25, 145–155. doi: 10.1093/jmt/25.3.145

Atzil, S., Hendler, T., and Feldman, R. (2011). Specifying the neurobiological basis of human attachment: brain, hormones, and behavior in synchronous and intrusive mothers. *Neuropsychopharmacology* 36, 2603–2615. doi: 10.1038/npp.2011.172

Barker, V., Romaniuk, L., Cardinal, R. N., Pope, M., Nicol, K., and Hall, J. (2015). Impulsivity in borderline personality disorder. *Psychol. Med.* 45, 1955–1964. doi: 10.1017/S0033291714003079

Bartz, J., Simeon, D., Hamilton, H., Kim, S., Crystal, S., Braun, A., et al. (2010). Oxytocin *can* hinder trust and cooperation in borderline personality. *Soc. Cogn. Affect. Neurosci.* 6, 556–563. doi: 10.1093/scan/nsq085

Beckes, L., and Coan, J. A. (2011). Social baseline theory: the role of social proximity in emotion and economy of action. *Soc. Pers. Psychol. Compass* 5, 976–988. doi: 10.1111/j.1751-9004.2011.00400.x

Beeney, J. E., Hallquist, M. N., Clifton, A. D., Lazarus, S. A., and Pilkonis, P. A. (2016). Social disadvantage and borderline personality disorder: a study of social networks. *Pers. Disord.* doi: 10.1037/per0000234. [Epub ahead of print].

Bodner, E., Cohen-Fridel, S., and Iancu, I. (2011). Staff attitudes toward patients with borderline personality disorder. *Compr. Psychiatry* 52, 548–555. doi: 10.1016/j.comppsych.2010.10.004

Bonde, L. O. (2005). "Approaches to researching music," in *Music Therapy Research. Quantitative and Qualitative Perspectives*, *2nd Edn*, ed B. Wheeler (Gilsum, NH: Barcelona Publishers), 489–525.

Bowlby, J. (1988). *A Secure Base: Parent-Child Attachment and Healthy Human Development*. New York, NY: Basic Books.

Brady, T. F., Konkle, T., and Alvarez, G. A. (2009). Compression in visual working memory: using statistical regularities to form more efficient memory representations. *J. Exp. Psychol. Gen.* 138, 487–502. doi: 10.1037/a0016797

Brown, E. C., and Brüne, M. (2012). The role of prediction in social neuroscience. *Front. Hum. Neurosci.* 6:147. doi: 10.3389/fnhum.2012.00147

Cheng, E., and Chew, E. (2008). Quantitative analysis of phrasing strategies in expressive performance: computational methods and analysis of performances of unaccompanied Bach for solo violin. *J. New Music Res.* 37, 325–338. doi: 10.1080/09298210802711660

Chew, E. (2012). *About Time: Strategies of Performance Revealed in Graphs. Visions of Research in Music Education* 20. Available online at: http://www-usr.rider.edu/∼vrme/

Choi-Kain, L. W., Fitzmaurice, G. M., Zanarini, M. C., Laverdière, O., and Gunderson, J. G. (2009). The relationship between self-reported attachment styles, interpersonal dysfunction, and borderline personality disorder. *J. Nerv. Ment. Dis.* 197, 816–821. doi: 10.1097/NMD.0b013e3181bea56e

Choi-Kain, L. W., Zanarini, M. C., Frankenburg, F. R., Fitzmaurice, G. M., and Reich, D. B. (2010). A longitudinal study of the 10-year course of interpersonal

features in borderline personality disorder. *J. Pers. Disord.* 24, 365–376. doi: 10.1521/pedi.2010.24.3.365

Collins, T. (2011). *Improved Methods for Pattern Discovery in Music, with Applications in Automated Stylistic Composition*. Ph. D thesis, Faculty of Mathematics, Computing and Technology, The Open University.

Collins, T., and Meredith, D. (2013). "Maximal translational equivalence classes of musical patterns in point-set representations," in *Proceedings of Mathematics and Computation in Music, Lecture Notes in Computer Science*, eds J. Yust, J. Wild, and A. Burgoyne (Montreal: Springer), 88–99.

Collins, T., Arzt, A., Frostel, H., and Widmer, G. (2016). "Using geometric symbolic fingerprinting to discover distinctive patterns in polyphonic music corpora," in *Computational Music Analysis*, ed D. Meredith (Berlin: Springer), 445–474.

Collins, T., Thurlow, J., Laney, R., Alistair, W., and Garthwaite, P. H. (2010). *A Comparative Evaluation of Algorithms for Discovering Translational Patterns in Baroque keyboard Works*. Utrecht.

Crandell, L. E., Patrick, M. P. H., and Hobson, R. P. (2003). "Still-face" interactions between mothers with borderline personality disorder and their 2-month-old infants. *Br. J. Psychiatry* 183, 239–247. doi: 10.1192/bjp.183.3.239

Dannenberg, R. B., Thom, B., and Watson, D. (1997). "A machine learning approach to musical style recognition," in *International Computer Music Conference* (Michigan Publishing), 344–347.

De Backer, J. (2008). Music and psychosis. *Nord. J. Music Ther.* 17, 89–104. doi: 10.1080/08098130809478202

De Backer, J., and Foubert, K. (2011). "Psychose und die innere Pulsierung," in *Wiener Schule der Differenziellen klinischen Musiktherapie: Ein Update*, eds J. Illner and M. Smetana (Vienna: Praesens Verlag), 13–30.

De Backer, J., and Sutton, J. (2014). *The Music in Music Therapy: Psychodynamic Music Therapy in Europe: Clinical, Theoretical and Research Approaches*. London; Philadelphia: Jessica Kingsley Publishers.

De Backer, J., Dileo, C., Erkkilä, J., Foubert, K., Brabant, O., and Letulé, N. (2016). "Clinical improvisation in music therapy: theory, practice, research and training," in *Abstracts Of The 10th European Music Therapy Conference 25*, eds C. Gold, K. Mössler, and T. Stegemann (Vienna: Nordic Journal of Music Therapy), 87–88.

De Backer, J., Foubert, K., and Van Camp, J. (2014). Lauschendes Spiel. Musiktherapeutische Interventionen in der Psychosenbehandlung. *PDP Psychodynam. Psychother.* 4, 256–263.

Depue, R. A., and Lenzenweger, M. F. (2001). "A neurobehavioral dimensional model of personality disorders," in *Handbook of Personality Disorders*, ed W. J. Livesley (New York, NY: Guilford Press), 136–176.

Drake, C., Jones, M. R., and Baruch, C. (2000). The development of rhythmic attending in auditory sequences: attunement, referent period, focal attending. *Cognition* 77, 251–288. doi: 10.1016/S0010-0277(00)00106-2

Dumas, G., Laroche, J., and Lehmann, A. (2014). Your body, my body, our coupling moves our bodies. *Front. Hum. Neurosci.* 8:1004. doi: 10.3389/fnhum.2014.01004

Eisenberg, N., Fabes, R. A., Guthrie, I. K., and Reiser, M. (2000). Dispositional emotionality and regulation: their role in predicting quality of social functioning. *J. Pers. Soc. Psychol.* 78, 136–157. doi: 10.1037/0022-3514.78.1.136

Eisenberg, N., Fabes, R. A., Shepard, S. A., Murphy, B. C., Guthrie, I. K., Jones, S., et al. (1997). Contemporaneous and longitudinal prediction of children's social functioning from regulation and emotionality. *Child Dev.* 68, 642–664. doi: 10.2307/1132116

Enzi, B., Doering, S., Faber, C., Hinrichs, J., Bahmer, J., and Northoff, G. (2013). Reduced deactivation in reward circuitry and midline structures during emotion processing in borderline personality disorder. *World J. Biol. Psychiatry* 14, 45–56. doi: 10.3109/15622975.2011.579162

Fairhurst, M. T., Janata, P., and Keller, P. E. (2013). Being and feeling in sync with an adaptive virtual partner: brain mechanisms underlying dynamic cooperativity. *Cereb. Cortex* 23, 2592–2600. doi: 10.1093/cercor/bhs243

Fairhurst, M. T., Janata, P., and Keller, P. E. (2014). Leading the follower: an fMRI investigation of dynamic cooperativity and leader–follower strategies in synchronization with an adaptive virtual partner. *Neuroimage* 84, 688–697. doi: 10.1016/j.neuroimage.2013.09.027

Feldman, R. (2007a). Parent-infant synchrony and the construction of shared timing; physiological precursors, developmental outcomes, and risk conditions. *J. Child Psychol. Psychiatry* 48, 329–354. doi: 10.1111/j.1469-7610.2006.01701.x

Feldman, R. (2007b). Parent–infant synchrony: biological foundations and developmental outcomes. *Curr. Dir. Psychol. Sci.* 16, 340–345. doi: 10.1111/j.1467-8721.2007.00532.x

Feldman, R. (2012). Parent-infant synchrony: a biobehavioral model of mutual influences in the formation of affiliative bonds. *Monogr. Soc. Res. Child Dev.* 77, 42–51. doi: 10.1111/j.1540-5834.2011.00660.x

First, M. B., Gibbon, M., Benjamin, L. S., Spitzer, R. L., and Williams, J. B. W. (1997). *User's Guide for the Structured Clinical Interview for DSM-IV Axis II Personality Disorders: SCID-II*. Washington, DC: American Psychiatric Press, Inc.

Fonagy, P., and Bateman, A. (2008). The development of borderline personality disorder—a mentalizing model. *J. Pers. Disord.* 22, 4–21. doi: 10.1521/pedi.2008.22.1.4

Fonagy, P., and Luyten, P. (2009). A developmental, mentalization-based approach to the understanding and treatment of borderline personality disorder. *Dev. Psychopathol.* 21, 1355–1381. doi: 10.1017/S0954579409990198

Fonagy, P., Luyten, P., and Strathearn, L. (2011). Borderline personality disorder, mentalization, and the neurobiology of attachment. *Infant Ment. Health J.* 32, 47–69. doi: 10.1002/imhj.20283

Fraley, R. C. (2002). Attachment stability from infancy to adulthood: meta-analysis and dynamic modeling of developmental mechanisms. *Pers. Soc. Psychol. Rev.* 6, 123–151. doi: 10.1207/S15327957PSPR0602_03

Fraley, R. C., Heffernan, M. E., Vicary, A. M., and Brumbaugh, C. C. (2011). The experiences in close relationships-relationship structures questionnaire: a method for assessing attachment orientations across relationships. *Psychol. Assess.* 23, 615–625. doi: 10.1037/a0022898

Fraley, R. C., Hudson, N. W., Heffernan, M. E., and Segal, N. (2015). Are adult attachment styles categorical or dimensional? A taxometric analysis of general and relationship-specific attachment orientations. *J. Pers. Soc. Psychol.* 109, 354–368. doi: 10.1037/pspp0000027

Fraley, R. C., Waller, N. G., and Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *J. Pers. Soc. Psychol.* 78, 350–365. doi: 10.1037/0022-3514.78.2.350

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 815–836. doi: 10.1098/rstb.2005.1622

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Grant, B. F., Chou, S. P., Goldstein, R. B., Huang, B., Stinson, F. S., Saha, T. D., et al. (2008). Prevalence, correlates, disability, and comorbidity of DSM-IV borderline personality disorder: results from the Wave 2 National Epidemiologic Survey on Alcohol and Related Conditions. *J. Clin. Psychiatry* 69, 533–545. doi: 10.4088/JCP.v69n0404

Grape, C., Sandgren, M., Hansson, L. O., Ericson, M., and Theorell, T. (2002). Does singing promote wellbeing?: an empirical study of professional and amateur singers during a singing lesson. *Integr. Physiol. Behav. Sci.* 38, 65–74. doi: 10.1007/BF02734261

Gratier, M. (2009). Du rythme expressif à la narrativité dans l'échange vocal mère-bébé. *Champ psy* 54:35. doi: 10.3917/cpsy.054.0035

Guedeney, A., Guedeney, N., Tereno, S., Dugravier, R., Greacen, T., Welniarz, B., et al. (2011). Infant rhythms versus parental time: promoting parent-infant synchrony. *J. Physiol.* 105, 195–200. doi: 10.1016/j.jphysparis.2011.07.005

Gunderson, J. G., and Lyons-Ruth, K. (2008). BPD's interpersonal hypersensitivity phenotype: a gene-environment-developmental model. *J. Pers. Disord.* 22, 22–41. doi: 10.1521/pedi.2008.22.1.22

Hambleton, R. K. (2005). "Issues, designs, and technical guidelines for adapting tests into multiple languages and coltures," in *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, eds R. K. Hambleton, P. Merenda, and C. D. Spielberger (Mahwah, NJ: Lawrence Erlbaum), 3–38.

Hannibal, N. (2014). "Implicit and explicit mentalisation in music therapy in the psychiatric treatment of people with borderline personality disorder," in *The Music in Music Therapy: Psychodynamic Music Therapy in Europe: Clinical, Theoretical and Research Approaches*, eds J. De Backer and J. Sutton (London; Philadelphia, PA: Jessica Kingsley Publishers), 211–223.

Hennig, H. (2014). Synchronization in human musical rhythms and mutually interacting complex systems. *Proc. Natl. Acad. Sci. U.S.A.* 111, 12974–12979. doi: 10.1073/pnas.1324142111

Herpertz, S. C., and Bertsch, K. (2015). A new perspective on the pathophysiology of borderline personality disorder: a model of the role of oxytocin. *Am. J. Psychiatry* 172, 840–851. doi: 10.1176/appi.ajp.2015.15020216

Hill, J., Stepp, S. D., Wan, M. W., Hope, H., Morse, J. Q., Steele, M., et al. (2011). Attachment, borderline personality, and romantic relationship dysfunction. *J. Pers. Disord.* 25, 789–805. doi: 10.1521/pedi.2011.25.6.789

Hobson, R. P., Patrick, M. P. H., Hobson, J. A., Crandell, L., Bronfman, E., and Lyons-Ruth, K. (2009). How mothers with borderline personality disorder relate to their year-old infants. *Br. J. Psychiatry* 195, 325–330. doi: 10.1192/bjp.bp.108.060624

Hove, M. J., and Risen, J. L. (2009). It's all in the timing: interpersonal synchrony increases affiliation. *Soc. Cogn.* 27, 949–960. doi: 10.1521/soco.2009.27.6.949

Iversen, J. R., and Patel, A. D. (2008). "The beat alignment test (BAT): Surveying beat processing abilities in the general population," in *The 10th International Conference on Music Perception, and Cognition (ICMPC 10)*, eds M. Miyazaki, Y. Hiraga, M. Adachi, Y. Nakajima, and M. Tsuzaki (Sapporo).

Iyer, V. (2004). Improvisation, temporality and embodied experience. *J. Conscious. Stud.* 11, 159–173.

Jeung, H., and Herpertz, S. C. (2014). Impairments of interpersonal functioning: empathy and intimacy in borderline personality disorder. *Psychopathology* 47, 220–234. doi: 10.1159/000357191

Keller, P. E., Knoblich, G., and Repp, B. H. (2007). Pianists duet better when they play with themselves: on the possible role of action simulation in synchronization. *Conscious. Cogn.* 16, 102–111. doi: 10.1016/j.concog.2005.12.004

Keller, P. E., Novembre, G., and Hove, M. J. (2014). Rhythm in joint action: psychological and neurophysiological mechanisms for real-time interpersonal coordination. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369:20130394. doi: 10.1098/rstb.2013.0394

Keller, P. E., Weber, A., and Engel, A. (2011). Practice makes too perfect: fluctuations in loudness indicate spontaneity in musical improvisation. *Music Percept.* 29, 109–114. doi: 10.1525/mp.2011.29.1.109

Kernberg, O. (1967). Borderline personality organization. *J. Am. Psychoanal. Assoc.* 15, 641–685. doi: 10.1177/000306516701500309

Kiel, E. J., Gratz, K. L., Moore, S. A., Latzman, R. D., and Tull, M. T. (2011). The impact *of* borderline personality pathology on mothers' responses to infant distress. *J. Fam. Psychol.* 25, 907–918. doi: 10.1037/a0025474

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., and Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science* 321, 806–810. doi: 10.1126/science.1156902

Kleinspehn-Ammerlahn, A., Riediger, M., Schmiedek, F., von Oertzen, T., Li, S.-C., and Lindenberger, U. (2011). Dyadic drumming across the lifespan reveals a zone of proximal development in children. *Dev. Psychol.* 47, 632–644. doi: 10.1037/a0021818

Knoche, A. (2009). "Affektregulierung als Ziel der musiktherapeutischen Arbeit mit Borderline-Patienten im Rahmen der Dialektisch-Behavioralen Therapie," in *Jahrbuch Musiktherapie*, ed D. M. Gesellschaft (Wiesbaden: Reichert), 59–98.

Kupski, G. (2007). Borderlinestörung und musiktherapie im kontext der dialektisch- behavioralen therapie. *Musikther. Umschau* 28, 17–27. doi: 10.13109/muum.2007.28.1.17

Large, E. W., and Jones, M. R. (1999). The dynamics of attending: how people track time-varying events. *Psychol. Rev.* 106, 119–159. doi: 10.1037/0033-295X.106.1.119

Large, E. W., Fink, P., and Kelso, S. J. (2002). Tracking simple and complex sequences. *Psychol. Res.* 66, 3–17. doi: 10.1007/s004260100069

Lazarus, S. A., and Cheavens, J. S. (2016). An examination of social network quality and composition in women with and without borderline personality disorder. *Pers. Disord.* doi: 10.1037/per0000201. [Epub ahead of print].

Lazarus, S. A., Cheavens, J. S., Festa, F., and Zachary Rosenthal, M. (2014). Interpersonal functioning in borderline personality disorder: a systematic review of behavioral and laboratory-based assessments. *Clin. Psychol. Rev.* 34, 193–205. doi: 10.1016/j.cpr.2014.01.007

Lee, J., and McFerran, K. S. (2015). Applying interpretative phenomenological analysis to video data in music therapy. *Qual. Res. Psychol.* 12, 367–381. doi: 10.1080/14780887.2014.960985

Levy, K. N. (2005). The implications of attachment theory and research for understanding borderline personality disorder. *Dev. Psychopathol.* 17, 959–986. doi: 10.1017/S0954579405050455

Levy, K. N., Clarkin, J. F., Yeomans, F. E., Scott, L. N., Wasserman, R. H., and Kernberg, O. F. (2006). The mechanisms of change in the treatment of borderline personality disorder with transference focused psychotherapy. *J. Clin. Psychol.* 62, 481–501. doi: 10.1002/jclp.20239

Levy, K. N., Johnson, B. N., Clouthier, T. L., Scala, W. J., and Temes, C. M. (2015). An attachment theoretical framework for personality disorders. *Can. Psychol.* 56, 197–207. doi: 10.1037/cap0000025

Lindsey, E. W., and Caldera, Y. M. (2014). Shared affect and dyadic synchrony among secure and insecure parent-toddler dyads. *Infant Child Dev.* 24, 394–413. doi: 10.1002/icd.1893

Loehr, J. D., and Palmer, C. (2011). Temporal coordination between performing musicians. *Q. J. Exp. Psychol.* 64, 2153–2167. doi: 10.1080/17470218.2011.603427

Loehr, J. D., Large, E. W., and Palmer, C. (2011). Temporal coordination and adaptation to rate change in music performance. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 1292–1309. doi: 10.1037/a0023102

Lowyck, B., Luyten, P., Vanwalleghem, D., Vermote, R., Mayes, L. C., and Crowley, M. J. (2016). What's in a face? Mentalizing in borderline personality disorder based on dynamically changing facial expressions. *Personal. Disord.* 7, 72–79.

Luck, G., Riikkilä, K., Lartillot, O., Erkkilä, J., Toiviainen, P., Mäkelä, A., et al. (2006). Exploring relationships between level of mental retardation and features of music therapy improvisations: a computational approach. *Nord. J. Music Ther.* 15, 30–48. doi: 10.1080/08098130609478149

Marsh, K. L., Richardson, M. J., and Schmidt, R. C. (2009). Social connection through joint action and interpersonal coordination. *Top. Cogn. Sci.* 1, 320–339. doi: 10.1111/j.1756-8765.2009.01022.x

Melges, F. T., and Swartz, M. S. (1989). Oscillations of attachment in borderline personality disorder. *Am. J. Psychiatry* 146, 1115–1120. doi: 10.1176/ajp.146.9.1115

Mills, P. F., Keller, P. E., Schultz, B. G., and van der Steen, M. C. (2015). Individual differences in temporal anticipation and adaptation during sensorimotor synchronization. *Timing Time Percept.* 3, 13–31. doi: 10.1163/22134468-03002040

Müllensiefen, D. L., Gingras, B., Musil, J., and Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PLoS ONE* 9:e101091. doi: 10.1371/journal.pone.0101091

Newman, L. K., Stevenson, C. S., Bergman, L. R., and Boyce, P. (2007). Borderline personality disorder, mother–infant interaction and parenting perceptions: preliminary findings. *Austral. N. Z. J. Psychiatry* 41, 598–605. doi: 10.1080/00048670701392833

Norgaard, M. (2011). Descriptions of improvisational thinking by artist-level jazz musicians. *J. Res. Music Educ.* 59, 109–127. doi: 10.1177/0022429411405669

Norgaard, M. (2014). How jazz musicians improvise. *Music Percept.* 31, 271–287. doi: 10.1525/mp.2014.31.3.271

Odell-Miller, H. (2011). Value of music therapy for people with personality disorders. *Ment. Health Pract.* 14, 34–35. doi: 10.7748/mhp2011.07.14.10.34.c8579

Pearce, N. (2016). Analysis of matched case-control studies. *Br. Med. J.* 352:969. doi: 10.1136/bmj.i969

Pecenka, N., and Keller, P. E. (2009). Auditory pitch imagery and its relationship to musical synchronization. *Ann. N. Y. Acad. Sci.* 1169, 282–286. doi: 10.1111/j.1749-6632.2009.04785.x

Peeters, G. (2013). *Roadmap for Music Information ReSearch*. Creative Commons BY-NC-ND 3.0 license.

Plitt, H. (2012). *Intersubjektivität Erleben: Musiktherapie als Chance für Borderline-Patienten*. Marburg: Tectum Wissenschaftsverlag.

Rankin, S. K., Large, E. W., and Fink, P. W. (2009). Fractal tempo fluctuation and pulse prediction. *Music Percept.* 26, 401–413. doi: 10.1525/mp.2009.26.5.401

Repp, B. H. (2001). Processes underlying adaptation to tempo changes in sensorimotor synchronization. *Hum. Mov. Sci.* 20, 277–312. doi: 10.1016/S0167-9457(01)00049-5

Repp, B. H. (2005). Sensorimotor synchronization: a review of the tapping literature. *Psychon. Bull. Rev.* 12, 969–992. doi: 10.3758/BF03206433

Repp, B. H. (2011). Tapping in synchrony with a perturbed metronome: the phase correction response to small and large phase shifts as a function of tempo. *J. Mot. Behav.* 43, 213–227. doi: 10.1080/00222895.2011.561377

Repp, B. H., and Keller, P. E. (2008). Sensorimotor synchronization with adaptively timed sequences. *Hum. Mov. Sci.* 27, 423–456. doi: 10.1016/j.humov.2008.02.016

Repp, B. H., and Su, Y. H. (2013). Sensorimotor synchronization: a review of recent research (2006–2012). *Psychon. Bull. Rev.* 20, 403–452. doi: 10.3758/s13423-012-0371-2

Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R. L., and Schmidt, R. C. (2007). Rocking together: dynamics of intentional and unintentional interpersonal coordination. *Hum. Mov. Sci.* 26, 867–891. doi: 10.1016/j.humov.2007.07.002

Roepke, S., Vater, A., Preißler, S., Heekeren, H. R., and Dziobek, I. (2013). Social cognition in borderline personality disorder. *Front. Neurosci.* 6:195. doi: 10.3389/fnins.2012.00195

Schmidt, R. C., and Turvey, M. T. (1994). Phase-entrainment dynamics of visually coupled rhythmic movements. *Biol. Cybern.* 70, 369–376. doi: 10.1007/BF00200334

Schotte, C. K., de Doncker, D., Vankerckhoven, C., Vertommen, H., and Cosyns, P. (1998). Self-report assessment of the DSM-IV personality disorders. Measurement of trait and distress characteristics: the ADP-IV. *Psychol. Med.* 28, 1179–1188. doi: 10.1017/S0033291798007041

Schumacher, K., and Calvet, C. (2007). "Entwicklungspsychologisch orientierte Kindermusiktherapie – am Beispiel der ≪ Synchronisation ≫ als relevantes Moment," in *Kindermusiktherapie*, eds U. Stiff and R. Tüpker (Göttingen: Vandenhoeck and Ruprecht), 27–61.

Scott, L. N., Levy, K. N., and Pincus, A. L. (2009). Adult attachment, personality traits, and borderline personality disorder features in young adults. *J. Pers. Disord.* 23, 258–280. doi: 10.1521/pedi.2009.23.3.258

Seres, I., Unoka, Z., and Kéri, S. (2009). The broken trust and cooperation in borderline personality disorder. *Neuroreport* 20, 388–392. doi: 10.1097/WNR.0b013e328324eb4d

Sevdalis, V., and Keller, P. E. (2011). Perceiving performer identity and intended expression intensity in point-light displays of dance. *Psychol. Res.* 75, 423–434. doi: 10.1007/s00426-010-0312-5

Shaver, P. R., and Mikulincer, M. (2005). Attachment theory and research: resurrection of the psychodynamic approach to personality. *J. Res. Pers.* 39, 22–45. doi: 10.1016/j.jrp.2004.09.002

Silk, K. R. (2000). "Overview of biological factors," in *The Psychiatric Clinics of North America: Borderline Personality Disorder,* ed J. Paris (Philadelphia, PA: W. B. Saunders), 61–75.

Sinnaeve, R., van den Bosch, L. M., and van Steenbergen-Weijenburg, K. M. (2015). Change in interpersonal functioning during psychological interventions for borderline personality disorder—a systematic review of measures and efficacy. *Pers. Ment. Health* 9, 173–194. doi: 10.1002/pmh.1296

Skodol, A. E., Gunderson, J. G., Pfohl, B., Widiger, T. A., Livesley, W. J., and Siever, L. J. (2002). The borderline diagnosis I: psychopathology, comorbidity, and personality structure. *Biol. Psychiatry* 51, 936–950. doi: 10.1016/S0006-3223(02)01324-0

Stamatatos, E., and Widmer, G. (2005). Automatic identification of music performers with learning ensembles. *Artif. Intell.* 165, 37–56. doi: 10.1016/j.artint.2005.01.007

Steinbeis, N., and Koelsch, S. (2009). Understanding the intentions behind man-made products elicits neural activity in areas dedicated to mental state attribution. *Cereb. Cortex* 19, 619–623. doi: 10.1093/cercor/bhn110

Stern, A. (1938). Psychoanalytic investigation and therapy in the borderline group of neuroses. *Psychoanal. Q.* 7, 467–489.

Strehlow, G., and Lindner, R. (2016). Music therapy interaction patterns in relation to borderline personality disorder (BPD) patients. *Nord. J. Music Ther.* 25, 134–158. doi: 10.1080/08098131.2015.1011207

Trondalen, G. (2006). "Bedeutsame Momente" in der Musiktherapie bei jungen Menschen mit Anorexia nervosa. *Musikther. Umschau* 2, 131–144.

Trondalen, G., and Skårderud, F. (2007). Playing with affects. *Nord. J. Music Ther.* 16, 100–111. doi: 10.1080/08098130709478180

Unoka, Z., Seres, I., Aspan, N., Bodi, N., and Keri, S. (2009). Trust game reveals restricted interpersonal transactions in patients with borderline personality disorder. *J. Pers. Disord.* 23, 399–409. doi: 10.1521/pedi.2009.23.4.399

Walton, A. E., Richardson, M. J., Langland-Hassan, P., and Chemero, A. (2015). Improvisation and the self-organization of multiple musical bodies. *Front. Psychol.* 6:313. doi: 10.3389/fpsyg.2015.00313

Washburn, A., DeMarco, M., de Vries, S., Ariyabuddhiphongs, K., Schmidt, R. C., Richardson, M. J., et al. (2014). Dancers entrain more effectively than non-dancers to another actor's movements. *Front. Hum. Neurosci.* 8:800. doi: 10.3389/fnhum.2014.00800

Waters, E., Merrick, S., Treboux, D., Crowell, J., and Albersheim, L. (2000). Attachment security in infancy and early adulthood: a twenty-year longitudinal study. *Child Dev.* 71, 684–689. doi: 10.1111/1467-8624.00176

Weertman, A., Arntz, A., and Kerkhofs, M. (2000). Structured clinical interview for DSM-IV personality disorders–Dutch Version. *PsycTESTS Dataset*. doi: 10.1037/t07828-000

Wheatley, T., Kang, O., Parkinson, C., and Looser, C. E. (2012). From mind perception to mental connection: synchrony as a mechanism for social understanding. *Soc. Personal. Psychol. Compass* 6, 589–606. doi: 10.1111/j.1751-9004.2012.00450.x

Wheeler, B., and Kenny, C. (2005). "Principles of qualitative research," in *Music Therapy Research. Quantitative and Qualitative Perspectives, 2nd Edn*, ed B. Wheeler (Gilsum, NH: Barcelona Publishers), 59–71.

Widiger, T. A., and Costa, P. T. (2002). "FFM personality disorder research," in *Personality Disorders and the Five-Factor Model of Personality 2nd Edn,* eds P. T. Costa, and T. A. Widiger (Washington, DC: American Psychological Association), 59–87.

Widmer, G. (2002). Machine discoveries: a few simple, robust local expression principles. *J. New Music Res.* 31, 37–50. doi: 10.1076/jnmr.31.1.37.8103

Widmer, G., and Goebl, W. (2004). Computational models of expressive music performance: the state of the art. *J. New Music Res.* 33, 203–216. doi: 10.1080/0929821042000317804

Wiltermuth, S. S., and Heath, C. (2009). Synchrony and cooperation. *Psychol. Sci.* 20, 1–5. doi: 10.1111/j.1467-9280.2008.02253.x

Wosch, T., and Wigram, T. (2007). *Microanalysis in Music Therapy: Methods, Techniques and Applications for Clinicians, Researchers, Educators and Students*. London: Jessica Kingsley Publishers.

Zimmerman, M., Sheeran, T., and Young, D. (2004). The diagnostic inventory for depression: a self-report scale to diagnose DSM-IV major depressive disorder. *J. Clin. Psychol.* 60, 87–110. doi: 10.1002/jclp.10207

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 03 August 2017
doi: 10.3389/fpsyg.2017.01255

Toward Studying Music Cognition with Information Retrieval Techniques: Lessons Learned from the OpenMIIR Initiative

Sebastian Stober *

Machine Learning in Cognitive Science Lab, Research Focus Cognitive Sciences, University of Potsdam, Potsdam, Germany

As an emerging sub-field of music information retrieval (MIR), music imagery information retrieval (MIIR) aims to retrieve information from brain activity recorded during music cognition–such as listening to or imagining music pieces. This is a highly interdisciplinary endeavor that requires expertise in MIR as well as cognitive neuroscience and psychology. The OpenMIIR initiative strives to foster collaborations between these fields to advance the state of the art in MIIR. As a first step, electroencephalography (EEG) recordings of music perception and imagination have been made publicly available, enabling MIR researchers to easily test and adapt their existing approaches for music analysis like fingerprinting, beat tracking or tempo estimation on this new kind of data. This paper reports on first results of MIIR experiments using these OpenMIIR datasets and points out how these findings could drive new research in cognitive neuroscience.

Keywords: music cognition, music perception, music information retrieval, deep learning, representation learning

OPEN ACCESS

Edited by:
Naresh N. Vempala,
Ryerson University, Canada

Reviewed by:
Michael Casey,
Dartmouth College, United States
Jean-Julien Aucouturier,
Centre National de la Recherche Scientifique (CNRS), France

***Correspondence:**
Sebastian Stober
sstober@uni-potsdam.de

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *24 October 2016*
Accepted: *10 July 2017*
Published: *03 August 2017*

Citation:
Stober S (2017) Toward Studying Music Cognition with Information Retrieval Techniques: Lessons Learned from the OpenMIIR Initiative. Front. Psychol. 8:1255. doi: 10.3389/fpsyg.2017.01255

1. INTRODUCTION

Music Information Retrieval (MIR) is a relatively young field of research that has emerged over the course of the last two decades. It brings together researchers from a large variety of disciplines who—in the broadest sense—investigate methods to retrieve and interact with music information. As the MIR community has grown, research questions also have become more diverse. The different kinds of data considered in MIR now comprise, for instance, symbolic representations, audio recordings, sheet music, playlists, (social) web data such as reviews or tweets, and usage meta-data.

As a very recent development, MIR researchers also have started to explore ways to detect and extract music information from brain activity recorded during listening to or imagining music pieces–a sub-field of MIR introduced as Music Imagery Information Retrieval (MIIR) in Stober and Thompson (2012). In the long term, research in this direction might lead to new ways of searching for music along the line of existing MIR approaches that, for instance, allow query by singing, humming, tapping, or beat-boxing. Inspired by recent successes in reconstructing visual stimuli (Miyawaki et al., 2008; Nishimoto et al., 2011; Cowen et al., 2014) and even dream imagery (Horikawa et al., 2013), it might eventually be possible to even reconstruct music stimuli from recorded brain activity.

In a broader context, Kaneshiro and Dmochowski (2015), for instance, mention transcription, tagging and annotation, audience following, and portable MIR applications as possible scenarios that could benefit from neuroimaging data such as EEG. Findings from MIIR can further support the development of Brain-Computer Interfaces (BCIs) that facilitate interaction with music in new

ways beyond basic search–such as Brain-Computer Music Interfaces (BCMIs) used to generate and control music (Miranda and Castet, 2014). Finally and most importantly, this paper aims to motivate an MIIR-driven approach to music cognition research that can lead to new insights about on how the human brain processes and encoders music.

The challenge of retrieving music information from recordings of brain activity can in principle be approached in the following naïve way: One could argue that as the brain processes perceived music or recreates this experience in imagination, it generates a transformed representation which is captured—to some extend—by the recording equipment. Hence, the recorded signal could in principle be seen as a mid-level representation of the original music piece that has been heavily distorted by two consecutive black-box filters—the brain and the recording equipment. This transformation involves and intermingles with several other brain processes unrelated to music perception and is further limited by the capabilities of the recording equipment which might additionally introduces signal artifacts.

This setting calls for sophisticated signal processing techniques, ideally developed in an intense interdisciplinary collaboration between MIR researchers and neuroscientists. In order to facilitate such a collaboration and contribute to new developments in this emerging field of research, the first OpenMIIR dataset was released as public domain in 2015 (Stober et al., 2015b). This article summarizes work over the course of a year since its publication. To this end, a brief overview of the dataset is provided in section 2. as well as related research in section 3. As the main part of this paper, section 4 covers our experiments. This is followed by a discussion in section 5. Finally, we draw conclusions and point out directions for future work in section 6.

2. THE OPENMIIR DATASET

The OpenMIIR dataset (Stober et al., 2015b) comprises Electroencephalography (EEG) recordings taken during music perception and imagination.[1] These data were collected from 10 subjects who listened to and imagined 12 short music fragments—each 7–16 s long—taken from well-known pieces. EEG was chosen as recording technique because it is much more accessible to MIR researchers than Magnoencephalography (MEG) and functional Magnetic Resonance Imaging (fMRI), with more and more affordable consumer-level devices becoming available. Furthermore, EEG has a good temporal resolution that can capture how music perception and imagination unfold over time and allows for analyzing temporal characteristics of the signal such as rhythmic information. The stimuli were selected from different genres and systematically span several musical dimensions such as meter, tempo, and the presence of lyrics. This way, various retrieval and classification scenarios can be addressed. As shown in **Table 1**, there are 3 groups with 4 stimuli each.

1. Stimuli 1–4 are from recordings of songs where a singing voice (lyrics) is present.
2. Stimuli 11–14 are from different recordings of the same songs as stimuli 1–4. These recordings do not contain a singing voice. Instead, the melody is played by one or more instruments.
3. Stimuli 21–24 are from recordings of purely instrumental pieces that do not have any lyrics and thus it is not possible to sing along.

All stimuli were normalized in volume and kept as similar in length as possible with care taken to ensure that they all contained complete musical phrases starting from the beginning of the piece. The pairs of recordings for the same song with and without lyrics were tempo-matched. The stimuli were presented to the participants in several conditions while EEG was recorded.

1. Stimulus perception with cue clicks
2. Stimulus imagination with cue clicks
3. Stimulus imagination without cue clicks
4. Stimulus imagination without cue clicks, with additional feedback from participants after each trial

Condition 1–3 trials were recorded directly back-to-back. The goal was to lock time and tempo between conditions 1 and 2 through the cue to help identifying overlapping features. Conditions 3 and 4 simulate a more realistic query scenario where the system cannot know the tempo and meter in advance. The presentation was divided into 5 blocks that each comprised all 12 stimuli in randomized order. In total, 60 trials (12 stimuli $\times$ 5 blocks) per condition were recorded for each subject.

EEG was recorded from 10 participants (3 male), aged 19–36, with normal hearing and no history of brain injury. A BioSemi Active-Two system was used with 64 + 2 EEG channels sampled at 512 Hz. Horizontal and vertical Electrooculography (EOG) channels were recorded to capture eye movements. The following common-practice pre-processing steps were applied to the raw EEG and EOG data using the MNE-python toolbox by Gramfort et al. (2013) to remove unwanted artifacts. We removed and interpolated bad EEG channels (between 0 and 3 per subject) identified by manual visual inspection. The data was then filtered with a bandpass keeping a frequency range between 0.5 and 30 Hz. This also removed any slow signal drift in the EEG. To remove artifacts caused by eye blinks, we computed independent components using extended Infomax independent component analysis (ICA) as described by Lee et al. (1999) and semi-automatically removed components that had a high correlation with the EOG channels. Afterwards, the 64 EEG channels were reconstructed from the remaining independent components without reducing dimensionality. Furthermore, the data of one participant was excluded for the experiments described in this paper because of a considerable number of trials with movement artifacts due to coughing. Finally, all trial channels were additionally normalized to zero mean and range $[-1, 1]$.

[1]The dataset is available at https://github.com/sstober/openmiir

TABLE 1 | Information about the tempo, meter, and length of the stimuli (without cue clicks).

ID	Group	Name	Meter	Length	Tempo in beats per minute (BPM)
1	Songs recorded with lyrics	Chim Chim Cheree	3/4	13.3 s	212
2		Take Me Out to the Ballgame	3/4	7.7 s	189
3		Jingle Bells (lyrics)	4/4	9.7 s	200
4		Mary Had a Little Lamb	4/4	11.6 s	160
11	Songs recorded without lyrics	Chim Chim Cheree	3/4	13.5 s	212
12		Take Me Out to the Ballgame	3/4	7.7 s	189
13		Jingle Bells	4/4	9.0 s	200
14		Mary Had a Little Lamb	4/4	12.2 s	160
21	Instrumental pieces	Emperor Waltz	3/4	8.3 s	178
22		Hedwig's Theme (Harry Potter)	3/4	16.0 s	166
23		Imperial March (Star Wars Theme)	4/4	9.2 s	104
24		Eine Kleine Nachtmusik	4/4	6.9 s	140
Mean				10.4 s	176

3. RELATED WORK

Retrieval based on brain wave recordings is still a very young and largely unexplored domain. EEG signals have been used to recognize emotions induced by music perception (Lin et al., 2009; Cabredo et al., 2012) and to distinguish perceived rhythmic stimuli (Stober et al., 2014). It has been shown that oscillatory neural activity in the gamma frequency band (20–60 Hz) is sensitiv to accented tones in a rhythmic sequence (Snyder and Large, 2005) and that oscillations in the beta band (20–30 Hz) increase in anticipation of strong tones in a non-isochronous sequence (Fujioka et al., 2009, 2012; Iversen et al., 2009). While listening to rhythmic sequences, the magnitude of steady state evoked potentials (SSEPs), i.e., reflecting neural oscillations entrained to the stimulus, changes for frequencies related to the metrical structure of the rhythm as a sign of entrainment to beat and meter (Nozaradan et al., 2011, 2012).

EEG studies by Geiser et al. (2009) have further shown that perturbations of the rhythmic pattern lead to distinguishable electrophysiological responses–commonly referred to as Event-Related Potentials (ERPs). This effect appears to be independent of the listener's level of musical proficiency. Furthermore, Vlek et al. (2011) showed that imagined auditory accents imposed on top of a steady metronome click can be recognized from ERPs. However, as usual for ERP analysis to deal with noise in the EEG signal and reduce the impact of unrelated brain activity, this requires averaging the brain responses recorded for many events. In contrast, retrieval scenarios usually only consider single trials. Nevertheless, findings from ERP studies can guide the design of single-trial approaches as demonstrated in subsection 4.1.

EEG has also been successfully used to distinguish perceived melodies. In a study conducted by Schaefer et al. (2011), 10 participants listened to 7 short melody clips with a length between 3.26 and 4.36 s. For single-trial classification, each stimulus was presented for a total of 140 trials in randomized back-to-back sequences of all stimuli. Using quadratically regularized linear logistic-regression classifier with 10-fold cross-validation, they were able to successfully classify the ERPs of single trials. Within subjects, the accuracy varied between 25 and 70%. Applying the same classification scheme across participants, they obtained between 35 and 53% accuracy. In a further analysis, they combined all trials from all subjects and stimuli into a grand average ERP. Using singular-value decomposition, they obtained a fronto-central component that explained 23% of the total signal variance. The related time courses showed significant differences between stimuli that were strong enough for cross-participant classification. Furthermore, a correlation with the stimulus envelopes of up to .48 was observed with the highest value over all stimuli at a time lag of 70–100 ms.

Results from fMRI studies by Herholz et al. (2012) and Halpern et al. (2004) provide strong evidence that perception and imagination of music share common processes in the brain, which is beneficial for training MIIR systems. As Hubbard (2010) concludes in his review of the literature on auditory imagery, *"auditory imagery preserves many structural and temporal properties of auditory stimuli"* and *"involves many of the same brain areas as auditory perception"*. This is also underlined by Schaefer (2011, p. 142) whose *"most important conclusion is that there is a substantial amount of overlap between the two tasks* [music perception and imagination]*, and that 'internally' creating a perceptual experience uses functionalities of 'normal' perception."* Thus, brain signals recorded while listening to a music piece could serve as reference data for a retrieval system in order to detect salient elements in the signal that could be expected during imagination as well.

A recent meta-analysis of Schaefer et al. (2013) summarized evidence that EEG is capable of detecting brain activity during the imagination of music. Most notably, encouraging preliminary results for recognizing purely imagined music fragments from EEG recordings were reported in Schaefer et al. (2009) where 4 out of 8 participants produced imagery that was classifiable (in a

binary comparison) with an accuracy between 70 and 90% after 11 trials.

Another closely related field of research is the reconstruction of auditory stimuli from EEG recordings. Deng et al. (2013) observed that EEG recorded during listening to natural speech contains traces of the speech amplitude envelope. They used ICA and a source localization technique to enhance the strength of this signal and successfully identify heard sentences. Applying their technique to imagined speech, they reported statistically significant single-sentence classification performance for 2 of 8 subjects with performance increasing when several sentences were combined for a longer trial duration.

More recently, O'Sullivan et al. (2015) proposed a method for decoding attentional selection in a cocktail party environment from single-trial EEG recordings of approximately one minute length. In their experiment, 40 subjects were presented with 2 classic works of fiction at the same time—each one to a different ear—for 30 trials. In order to determine which of the 2 stimuli a subject attended to, they reconstructed both stimuli envelopes from the recorded EEG. To this end, they trained two different decoders per trial using a linear regression approach—one to reconstruct the attended stimulus and the other to reconstruct the unattended one. This resulted in 60 decoders per subject. These decoders where then averaged in a leave-on-out cross-validation scheme. During testing, each decoder would predict the stimulus with the best reconstruction from the EEG using the Pearson correlation of the envelopes as measure of quality. Using subject-specific decoders averaged from 29 training trials, the prediction of the attended stimulus decoder was correct for 89% of the trials whereas the mean accuracy of the unattended stimulus decoder was 78.9%. Alternatively, using a grand-average decoding method that combined the decoders from every other subject and every other trial, they obtained a mean accuracy of 82 and 75% respectively.

4. EXPERIMENTS

Our initial analyses of the OpenMIIR recordings was largely exploratory. Hence, the following subsections cover three very different approaches:

1. ERP-inspired single-trial analysis (subsection 4.1),
2. reconstruction of the audio stimulus envelope from the EEG (subsection 4.2), and
3. extraction of stimulus-related brain activity from the EEG recordings (subsection 4.3).

These approaches increase in complexity, ranging from hand-crafted design to representation learning, i.e., a machine learning pipeline that also includes learning suitable features from the raw EEG data.[2]

The experiments were implemented in Python with the exception of the Matlab code for the tempo estimation experiment described in subsubsection 4.3.5. For neural network training, the framework Theano (Al-Rfou et al., 2016) was used in combination with Blocks and Fuel (van Merriënboer et al., 2015). The code to run the experiments and to generate the plots shown in this paper is made available as open source and linked from the OpenMIIR website. As the OpenMIIR dataset is public domain, this assures full reproducibility of the results presented here.

4.1. ERP-Inspired Single-Trial Tempo Analysis

Our first experiment was inspired by traditional ERP analysis but also incorporated autocorrelation as a common MIR approach to tempo estimation (e.g., Ellis, 2007). This experiment has been described in detail in Sternin et al. (2015). Recordings from 5 participants were used that were available at this point in time. Additionally to the pre-processing steps described in section 2, the EEG recordings were down-sampled to 64 Hz.

4.1.1. Initial ERP-Analysis

We started with a basic ERP analysis and focused on the trials recorded for conditions 1–3. Beat annotations were obtain for all beats within the audio stimuli using the dynamic beat tracker described in Ellis (2007) and provided by the *librosa* library.[3] To this end, the beat tracker was initialized with the known tempo of each stimulus. The quality of the automatic annotations was verified through sonification.

Given the beat annotations of the stimuli and assuming that the participants would imagine the stimuli at a similar tempo in conditions 2 and 3, we computed bar-aligned ERPs using non-overlapping epochs from 100 ms before to 2.4 s after a downbeat annotation. This length was required to capture slightly more than a single bar for the slowest stimulus–number 23 with a bar length of more than 2.3 s. As expected, the resulting averaged ERPs differed considerably between participants, stimuli, and conditions. Nevertheless, we often observed a periodicity in the averaged signal proportional to the bar length. **Figure 1** shows example ERPs for a specific participant and stimulus where this is clearly visible in all conditions.

In order to analyze this periodicity, we computed the autocorrelation curves by comparing each signal with itself at a range of time lags. To this end, we aggregated all 64 EEG channels into a mean signal. We further chose time lags corresponding to the bar tempo range of the stimuli. The lower end of 24 BPM was determined by the choice of the epoch length. Using longer epochs would allow for extending the tempo range to slower tempi, but this would be at the expense of fewer epochs available for averaging.

4.1.2. Limitations and Potential Pitfalls

In general, more distinct peaks in the autocorrelation were observed in the perception condition. For the two imagination conditions, peaks were more blurred as can also be seen in **Figure 1**. This is most likely caused by the lack of a time locking mechanism, which allows the imagination tempo to vary—causing bar onsets to deviate from the stimulus-based annotations. This hypothesis is also backed by the observation

[2] An introduction and overview of representation learning is, for instance, provided by Bengio et al. (2013).

[3] https://github.com/bmcfee/librosa

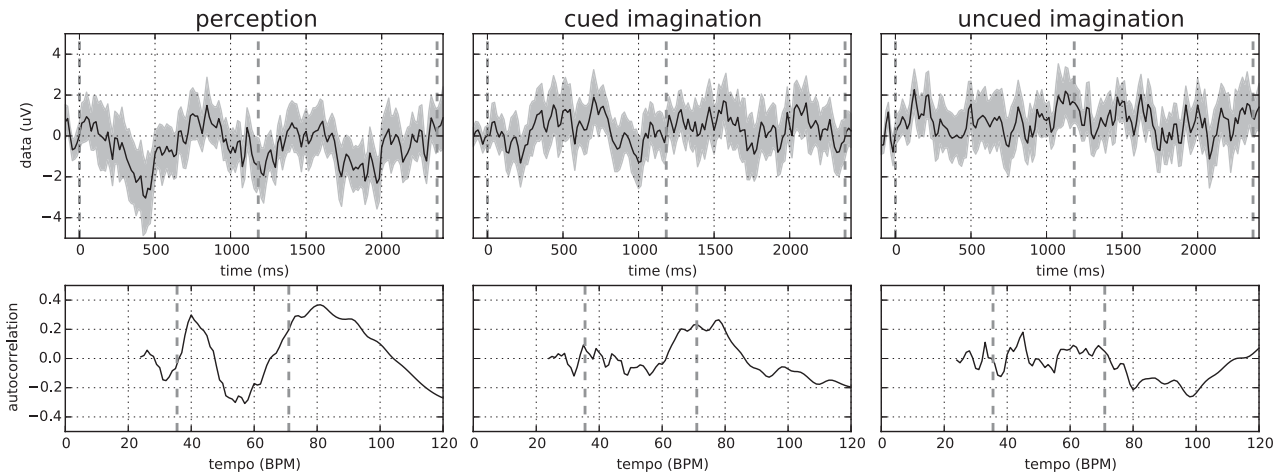


FIGURE 1 | (Top) Mean and standard deviation over all 64 EEG channels of the bar-aligned ERPs (without epochs overlap) for "Chim Chim Cheree (lyrics)" in conditions 1–3 for participant P01. Each ERP was averaged over 25 epochs (5 from each trial). **(Bottom)** Corresponding autocorrelation scores in the relevant tempo range. Dashed lines indicate downbeats **(Top)** and the approximate bar tempo of the stimulus plus its lower tempo harmonic **(Bottom)**. Originally published in Sternin et al. (2015).

that artificially jittering the bar onsets results in a decrease in autocorrelation.

The computation of the ERPs benefits from a constant stimulus tempo. The tempo values provided in **Table 1** refer to the initial tempo which is also used by the tempo cue. In some stimuli, however, the tempo is not exactly constant but changes slightly over time. In stimulus 22, for instance, the tempo temporarily drops after the first half of the theme at around 8 s. Such deviations further impact the quality of bar-aligned ERPs because of the variable timing within the individual bars.

As a very important detail of the bar-aligned ERP analysis, it is essential to ensure that the bar-aligned epochs do not overlap by rejecting some of the epochs. If they overlap, a single data segment can contribute to multiple epochs at different time points. This can induce misleading autocorrelation peaks that are not supported by the raw data.

4.1.3. Analyzing Single Trials

Based on the ERP-based observations, the question was whether the tempo could similarly be estimated through autocorrelation from single trials. This posed several challenges. First, there were too few bar-aligned epochs in a single trial to use ERPs. Second, neither the tempo of the stimulus nor the beat annotations should be known a priori in a realistic setting. Therefore, there were no reference points for extracting bar-aligned epochs. Moreover, the problem of possible tempo variance in the imagination conditions needed to be addressed.

Figure 2 illustrates our proposed solution to this problem. A 2.5-seconds sliding window is moved over the mean EEG signal aggregated over all channels. At each position with a hop size of 5 samples at 64 Hz, an autocorrelation curve is computed. The curves for the individual window segments are stacked into a two-dimensional matrix with the first dimension corresponding to the window offset in the trial signal and the second dimension corresponding to the possible tempo values. Hence, each matrix value holds the score for a certain tempo at one specific point in the trial. The scores in the matrix are finally aggregated deriving an estimated tempo value for the trial. While the mean and maximum over all matrix rows often produced significant peaks in the aggregated autocorrelation curve as illustrated in **Figure 2**, the following heuristic has led to slightly more stable results:

1. In each row, find the pair of tempo values with the maximal combined score.
2. Select the median of all selected pairs.
3. From this pair, return the tempo value with the higher mean value over all rows.

For the evaluation of our approach, we computed the mean absolute error of the estimated tempo and the actual tempo. We also considered the tempo harmonic below and above the correct value, i.e., half or twice the tempo, as a correct result. The prediction error, averaged across all stimuli, varied considerably between participants ranging from 7.07, 7.15, and 8.11 in the three conditions for participant P14 up to 9.81, 10.04, and 12.58 for P12. Furthermore, the results clearly showed a trend that tempo was easier to predict for some stimuli, such as "Chim Chim Cheree" (ID 1 and 11) and "Mary had a little lamb" (ID 4 and 14), than for others. The slowest stimulus, the "Imperial March" (ID 23) had the highest variation of prediction accuracy. These initial results eventually encouraged further research into estimating the stimulus tempo from the EEG using more sophisticated signal processing techniques. This is further described in subsubsection 4.3.5.

4.2. Audio Stimulus Envelope Reconstruction

In our second experiment, we attempted to reconstruct the audio stimulus envelopes from the EEG signals, i.e., reversing the

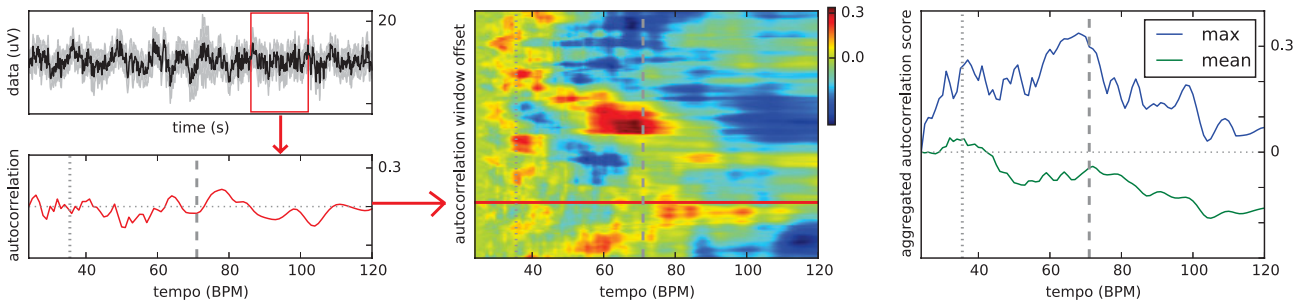


FIGURE 2 | Schema of the proposed tempo estimation technique. All plots refer to the first of the five trials contributing to the imagination ERP of "Chim Chim Cheree (lyrics)" in **Figure 1**, middle. **Left top:** EEG waveform (mean of all 64 channels) for the whole trial with the red box indicating the sliding window of 2.5 s. **Left bottom:** Autocorrelation curve for this specific segment of the trial. **Middle:** Vertically stacked autocorrelation curves for the whole trial with the red horizontal line indicating the position of the sliding window shown on the left. **Right:** Aggregated autocorrelation scores (mean and max) for the whole trial. Dashed vertical lines indicate the stimulus bar tempo. Dotted vertical lines refer to half the bar tempo. Originally published in Sternin et al. (2015).

black-box signal transformations by the brain and the recording equipment. An approximate reconstruction of the envelopes would be a very useful feature for further retrieval steps such as beat tracking, tempo prediction or stimulus identification. Furthermore, it could also be directly sonified—for instance by shaping a white-noise base signal with the up-sampled envelope. This would be helpful for analysis and for interactive scenarios like brain-computer interfacing where (auditory) feedback is desirable.

Using the method described by O'Sullivan et al. (2015), we attempted to reconstruct and classify the audio envelopes shown in **Figure 3**. These envelopes were computed by applying the Hilbert transform to the mono audio signal of the stimuli, down-sampling to 64 Hz and low-pass filtering at 8 Hz. The EEG recordings were also down-sampled to 64 Hz matching the envelope sample rate. This rate was chosen to reduce dimensionality and thus limit the number of regression parameters.

The linear reconstruction technique used in O'Sullivan et al. (2015) learns a filter matrix with individual weights for each channel at a range of time lags based on the cross-correlation between the EEG channels and the stimulus envelope. This matrix is then used to convolve the EEG signal to produce the reconstructed stimulus envelope. The size of the matrix and thus the number of parameters to be fit depends on the number of EEG channels and the maximum time lag to be considered.

Directly applying this technique did unfortunately not lead to satisfying results. For the trial-specific decoders, the correlation of the reconstruction and the stimulus envelope was only 0.11 on average with a very high variance of 0.52. Results were also very unstable, i.e., minimally changing the length of the time-lag window generally resulted in very different decoder weights. This eventually produced very poor results when decoder matrices were added together during training, rendering them useless for classification.

We suspect two main reasons for this outcome: Firstly, the trials might be too short for the algorithm to produce stable decoder matrices and secondly, the music envelopes differ significantly from those for speech. We tried to address the second point by using envelopes computed from filtered stimuli versions that emphasized the main voice (using an "inverse-karaoke" filter as described in Duda et al., 2007) and artificial "beat envelopes" derived from the beat and downbeat annotations shown in **Figure 3**. However, this did not lead to an improvement.

Limiting the maximum time lag to 375 ms and reducing the number of channels through PCA, we were able to reduce the number of parameters and the resulting tendency of over-fitting the filter matrix to the training data. However, the envelope reconstruction quality remained very poor and the resulting (leave-one-out) classification accuracy was not statistically significant. Based on these observations, we concluded that the tested approach which worked well for speech reconstruction is not transferable to our music stimuli. We hypothesize that this is caused by the lack of signal sparsity of the music stimuli.

4.3. Extracting Music-Related Brain Activity

This experiment aimed to extract brain activity that is related to stimulus perception and imagination using techniques from the field of deep representation learning. Note that this is a much broader focus than the attempted stimulus envelope reconstruction from the previous experiment. Naturally, any EEG signal component correlated with the stimulus envelope would be related to stimulus perception or imagination. But there is potentially much more brain activity that is also related to the music stimuli but not directly helpful for their reconstruction.

The basic pre-processing steps briefly described in section 2 aimed to improve the general signal quality by removing common EEG artifacts. However, there is still the problem that the EEG naturally also records brain activity that is unrelated to music perception or imagination. These signals can be considered as noise with respect to the specific focus of interest. Separating this background noise from the music-related brain activity is a very challenging task. Figuratively speaking, this could be compared to a cocktail-party situation where a listener would like to attend to a specific speaker in a room with many independently ongoing conversations. As an additional complication, the listener is not in the same room as the speakers

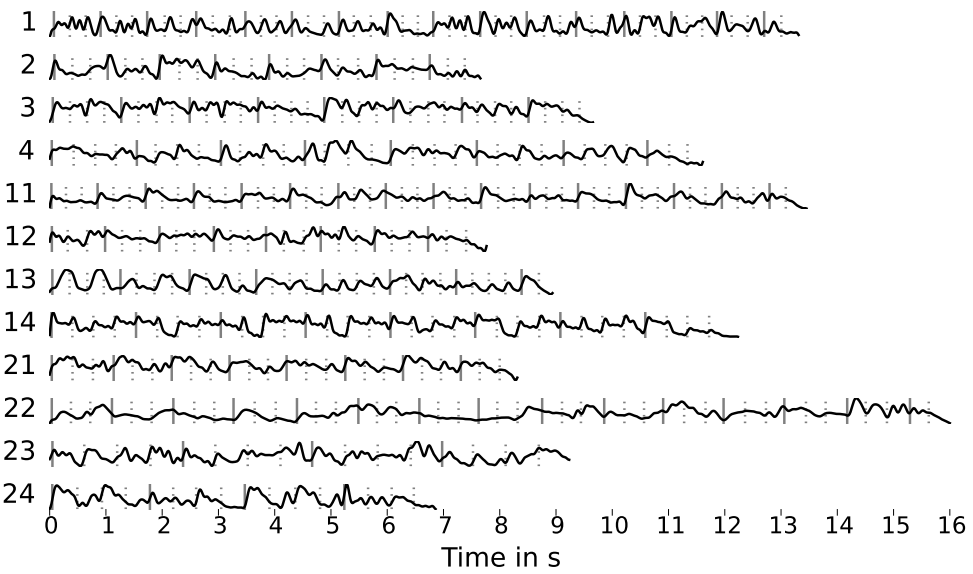


FIGURE 3 | Stimulus envelopes (sampled at 64 Hz, low-pass filtered at 8 Hz) with markers for beats (dashed lines) and downbeats (solid lines) obtained using the dynamic beat tracker by Ellis (2007) as part of the *librosa* library.

but in the next room separated by a thick wall—analogously to the EEG equipment that can only measure brain activity from the outside with the skull in between.

This challenge calls for sophisticted signal processing techniques. One newly emerging option is using so-called deep artificial neural networks that over the last decade have become very popular in various application domains such as computer vision, automatic speech recognition, natural language processing, and bioinformatics where they produce state-of-the-art results on various tasks. These networks are able to learn (hierarchies of increasingly) complex features from raw data which is referred to as (deep) representation learning. The learned feature representations can then be used to solve machine learning problems such as a classification tasks. We hypothesized that this approach could also be applied to EEG analysis.

The main problem with applying deep neural networks for EEG analysis is the limited amount of data for training. If all perception trials are clipped to match the length of the shortest stimulus, excluding the cue clicks, the total amount of EEG data recorded for the perception conditions is 63 min from 540 trials. At the same time, each trial has more than 225,000 dimensions at the original sampling rate of 512 Hz. This is very unlike the typical scenarios where deep neural networks are successful.[4] In such a setting with potentially many network parameters (due to the number of input dimensions) and only a small set of training instances, the neural net is very likely to overfit. I.e., it adapts too much to the training data which results in a poor generalization performance.

We addressed this challenge by focusing on small nets that have few model parameters and by developing a special pre-training technique called similarity-constraint encoding for representation learning. The series of representation learning experiments that eventually led to this technique is described in detail in Stober et al. (2015a). In the following, we summarize the main idea.

4.3.1. Similarity-Constraint Encoding

The idea of similarity-constraint encoding (SCE) is derived from auto-encoder pre-training (Bengio et al., 2007). An auto-encoder is a neural network that is trained to reconstruct its inputs while its internal representation is limited to make this a non-trivial task—for instance, through a structural bottleneck or regularization of weights or activations. Additionally, the inputs can be corrupted by adding random noise which can result in more robust features (Vincent et al., 2010). This approach has been successfully applied for learning compressed feature representations—usually during an unsupervised pre-training phase—in many domains such as for learning high-level image features (Le, 2013), coding speech spectrograms (Deng et al., 2010) or sentiment analysis (Socher et al., 2011).

EEG data already contain noise from various sources. Furthermore, only a small portion of the recorded brain activity is usually relevant in the context of an experiment. Given only a small dataset, a basic auto-encoder would learn features that represent the full EEG data including noise and irrelevant brain activity. This limits the usefulness of the learned features. For better features, the encoding needs to be more selective. To this end, side information can be used. Demanding that trials belonging to the same class[5] are encoded similarly facilitates learning features representing brain activity that is stable across trials. Features to be used in classification tasks should

[4]For comparison, a 224-by-224 RGB image in the Imagenet dataset has roughly 150,000 dimensions—about two-thirds of the size of the EEG trials. However, Imagenet contains millions of labeled images for training.

[5]There are several ways to assign the trials to classes based on the stimulus meta-data such as the stimulus id, the meter, or the presence or absence of lyrics.

furthermore allow for distinguishing between the respective classes. This can be achieved by a training objective that also considers how trials from other classes are encoded.

In the most basic form, the encoded representations of two trials belonging to the same class are compared with an encoded trial from a different class. The desired outcome of this comparison can be expressed as a *relative similarity constraint* as introduced in Schultz and Joachims (2003). A relative similarity constraint (a, b, c) describes a relative comparison of the trials a, b, and c in the form "a is more similar to b than a is to c." Here, a is the *reference trial* for the comparison. There exists a vast literature on using such constraints to learn similarity measures in general and for applications within MIR specifically (Lübbers and Jarke, 2009; McFee and Lanckriet, 2010; Stober, 2011; Wolff and Weyde, 2014). Based on this formalization, we define a cost function for learning a feature encoding by combining all pairs of trials (a, b) from the same class with all trials c belonging to different classes and demanding that a and b are more similar. The resulting set of trial triplets is then used to train a similarity-constraint encoder network as illustrated in **Figure 4**.

All trials within a triplet that constitutes a similarity constraint are processed using the same encoder pipeline. This results in three internal feature representations. Based on these, the reference trial is compared with the paired trial and the trial from the other class resulting in two similarity scores. We use the dot product as similarity measure because this matches the way patterns are compared in a neural network classifier and it is also suitable to compare time series. More complex approaches are possible as well, as long as they allow training through backpropagation. The output layer of the similarity constraint encoder finally predicts the trial with the highest similarity score without further applying any additional affine transformations. The whole network can be trained like a common binary classifier, minimizing the error of predicting the wrong trial as belonging to the same class as the reference. The only trainable part is the shared encoder pipeline. This pipeline can be arbitrarily complex—e.g., also include recurrent connections within the pipeline.

After pre-training, the output of the encoder pipeline can be used as feature representation to train a classifier for identifying the actual classes (in contrast to the artificially constructed binary classification problem for pre-training). Alternatively, as we will show later, the features could also be used to train a classifier for different classes than the ones originally used to construct the triplets during pre-training.

4.3.2. Encoder Pipeline and Classifiers

For all SCE experiments described in the following, the encoder pipeline consisted of a single convolutional layer with a single filter and without a bias term. This filter aggregated the 64 raw EEG channels into a single waveform processing one sample (over all channels) at a time. I.e. it had the shape 64×1 (channels $\times$ samples) and thus a very small number of parameters. The hyperbolic tangent (tanh) was used as activation function because its output range matched the value range of the network inputs ($[-1,1]$). No pooling was applied. The number of network and learning hyper parameters was kept as low as possible to minimize their impact.

A linear support vector machine classifier (SVC) was trained using Liblinear (Fan et al., 2008) on

- baseline (1): the raw EEG data,
- baseline (2): the averaged EEG data (mean over all channels as a naïve filter), and
- the output of the pre-trained encoder pipeline.

With this setting, an increase in the stimulus classification accuracy over the baselines can be attributed to a reduction of the signal-to-noise ratio by the encoder pipeline. This could then be interpreted as evidence that the encoder has successfully picked up music-related brain activity.

As additional classifier, a simple neural network (NN) was trained on the encoder pipeline output. This network consisted of a single fully-connected layer with a Softmax non-linearity. No bias term was used. This resulted in one temporal pattern learned for each of the classes, which could then be analyzed. For further comparison, we also trained and end-to-end neural network that had the same structure as the encoder pipeline combined with the neural network classifier but was initialized randomly instead of pre-training. All tested methods are listed in **Table 2**.

4.3.3. Training and Evaluation Scheme

A nested cross-validation scheme as shown in **Figure 5** was chosen that allowed for using each one of the 540 trials for testing once. The outer 9-fold cross-validation was performed across subjects, training on 8 and testing on the 9th subject. The inner 5-fold cross-validation was used for model selection based on 1 of the 5 trial blocks. Training was divided into two phases.

In the first phase, the encoder pipeline was trained using the proposed similarity-constraint encoding technique with the hinge loss as cost function. Stochastic gradient descent (SGD) with a batch size of 1,000 and the Adam (Kingma and Ba, 2014) step rule was used. Training was stopped after 10 epochs and the model with the lowest binary classification error on the validation triplets was selected. Triplets were constructed such that all trials within a triplet belonged to the same subject as the simple encoder pipeline likely could not easily compensate inter-subject differences. The validation triplets consisted of a reference trial from the validation trials and the other two trials drawn from the combined training and validation set of the inner cross-validation. This way, a reasonable number of validation triplets could be generated without sacrificing too many trials for validation.[6] The final encoder filter weights were computed as mean of the 5-fold models. The output of this filter was used to compute the features for the second training phase.

In the second phase, the two classifiers were trained. For the SVC, the optimal value for the parameter C that controls the trade-off between the model complexity and the proportion of non-separable training instance was determined through a grid search during the inner cross-validation. For the neural

[6] At least 2 of the 5 trials per class and subject are required to construct within-subject triplets.

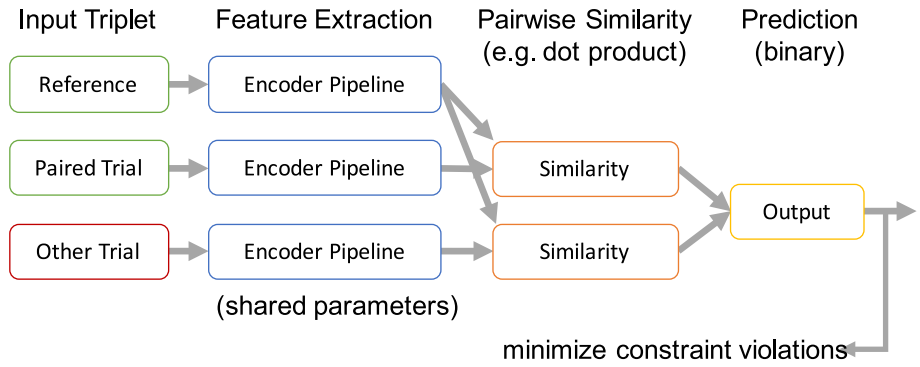


FIGURE 4 | Processing scheme of a similarity-constraint encoder. Originally published in Stober (2017).

TABLE 2 | Accuracies for the three classification tasks: stimulus (12 classes), group (3 classes) and meter (2 classes).

Classifier	Features	Classification Accuracy & Significance		
		Stimulus (12) (%)	Group (3) (%)	Meter (2) (%)
Chance of correct classification for a single trial		8.33	33.33	50.00
Chance accuracy at $p = 0.001$ for 560 trials w.r.t. cumulative binomial distribution		12.22	39.63	56.67
SVC	Raw EEG	18.52 ***	40.37 **	62.04 **
SVC	Raw EEG channel mean	12.41 ****	38.70 ***	58.52 ****
End-to-end NN	Raw EEG	18.15 ***	37.41 ****	60.56 ***
Dummy	Output of stimulus classifier		38.89 ***	59.63 ***
SVC (reference)	Stimulus SCE features	27.59	48.89	69.44
NN	Stimulus SCE features	27.22	48.89	67.78
SVC	Group SCE features		35.37 ****	
NN	Group SCE features		34.63 ****	
SVC	Meter SCE features			60.19 ***
NN	Meter SCE features			58.88 ****

*Chance accuracy values are provided for comparison. Significance levels are indicated against the best performing approach (highlighted in red) using McNemar's tests ($n = 540$, mid-p). **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$.*

network classifier, 5 fold models were trained for 100 epochs using SGD with batch size 120, the Adam step rule, and the hinge loss as cost function. The best models were selected based on the classification performance on the validation trials and then averaged to obtain the final classifier.

4.3.4. Stimulus Identification

As first classification task, we investigated stimulus identification. There are 12 perfectly balanced classes–one for each stimulus. **Figure 6** shows the filters learned in the pre-training phase of each outer cross-validation fold as well as the standard deviation and ranges for the weights within the 5 inner cross-validation folds. The filter weights only differed in small details across folds. However, sometimes the polarity of the weights had flipped. To avoid cancellation effects during aggregation, the polarity was normalized based on the sign of the weight for channel T7 (next to the left ear), which always had a high absolute value.

The magnitude of the channel weights in the pre-trained filters (which are further aggregated in **Figure 7**) indicates how much the respective EEG channels are contributing to the aggregated signal. The electrodes within the dark red areas that appear bilaterally towards the back of the head lie directly over the auditory cortex. These electrodes may be picking up on brain activation from the auditory cortex that is modulated by the perception of the stimuli. The electrodes within the blue areas that appear more centrally may be picking up on the cognitive processes that occur as a result of the brain processing the music at a higher level.

However, as pointed out by Haufe et al. (2014), model parameters for classification or decoding should not be directly interpreted in terms of the brain activity as they depend on all noise components in the data, too. Instead, a forward model should be derived that explains how the measured signals were generated from the neural sources. We applied the proposed regression approach and trained a deconvolutional filter that

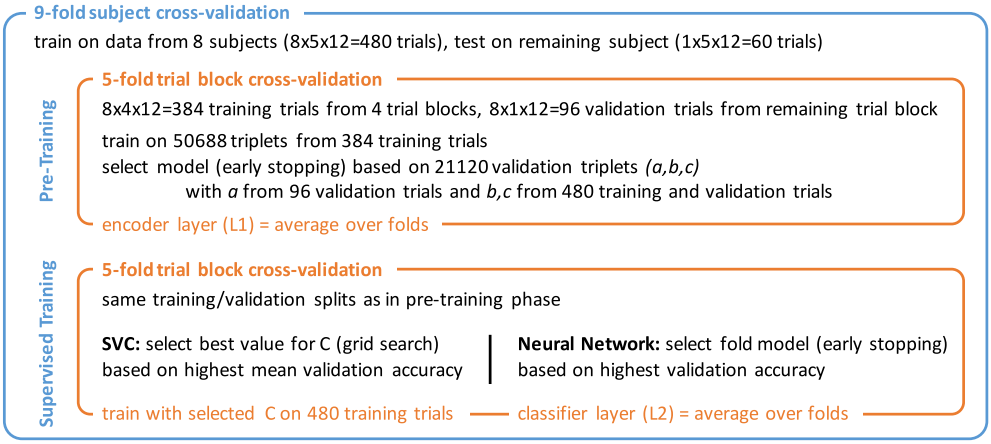


FIGURE 5 | Nested cross-validation scheme with pre-training and supervised training phase. Triplet numbers for SCE pre-training refer to the 12-class stimulus identification task.

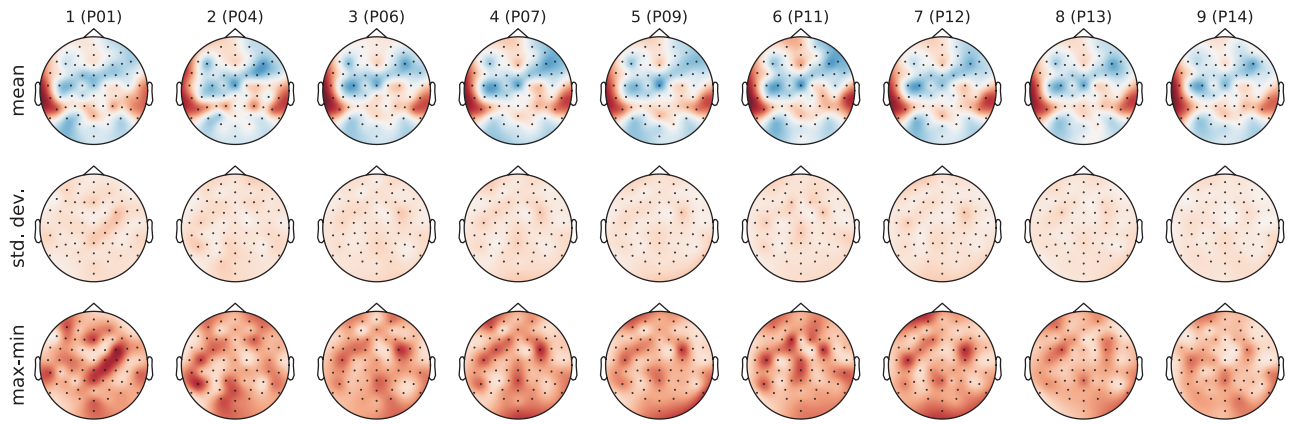


FIGURE 6 | Topographic visualization of the learned filter weights aggregated over the 5 inner cross-validation folds within each outer cross-validation fold. **(Top)** mean. **(Middle)** standard deviation. **(Bottom)** range (maximum–minimum). Columns correspond to outer cross-validation folds with the id of the test subject as column label. The mean filters in the top row were used to compute the features for the supervised training phase. All plots use the same color map and range.

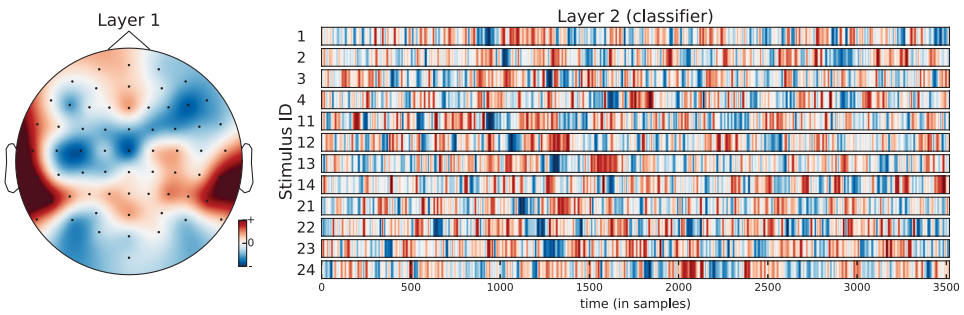


FIGURE 7 | Visualization of the average neural network parameters (from the 9 outer cross-validation folds) for stimulus classification. Layer 1: mean of convolutional layers from the pre-trained encoders (SCE), i.e., mean over the top row in **Figure 6**. The filter weights only differed in small details across folds. Layer 2: mean of classifier layers trained in the supervised phase.

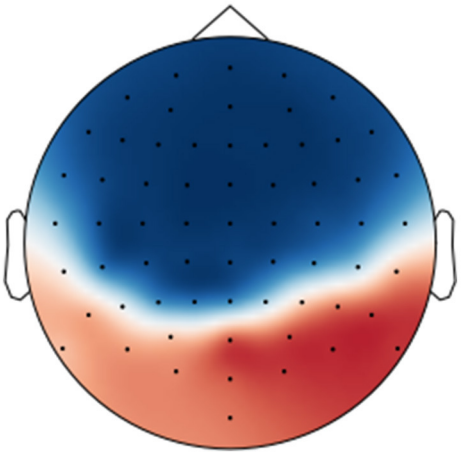


FIGURE 8 | Visualization of the forward model (deconvolutional filter) trained to reconstruct the originally recorded EEG signals from the encoder output over all perception trials.

reconstructs the original EEG signal from the encoder output by minimizing the mean squared error between the reconstructed and the actual signal over all trials. For each trial, we used the encoder from the respective outer cross-validation fold. The resulting deconvolutional filter is shown in **Figure 8**.

Table 2 (column "stimulus") lists the classification accuracy for the tested approaches. Remarkably, all values were significantly above chance. Even for baseline 2, the value of 12.41% was significant at $p = 0.001$. This significance value was determined by using the cumulative binomial distribution to estimate the likelihood of observing a given classification rate by chance. To evaluate whether the differences in the classification accuracies produced by the different methods are statistically significant, McNemar's tests using the "mid-p" variant suggested in Fagerland et al. (2013) were applied. The obtained significance levels are indicated in **Table 2** for a comparison with the best performing approach—using SVC in combination with the SCE features learned for stimulus identification. The very significant improvement of the classification accuracy over the two baselines and the neural network trained end-to-end is a strong indicator for a reduction of the signal-to-noise ratio. Notably, the pre-trained filter is very superior to the naïve filter of baseline 2 that was actually harmful judging from the drop in accuracy.

The confusion matrices for the classifiers trained on the encoder output are shown in **Figure 9**. Apart from the main diagonal, two parallel diagonals can be seen that indicate confusion between stimuli 1–4 and their corresponding stimuli 11–14, which are tempo-matched recordings of songs 1–4 without lyrics. Analyzing the averaged neural network parameters visualized in **Figure 7** shows similar temporal patterns for these stimuli pairs.[7] A detailed analysis of the network layer activations as shown in **Figure 10** reveals noticeable peaks in the encoder output and matching weights with high magnitude in the classifier layer that often coincide with downbeats–i.e., the first beat within each measure, usually with special musical emphasis. These peaks are not visible in the channel-averaged EEG (baseline 2). Thus, it can be concluded that the encoder filter has successfully extracted a component from the EEG signal that contains musically meaningful information.

Both, the systematic confusion of stimuli 1–4 with their corresponding tempo-matched versions without lyrics (stimuli 11–14) as well as the temporal patterns learned by the neural network classifier are strong indicators against a possible "horse" classifier. Sturm (2014) defines a "horse" as "a system appearing capable of a remarkable human feat [...] but actually working by using irrelevant characteristics (confounds)." In this specific context, a "horse" might base the classification on signal components unrelated to music cognition. An additional behavioral experiment where 8 subjects judged the similarity of each stimulus pair confirmed the parallel diagonals observed in the confusion matrices. Measuring the time required to recognize the individual music stimuli yielded average values of 1–3 s that did not correlate with the third-downbeat peaks in the temporal patterns of the classifier. This suggests that the peak is not related to brain activity caused by stimulus recognition but rather by musical features of the stimuli.

4.3.5. Tempo Estimation Revisited

In a follow-up experiment published in Stober et al. (2016) that also picks up the thread from our tempo analysis experiment described in subsection 4.1, we used the stimulus SCE features as input to a sophisticated tempo estimation technique provided by the Tempogram Toolbox.[8] This technique has been originally developed for analyzing audio recordings. To compute a tempogram, a given music audio signal is first transformed into a novelty curve that captures note onset information—for instance, as the positive part of a spectral flux as described in Grosche and Müller (2011a). Through a short-time Fourier analysis of the novelty curve, the audio tempogram is derived that reveals how dominant different tempi are at a given time point in the audio signal. Aggregating a tempo histogram along the time axis yields a tempo histogram where peaks indicate the predominant tempo within the piece.

We applied the same processing pipeline for the perception EEG data of participants P09 to P14 by directly interpreting the EEG signal filtered by the SCE encoder pipeline as novelty curve. We were able to observe peaks in the derived tempo histograms that sometimes highly correlated with the stimulus tempo. Averaging tempogram histograms over trials and participants overall stabilized the tempo estimation. Remarkably, results seemed to strongly depend on the music stimuli. For the first 8 stimuli (1–4 and 11–14), i.e., the songs recorded with and

[7]The average model is only for illustration and analysis. For testing, the respective outer cross-validation fold model was used for each trial.

[8]The Tempogram Toolbox contains MATLAB implementations for extracting various types of tempo and pulse related audio representations (Grosche and Müller, 2011b) A free implementation can be obtained at https://www.audiolabs-erlangen.de/resources/MIR/tempogramtoolbox.

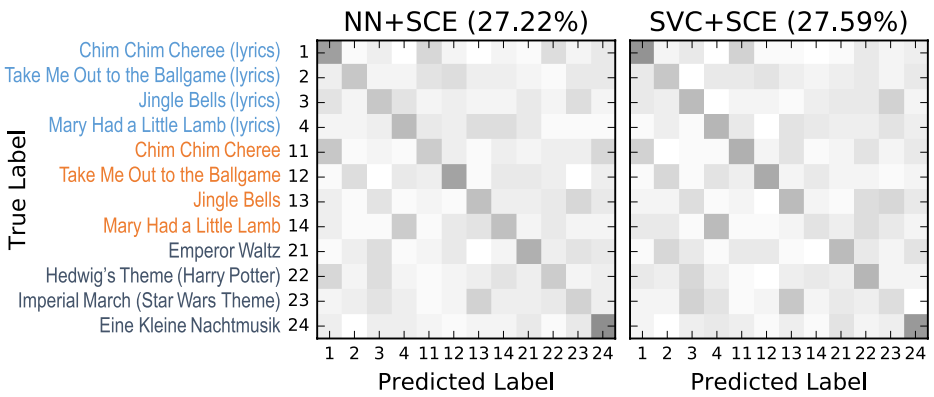


FIGURE 9 | 12-class confusion matrices for the music stimuli (listed on the left) for the classifiers trained on the stimulus SCE output. Middle: SVC. Right: Neural network classifier (NN). Results were aggregated from the 9 outer cross-validation folds ($n = 540$). Originally published in Stober (2017).

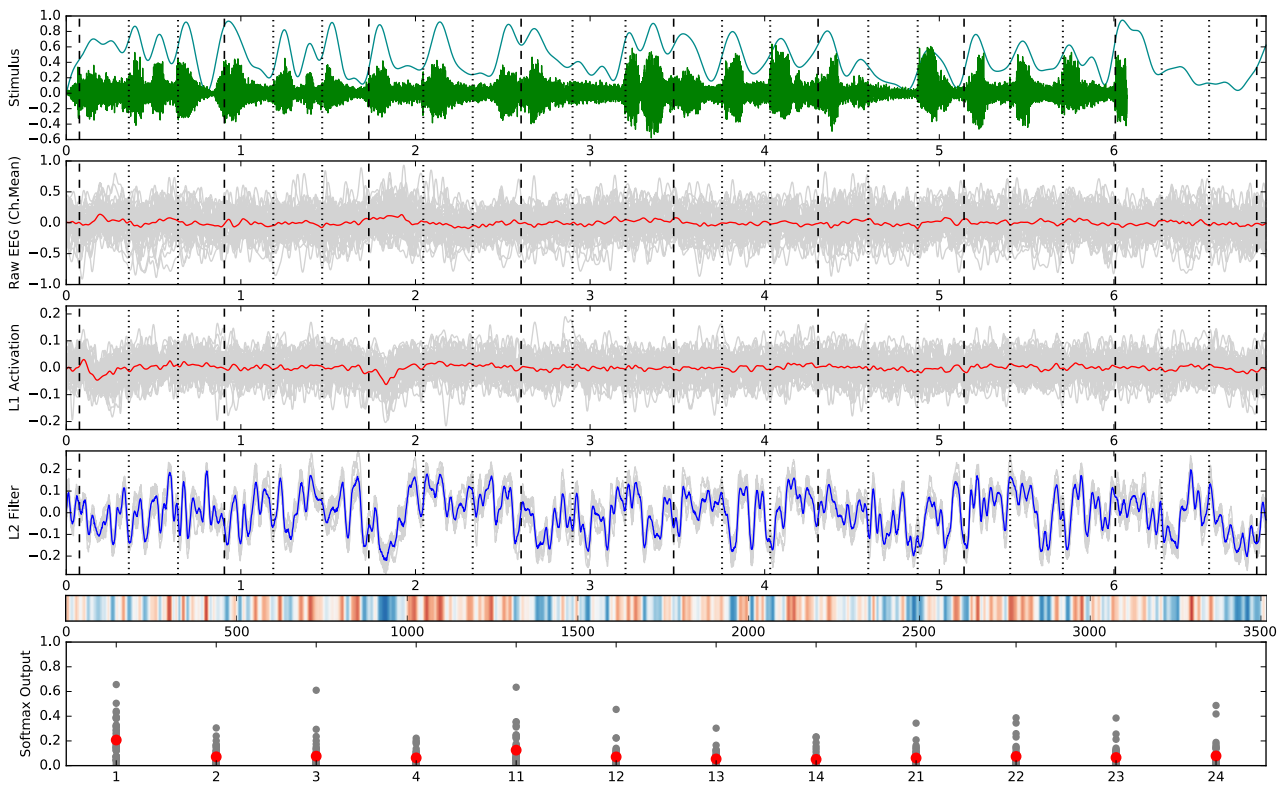


FIGURE 10 | Detailed analysis of all trials belonging to stimulus 1. Vertical marker lines indicate beats (dotted) and downbeats (dashed). The horizontal axis in rows 1–5 corresponds with the time in seconds or samples (row 5). **(Top)** Audio stimulus (green) and envelope (cyan). **(2nd row)** Raw EEG averaged over all 64 channels per trial (gray) and overall mean (red). This is identical to the SVC input for baseline 2. **(3rd row)** Encoder output (activation) for the individual trials (gray) and overall mean (red). **(4th row)** Patterns learned by the neural network classifier for this class in the 9 folds of the outer cross-validation (gray) and overall mean (blue). **(5th row)** Alternative visualization (as in **Figure 7**) of the averaged pattern from row 4. **(Bottom)** Softmax output of the neural network classifier for the individual trials (gray) and overall mean (red) with class labels on the horizontal axis. All outputs were generated using the respective test trials for each fold model in the outer cross-validation.

without lyrics, the tempo extraction seemed to work better than for the last 4 (21–24), i.e., the instrumental pieces. Exploring this effect was beyond the scope of this small study. To uncover and properly understand the underlying factors, a large-scale music perception experiment using stimuli with systematically adapted tempi would be needed. Possible reasons might be the complexity of the music stimuli, the presence of lyrics, the participants, or the applied methodology

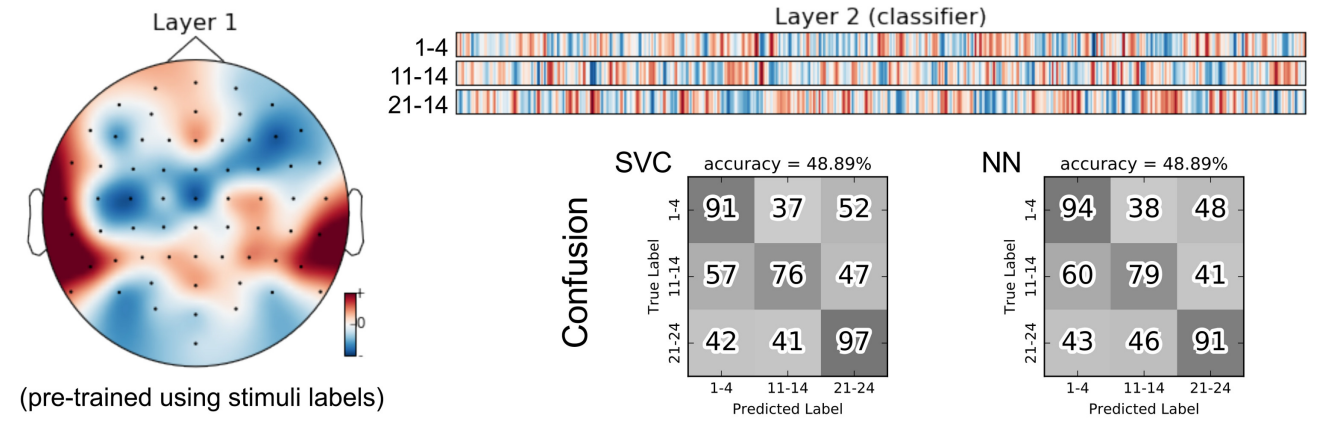


FIGURE 11 | Visualization of the average neural network parameters (from the 9 outer cross-validation folds) for group classification (stimuli 1–4: songs recorded with lyrics, stimuli 11–14: songs recorded without lyrics, stimuli 21–24: instrumental pieces). Layer 1: mean of convolutional layers from the pre-trained encoders (SCE) using the stimuli labels. Layer 2: mean of classifier layers trained in the supervised phase.

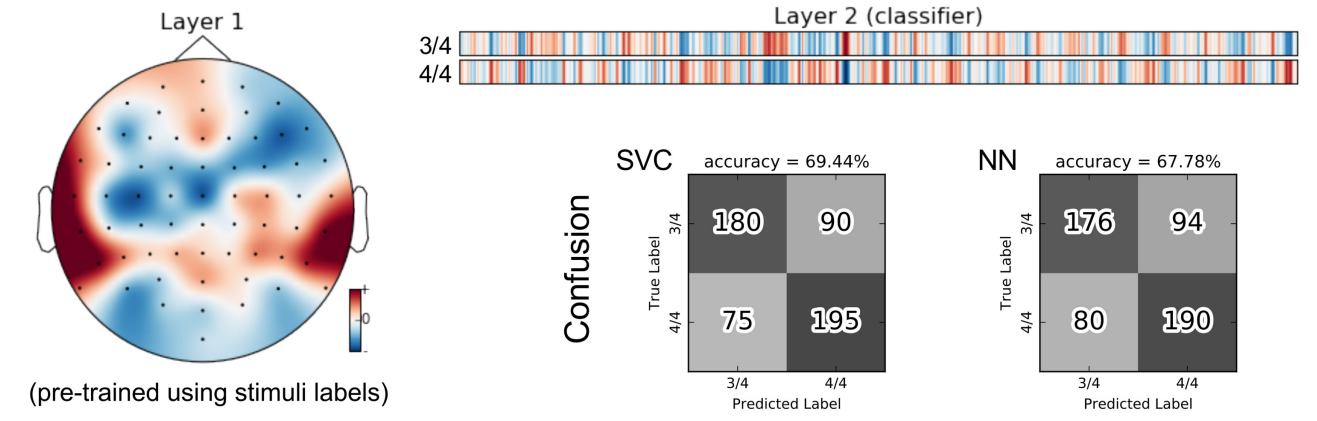


FIGURE 12 | Visualization of the average neural network parameters (from the 9 outer cross-validation folds) for meter classification. Layer 1: mean of convolutional layers from the pre-trained encoders (SCE) using the stimuli labels. Layer 2: mean of classifier layers trained in the supervised phase.

and techniques. Investigating these issues could be a starting point for interdisciplinary research between MIR and music cognition.

4.3.6. Group Classification

As described in section 2 and shown in **Table 1**, the 12 music stimuli can be grouped into 3 groups of 4 stimuli each: songs recorded with lyrics (stimuli 1–4), songs recorded without lyrics (stimuli 11–14), and instrumental pieces (stimuli 21–24). Using these three perfectly balanced classes, 184,320 training triplets and 72,960 validation triplets were available for each inner cross-validation fold during SCE pre-training. Here, the SCE pre-training did not result in a suitable feature representation as indicated by the inferior classification accuracy compared to the baselines shown in **Table 2** (column "groups"). As a likely reason, the SCE learning problem may me ill-posed, i.e., the encoder pipeline may not have been sufficiently complex to learn a transformation of the raw EEG that makes trials within groups more similar to each other than to trials from the other groups.

As an alternative, we trained the group classifiers on the feature representation from the stimulus classification task. This resulted in a substantial increase in classification accuracy of roughly 10%. We further added a "dummy" baseline classifier that just derived the group class labels from the predicted stimulus labels. The difference in accuracy indicates that the stimulus SCE features seem to capture some relevant information for the group classification task beyond what is necessary to recognize the stimuli. Similarly to **Figures 7** and **9** for the stimulus classification task, **Figure 11** shows the parameters of the neural network classifier averaged over the 9 outer cross-validation folds as well as the confusion matrices for the two group classifiers trained on the stimulus SCE features. The temporal patterns learned by the classifier are currently subject of further analysis.

4.3.7. Meter Classification

There are two perfectly balanced classes with respect to meter as half of the stimuli are in 3/4 meter and the others are in

1. Select features of interest and determine their respective values for the stimuli
   a. based on MIR feature extraction tools
   b. defined by experts
   c. determined in behavioral experiments
2. Compute a similarity matrix from pairwise stimulus similarities
3. Derive similarity constraints for training
4. Define an encoder pipeline based on hypotheses about the cognitive processes of interest. This can include:
   a. applying pre-processing techniques like transforming the signal into a time-frequency representation or computing signal components using ICA
   b. focusing on specific data such as selecting specific channels, signal components or frequency bands
   c. determining the encoder hyper parameters such as the number and kind of layers for artificial neural networks
5. Train the retrieval model (such as the encoder combined with a classifier)
6. Analyze the trained model and its performance including
   a. error and misclassification patterns (like the confusion matrices in **Figure 9**)
   b. emerging encoder parameters (like the filters visualized in **Figure 7**) as well as corresponding forward models (such as **Figure 8**)
   c. patterns in encoder output (as shown in **Figure 10**)
7. Iterate with revised features (step 1) or hypotheses about the cognitive processes (step 4)

FIGURE 13 | Outline of the proposed MIR-driven research approach using similarity-constraint encoding as a specific example.

4/4 meter. With these class labels, 211,968 training triplets and 83,520 validation triplets are available for each inner cross-validation fold during SCE pre-training. As for the group classification, the resulting feature representation is not helpful for this classification task. Instead, using the stimulus SCE features again results in the best performance that is roughly 9% higher. **Figure 12** shows the parameters of the neural network classifier averaged over the 9 outer cross-validation folds as well as the confusion matrices for the two group classifiers trained on the stimulus SCE features.

The inferior performance of the meter SCE features may again be attributed to complexity limitations of the simple convolutional encoder pipeline. We are currently investigating more complex encoders that also incorporate recurrent components to capture temporal patterns within the encoder already.

4.3.8. Classifying EEG from the Imagination Conditions

All SCE-based experiments described above focused on perception data. Applying the same pre-training technique to the data from the imagination conditions has so far not led to significant classification results or to the discovery of meaningful or interesting patterns. Also, using the encoder trained on the perception data to filter the imagination trials before training the classifier was not successful. As possible reason for this, we suspect—at least for the current encoder design—that timing and synchronization in the imagination trials are insufficiently accurate. This makes it hard to learn an encoder that produced similar temporal patterns or—given a successfully pre-trained encoder—to learn temporal patterns for classification that generalize well. Different encoder designs that can compensate temporal variance may lead to better results. This needs to be further investigated. However, focusing on the perception data for now in order to improve the analysis methods appears to be more promising.

5. DISCUSSION

5.1. Proposal of an MIR-Driven Research Approach

Based on the findings from our representation learning experiments described in subsection 4.3, we can derive the following general MIR-driven approach to analyzing music perception and imagination data as outlined in **Figure 13**. We start by choosing a specific music feature—that necessarily has to be present in the respective music stimuli—and attempt to retrieve it from the recorded brain signals. Representation learning techniques like similarity-constraint encoding allow for finding signal filters that extract relevant components from the recorded brain signals given that we have chosen a suitable encoder pipeline. This choice should be hypothesis-driven and informed by findings from cognitive neuroscience. If the trained encoder pipeline indeed improves the signal-to-noise ratio and consequently the retrieval performance, this can be seen as supporting evidence for the hypothesis that guided the encoder design. Analyzing the emerging network parameters and activation patterns might further allow for learning more about the underlying cognitive processes. Failure could be attributed to poor encoder design choices and question the underlying hypothesis, or it could be caused by limitations of the dataset.

(For instance, there might be a bias within the dataset caused by the choice of the stimuli or the participants.) The impact of the latter should naturally be minimized through the study design.

5.2. Interpretation of Temporal Classifier and Activation Patterns

The neural networks trained so far seem still simple enough to allow for interpretation of the learned parameters by domain experts and facilitate findings about the cognitive processes. Most remarkably, the temporal patterns learned by the neural network classifier for the stimulus identification task show prominent signal peaks at the third downbeat (i.e., the beginning of the third bar) for almost all stimuli. They can be clearly recognized in the visualization of the averaged model parameters in **Figure 7**. There are also noticeable matching peaks in the encoder filter activation as shown in **Figure 10** for one of the stimuli. This raises the question which cognitive process could explain these patterns and calls for further investigation by domain experts from music cognition.

5.3. Lessons Learned from the OpenMIIR Study Design

In the way it has been used so far, similarity-constraint encoding imposes a strong regularization assumption that requires a very tight synchronization of the trials to identify good filter parameters. This is problematic if synchronization between the stimuli and the recorded EEG signals is poor like in the imagination conditions. Different encoder designs–for instance including temporal pooling operations–might be able to compensate the lack of tight synchronizations. But generally, it seems very desirable for representation learning to ensure synchronization by experimental design in the first place. In the design of our study to collect the first OpenMIIR dataset, we decided against having a metronome click for synchronization during imagination trials in order to avoid artifacts caused by the audio stimulation. It seems now like the downside of having such artifacts is outweighed by the possible benefits of tightly controlling the imagination tempo. Of course, the added metronome clicks in the background would have to be exactly identical in tempo, loudness etc. for all stimuli. Otherwise, they would easily allow for distinguishing the stimuli by a "horse" classifier.[9] Hence, all stimuli would need to be in the same tempo (or multiples) as the click.

Another issue is the variable length of the stimuli caused by using full musical phrases in the original study. We ended up cutting all trials to the length of the shortest one for our representation learning approach. Zero-padding the shorter trials instead would have easily given away their identity leading to useless feature representations. To avoid recording likely unused EEG data, it seems more desirable to have equal-length trials–even as this means to stop in the middle of a musical phrase. Having tempo-synced stimuli makes finding good cut points already easier.

Furthermore, switching to imagination trials with a metronome click would also rule out conditions 3 and 4 as listed in section 2. With half as many conditions, already twice as many trials could be recorded in the same time. Additionally reducing the number of stimuli and the stimuli length would allow for further increasing the number of trials per class and condition. This could allow us to collect enough data for learning within-subject feature representations and retrieval models. Based on these considerations, we are currently designing a follow-up OpenMIIR study to collect another EEG dataset.

[9]Cf. Section 4.3.4 for a discussion of the "horse" phenomenon.

6. CONCLUSIONS AND OUTLOOK

Less than four years have passed since the subject of MIIR was first discussed during the "Unconference" (Anglade et al., 2013) at the International Society of Music Information Retrieval Conference (ISMIR) in 2012. ISMIR 2016 already featured a well-attended tutorial on the "Introduction to EEG Decoding for Music Information Retrieval Research" and for the first time, the annual seminar on Cognitively based Music Informatics Research (CogMIR) was co-located as a satellite event which drew the attention of many main-conference attendees. This is evidence for the increasing interest within the MIR community to combine MIR and music cognition research.

The goal of the OpenMIIR initiative is to foster interdisciplinary exchange and collaborations between these two fields. To this end, we introduced the OpenMIIR dataset in 2015—an public-domain EEG dataset intended to enable MIR researchers to venture into the domain of music imagery and develop novel methods without the need for special EEG equipment. This paper summarized our findings from a first series of largely exploratory experiments addressing several MIIR tasks with this dataset. For some tasks—especially when working with data from the imagination conditions—our approaches failed or did not perform as expected. We have hypothesized why this might be the case and derived ideas for a follow-up EEG study to collect a second dataset.

A first success of our efforts is our proposed similarity-constraint encoding approach for extracting music-related brain activity of EEG recordings. Using this technique, we were able to train simple spatial filters that significantly improve the signal-to-noise ratio for the perception data in several classification tasks. There is a lot of potential for improving the classification accuracy by using more complex encoders that possibly comprise multiple layers of neurons and recurrent connections. Investigating such options is one major direction of our ongoing research efforts. We have also obtained encouraging first results by applying MIR techniques from the Tempogram Toolbox for estimating the stimulus tempo from the perception EEG recordings. This experiment nicely showcases how well-established MIR techniques for music audio analysis can also be applied to music cognition data.

We hope that our work described here inspires other MIR researchers to try their methods in this emerging interdisciplinary field and encourages music cognition researchers to share their datasets and engage in an exchange with the MIR community. Everybody interested is invited to

contribute and collaborate within the OpenMIIR initiative. Further information about the OpenMIIR initiative can be found at https://openmiir.github.io where apart from the OpenMIIR dataset itself, the code to run the described experiments is shared and constantly being updated.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this article and approved it for publication.

FUNDING

The OpenMIIR dataset and most of the work presented here are the result of an ongoing joint research effort between the Owen Lab and the Music and Neuroscience Lab at the Brain and Mind Institute of the University of Western Ontario, and the Machine Learning in Cognitive Science Lab at the University of Potsdam. This work has been supported by a fellowship within the Postdoc-Program of the German Academic Exchange Service (DAAD), the Canada Excellence Research Chairs (CERC) Program, an National Sciences and Engineering Research Council (NSERC) Discovery Grant, an Ontario Early Researcher Award, the James S. McDonnell Foundation, and the donation of a Geforce GTX Titan X graphics card from the NVIDIA Corporation. The work involving the Tempogram Toolbox in collaboration with the the International Audio Laboratories Erlangen—a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS—was supported by the German Research Foundation (DFG MU 2686/6-1, DFG MU 2686/7-1). The publication of this article was supported by the German Research Foundation (DFG) and the Open Access Publishing Fund of the University of Potsdam.

ACKNOWLEDGMENTS

The author would like to thank all collaborators who were involved in the work presented here: Avital Sternin, Jessica A. Grahn, Adrian M. Owen, Thomas Prätzlich, and Meinard Müller. Further thanks go to Dawn Pavich, Siegfried Wrobel, and Haitao Yang for their administrative and technical support, and to all the study participants.

REFERENCES

Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., et al. (2016). Theano: a python framework for fast computation of mathematical expressions. *arXiv:1605.02688*.

Anglade, A., Humphrey, E., Schmidt, E., Stober, S., and Sordo, M. (2013). Demos and late-breaking session of the thirteenth international society for music information retrieval conference (ismir 2012). *Comput. Music J.* 37, 91–93. doi: 10.1162/COMJ_r_00171

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *Patt. Anal. Mach. Intell. IEEE Trans.* 35, 1798–1828. doi: 10.1109/TPAMI.2013.50

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). "Greedy layer-wise training of deep networks," in *Advances in Neural Information Processing Systems 19*.

Cabredo, R., Legaspi, R. S., Inventado, P. S., and Numao, M. (2012). "An emotion model for music using brain waves," in *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012, Mosteiro S. Bento Da VitÃşria*, eds F. Gouyon, P. Herrera, L. G. Martins, and M. MÃijller (Porto), 265–270.

Cowen, A. S., Chun, M. M., and Kuhl, B. A. (2014). Neural portraits of perception: reconstructing face images from evoked brain activity. *Neuroimage* 94, 12–22. doi: 10.1016/j.neuroimage.2014.03.018

Deng, L., Seltzer, M. L., Yu, D., Acero, A., Mohamed, A.-R., and Hinton, G. E. (2010). *Binary Coding of Speech Spectrograms Using a Deep Auto-Encoder*. Interspeech.

Deng, S., Srinivasan, R., and D'Zmura, M. (2013). *Cortical Signatures of Heard and Imagined Speech Envelopes*. Technical report, DTIC.

Duda, A., Nürnberger, A., and Stober, S. (2007). "Towards query by singing/humming on audio databases," in *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007*, eds S. Dixon, D. Bainbridge, and R. Typke (Vienna), 331–334.

Ellis, D. P. W. (2007). Beat tracking by dynamic programming. *J. New Music Res.* 36, 51–60. doi: 10.1080/09298210701653344

Fagerland, M., Lydersen, S., and Laake, P. (2013). The mcnemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Med. Res. Methodol.* 13:1. doi: 10.1186/1471-2288-13-91

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: a library for large linear classification. *J. Mach. Learn. Res.* 9, 1871–1874.

Fujioka, T., Trainor, L. J., Large, E. W., and Ross, B. (2009). Beta and gamma rhythms in human auditory cortex during musical beat processing. *Anna. N.Y. Acad. Sci.* 1169, 89–92. doi: 10.1111/j.1749-6632.2009.04779.x

Fujioka, T., Trainor, L. J., Large, E. W., and Ross, B. (2012). Internalized timing of isochronous sounds is represented in neuromagnetic beta oscillations. *J. Neurosci.* 32, 1791–1802. doi: 10.1523/JNEUROSCI.4107-11.2012

Geiser, E., Ziegler, E., Jancke, L., and Meyer, M. (2009). Early electrophysiological correlates of meter and rhythm processing in music perception. *Cortex* 45, 93–102. doi: 10.1016/j.cortex.2007.09.010

Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., et al. (2013). MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* 7:267. doi: 10.3389/fnins.2013.00267

Grosche, P., and Müller, M. (2011a). Extracting predominant local pulse information from music recordings. *IEEE Trans. Audio Speech Lang. Process.* 19, 1688–1701. doi: 10.1109/TASL.2010.2096216

Grosche, P., and Müller, M. (2011b). "Tempogram toolbox: MATLAB implementations for tempo and pulse analysis of music recordings," in *Late-Breaking News of the International Society for Music Information Retrieval Conference (ISMIR)* (Miami, FL).

Halpern, A. R., Zatorre, R. J., Bouffard, M., and Johnson, J. A. (2004). Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia* 42, 1281–1292. doi: 10.1016/j.neuropsychologia.2003.12.017

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., et al. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* 87, 96–110. doi: 10.1016/j.neuroimage.2013.10.067

Herholz, S., Halpern, A., and Zatorre, R. (2012). Neuronal correlates of perception, imagery, and memory for familiar tunes. *J. Cogn. Neurosci.* 24, 1382–1397. doi: 10.1162/jocn_a_00216

Horikawa, T., Tamaki, M., Miyawaki, Y., and Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science* 340, 639–642. doi: 10.1126/science.1234330

Hubbard, T. L. (2010). Auditory imagery: empirical findings. *Psychol. Bull.* 136, 302–329. doi: 10.1037/a0018436

Iversen, J. R., Repp, B. H., and Patel, A. D. (2009). Top-down control of rhythm perception modulates early auditory responses. *Anna. N.Y. Acad. Sci.* 1169, 58–73. doi: 10.1111/j.1749-6632.2009.04579.x

Kaneshiro, B., and Dmochowski, J. P. (2015). "Neuroimaging methods for music information retrieval: current findings and future prospects," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR'15)*, 538–544.

Kingma, D., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv:1412.6980*.

Le, Q. (2013). "Building high-level features using large scale unsupervised learning," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8595–8598.

Lee, T., Girolami, M., and Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Comput.* 11, 417–441. doi: 10.1162/089976699300016719

Lin, Y.-P., Jung, T.-P., and Chen, J.-H. (2009). "EEG dynamics during music appreciation," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009. EMBC 2009* (Minneapolis, MN: IEEE), 5316–5319. doi: 10.1109/IEMBS.2009.5333524

Lübbers, D., and Jarke, M. (2009). "Adaptive multimodal exploration of music collections," in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR'09)*, 195–200.

McFee, B., and Lanckriet, G. R. G. (2010). "Metric learning to rank," in *Proceedings of the 27th International Conference on Machine Learning (ICML'10)* (Haifa), 775–782.

Miranda, E. R., and Castet, J. (eds.). (2014). *Guide to Brain-Computer Music Interfacing*. London: Springer.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-A., Morito, Y., Tanabe, H. C., et al. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron* 60, 915–929. doi: 10.1016/j.neuron.2008.11.004

Nishimoto, S., Vu, A., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol* 21, 1641–1646. doi: 10.1016/j.cub.2011.08.031

Nozaradan, S., Peretz, I., Missal, M., and Mouraux, A. (2011). Tagging the neuronal entrainment to beat and meter. *J. Neurosci. Off. J. Soc. Neurosci.* 31, 10234–10240. doi: 10.1523/JNEUROSCI.0411-11.2011

Nozaradan, S., Peretz, I., and Mouraux, A. (2012). Selective neuronal entrainment to the beat and meter embedded in a musical rhythm. *J. Neurosci.* 32, 17572–17581. doi: 10.1523/JNEUROSCI.3203-12.2012

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., et al. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* 25, 1697–1706. doi: 10.1093/cercor/bht355

Schaefer, R. S. (2011). *Measuring the Mind's Ear EEG of Music Imagery*. PhD thesis, S.l., Nijmegen.

Schaefer, R. S., Blokland, Y., Farquhar, J., and Desain, P. (2009). "Single trial classification of perceived and imagined music from EEG," in *Proceedings of the 2009 Berlin BCI Workshop* (Berlin).

Schaefer, R. S., Desain, P., and Farquhar, J. (2013). Shared processing of perception and imagery of music in decomposed EEG. *Neuroimage* 70, 317–326. doi: 10.1016/j.neuroimage.2012.12.064

Schaefer, R. S., Farquhar, J., Blokland, Y., Sadakata, M., and Desain, P. (2011). Name that tune: decoding music from the listening brain. *Neuroimage* 56, 843–849. doi: 10.1016/j.neuroimage.2010.05.084

Schultz, M., and Joachims, T. (2003). "Learning a distance metric from relative comparisons," in *Advances in Neural Information Processing Systems*, 41–48.

Snyder, J. S., and Large, E. W. (2005). Gamma-band activity reflects the metric structure of rhythmic tone sequences. *Cogn. Brain Res.* 24, 117–126.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011). "Semi-supervised recursive autoencoders for predicting sentiment distributions," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics), 151–161.

Sternin, A., Stober, S., Grahn, J. A., and Owen, A. M. (2015). "Tempo estimation from the eeg signal during perception and imagination of music," in *1st International Workshop on Brain-Computer Music Interfacing/11th International Symposium on Computer Music Multidisciplinary Research (BCMI/CMMR'15)* (Plymouth).

Stober, S. (2011). "Adaptive distance measures for exploration and structuring of music collections," in *Proceedings of AES 42nd Conference on Semantic Audio* (Ilmenau), 275–284.

Stober, S. (2017). "Learning discriminative features from electroencephalography recordings by encoding similarity constraints," in *Proceedings of 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)* (New Orleans, LA), 6175–6179. doi: 10.1109/ICASSP.2017.7953343

Stober, S., Cameron, D. J., and Grahn, J. A. (2014). "Using convolutional neural networks to recognize rhythm stimuli from electroencephalography recordings," in *Advances in Neural Information Processing Systems 27*, 1449–1457.

Stober, S., Prätzlich, T., and Müller, M. (2016). "Brain beats: tempo extraction from eeg data," in *17th International Society for Music Information Retrieval Conference (ISMIR'16)*.

Stober, S., Sternin, A., Owen, A. M., and Grahn, J. A. (2015a). Deep feature learning for EEG recordings. *arXiv:1511.04306*.

Stober, S., Sternin, A., Owen, A. M., and Grahn, J. A. (2015b). "Towards music imagery information retrieval: introducing the openmiir dataset of EEG recordings from music perception and imagination," in *16th International Society for Music Information Retrieval Conference (ISMIR'15)*, 763–769.

Stober, S., and Thompson, J. (2012). "Music imagery information retrieval: Bringing the song on your mind back to your ears," in *13th International Conference on Music Information Retrieval (ISMIR'12) - Late-Breaking & Demo Papers*.

Sturm, B. L. (2014). A simple method to determine if a music information retrieval system is a "Horse." *IEEE Trans. Multimedia* 16, 1636–1644. doi: 10.1109/TMM.2014.2330697

van Merriënboer, B., Bahdanau, D., Dumoulin, V., Serdyuk, D., Warde-Farley, D., Chorowski, J., et al. (2015). Blocks and fuel: frameworks for deep learning. *arXiv:1506.00619*.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* 11, 3371–3408.

Vlek, R. J., Schaefer, R. S., Gielen, C. C. A. M., Farquhar, J. D. R., and Desain, P. (2011). Shared mechanisms in perception and imagery of auditory accents. *Clin. Neurophysiol.* 122, 1526–1532. doi: 10.1016/j.clinph.2011.01.042

Wolff, D., and Weyde, T. (2014). Learning music similarity from relative user ratings. *Inform. Retr.* 17, 109–136. doi: 10.1007/s10791-013-9229-0

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 14 July 2017
doi: 10.3389/fpsyg.2017.01179

Music of the 7Ts: Predicting and Decoding Multivoxel fMRI Responses with Acoustic, Schematic, and Categorical Music Features

Michael A. Casey *

Bregman Music and Audio Lab, Computer Science and Music Departments, Dartmouth College, Hanover, NH, United States

Underlying the experience of listening to music are parallel streams of auditory, categorical, and schematic qualia, whose representations and cortical organization remain largely unresolved. We collected high-field (7T) fMRI data in a music listening task, and analyzed the data using multivariate decoding and stimulus-encoding models. Twenty subjects participated in the experiment, which measured BOLD responses evoked by naturalistic listening to twenty-five music clips from five genres. Our first analysis applied machine classification to the multivoxel patterns that were evoked in temporal cortex. Results yielded above-chance levels for both stimulus identification and genre classification–cross-validated by holding out data from multiple of the stimuli during model training and then testing decoding performance on the held-out data. Genre model misclassifications were significantly correlated with those in a corresponding behavioral music categorization task, supporting the hypothesis that geometric properties of multivoxel pattern spaces underlie observed musical behavior. A second analysis employed a spherical searchlight regression analysis which predicted multivoxel pattern responses to music features representing melody and harmony across a large area of cortex. The resulting prediction-accuracy maps yielded significant clusters in the temporal, frontal, parietal, and occipital lobes, as well as in the parahippocampal gyrus and the cerebellum. These maps provide evidence in support of our hypothesis that geometric properties of music cognition are neurally encoded as multivoxel representational spaces. The maps also reveal a cortical topography that differentially encodes categorical and absolute-pitch information in distributed and overlapping networks, with smaller specialized regions that encode tonal music information in relative-pitch representations.

Keywords: multivariate, fMRI, naturalistic, music-informatics, stimulus-encoding, genre, melody, harmony

OPEN ACCESS

Edited by:
Naresh N. Vempala,
Ryerson University, Canada

Reviewed by:
Maria Grazia Di Bono,
University of Padua, Italy
Suzanne T. Witt,
Linköping University, Sweden

***Correspondence:**
Michael A. Casey
mcasey@dartmouth.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *29 October 2016*
Accepted: *28 June 2017*
Published: *14 July 2017*

Citation:
Casey MA (2017) Music of the 7Ts: Predicting and Decoding Multivoxel fMRI Responses with Acoustic, Schematic, and Categorical Music Features. Front. Psychol. 8:1179. doi: 10.3389/fpsyg.2017.01179

1. INTRODUCTION

Humans possess an effortless proclivity to enjoy musical experiences in a wide variety of styles and acoustic configurations. Being moved by, or moving to music requires mental processing that is sensitive to specific auditory and schematic information–the precise features of which, as well as their cortical organization, are yet to be properly understood. Substantial progress has been

made in eliciting the tuning response of groups of voxels to acoustic features in primary auditory areas (Aertsen and Johannesma, 1981; Eggermont et al., 1981; Cariani and Delgutte, 1996; Bendor and Wang, 2005; McDermott and Oxenham, 2008). However, far less is known about responses to categorical and schematic music features–such as genre categories and pitch classes–and about music representations that are encoded outside of primary auditory areas. We address the gap in understanding the fundamental neural codes underlying music cognition by combining methods from three research fields: (i) music cognition, (ii) music information retrieval, and (iii) multivoxel pattern analysis applied to high-field functional magnetic resonance imaging (fMRI).

1.0.1. Multidimensional Representations in Music Cognition and Music Informatics

Results of music cognition research show that multidimensional geometries are implicated in the encoding of musical attributes and in processes of anticipation and reward during music perception. Examples of such geometries include the pitch spiral, torus, and tonnetz models of tonal pitch cognition (Shepard, 1964; Krumhansl, 1990; Tymoczko, 2012), and simplex models of categorical rhythm perception (Honing, 2012). These studies demonstrated that common behavioral responses to music are predicted by models employing statistical learning within multidimensional geometric spaces.

Likewise, music information retrieval systems learn embeddings of musical features in multidimensional spaces, the geometric properties of which are used to successfully predict behavior such as music categorization and musical preferences (Bartsch and Wakefield, 2001; Tzanetakis et al., 2002). Such representations are widely adopted for products and services relating to music consumption (Casey et al., 2008). Hence, a portion of the information in music is inherently geometric, and the properties of such geometries correspond with human behavior.

1.1. Prior Work

1.1.1. Voxel Encoding and Decoding Models

Direct testing of hypotheses about cognitive representations of music and their topographies can be achieved with stimulus-model-based encoding and decoding. Janata et al. (2002) used the geometric pitch-torus model described by Krumhansl (1990), which preserves pitch-distance relationships as perceived by listeners. In their fMRI study, moment-to-moment pitch information of the stimulus–a clarinet melody cycling through all keys–was projected onto a pitch torus using an artificial neural network model (self-organizing map), and the model outputs were used as inputs to a regression model with fMRI voxel responses as the dependent variables. Clusters of significant model predictions were found in pre-frontal cortex, predominantly in rostral and ventral reaches of superior frontal gyrus (SFG). Also utilizing schematic stimulus-model-based encoding, Foster and Zatorre (2010) studied absolute- and relative-pitch representations in a melody-transposition memory task. Their results implicated the intraparietal sulcus (IPS) in comparing two differently transposed melodies.

Expanding the scope of topographical mapping of music features, Alluri et al. (2012) used 25 acoustic features automatically extracted from a single naturalistic musical work–a tango of 8 min duration–to investigate voxel responses to timbral, rhythmic, and tonal features voxel-wise for large cortical and subcortical volumes. Results showed anatomically distinct responses between the three feature groups. Timbral features were implicated in HG, STG, rolandic operculum (ROL), supramarginal gyrus (SMG), superior temporal pole (STP), and the cerebellum; rhythmic and tonal features were found in STG, inferior temporal gyrus (ITG), precuneus, and several subcortical limbic areas–including the left hemispheric amygdala, hippocampus and putamen, mid-cingulate gyrus, supplementary motor area, and the insula. In a further study, they were able to predict voxel responses in bilateral auditory cortex to two music medleys (Alluri et al., 2013), showing significant accuracy of voxel response predictions for auditory, limbic, motor, somatosensory, and frontal areas. In a related work, Toiviainen et al. (2014) demonstrated decoding of acoustic features, predicting the stimulus feature from the voxel response. They found contributions from STG, HG, ROL, and cerebellum contributed to the decoding of timbral features. Bilateral STG, right HG, and hippocampus were significant for rhythmic features. Tonal features, however, were not predicted above chance levels in their study, leaving open the question of whether multivoxel patterns are required to accurately decode neural representations of tonality.

1.1.2. Multivoxel Pattern Analysis

Multivariate pattern analysis (MVPA) treats voxels as the dimensions of continuously-valued feature spaces, such that stimulus-evoked activations are distributed and overlapping between distinct conditions (Haxby et al., 2001; Kriegeskorte et al., 2006; Kriegeskorte, 2011; Stelzer et al., 2013). MVPA models of information representation may recruit the same set of voxels in two or more stimulus conditions with different response levels in each (Haxby et al., 2014).

Applying multivoxel pattern analysis to music, Casey et al. (2012) showed that timbral features based on cepstral coefficients most accurately predicted voxel patterns in primary and secondary auditory areas: Heschl's gyrus (HG), superior temporal gyrus (STG), and superior temporal sulcus (STS). Guntupalli (2013) repeated the experiment of Casey et al. (2012), and additionally performed whole-brain hyperalignment to create between-subject models of stimulus encoding and reconstruction for spectral and timbral acoustic features. Lee et al. (2011) also used voxel-based decoding to classify melodic contour of ascending and descending major and minor scales.

1.2. Hypothesis

Our central hypothesis is that distinct musical attributes are neurally encoded as multivoxel representational spaces. The dimensions of these spaces are individual voxel responses that, when analyzed together in a region, yield properties corresponding to musical behaviors. As such, we would expect machine learning models to statistically infer and generalize the patterns in these encodings, thus yielding accurate decoding of

music information from multivoxel patterns elicited by novel stimuli (decoding models) and accurate predictions of multivoxel patterns for features of novel stimuli (stimulus-model-based encoding).

We also hypothesize that, for naturalistic music listening, multivoxel representational spaces will span the hierarchy of music information from the most general–such as musical style and genre–to the specific–such as melody encoded as relative pitch classes. We further hypothesize that distinct musical features will be differentially encoded across regions where music information is processed, including temporal, pre-frontal, frontal, parietal, occipital, hippocampal, and cerebellar regions, as implied by the prior research outlined above.

We focus our investigation of multivoxel representations on different levels of musical representation: high-level categorical features (5-category music genre), schematic melody features in absolute- and relative-pitch representations, and harmony features encoded by acoustic pitch-class profiles, also called chromagrams. The remainder of the paper proceeds as follows: Section 2 describes the stimuli, experimental procedure, and fMRI data collection and processing; Section 2.4 details the data analysis methods; results are presented in Section 3 followed by discussion of the results and their implication for music cognition in Section 4; and we conclude in Section 5 by outlining directions for our future research.

2. MATERIALS AND METHODS

2.1. Participants

We used the public OpenFMRI dataset published in Hanke et al. (2014) and Hanke et al. (2015). The subject pool consisted of 20 right-handed participants (mean age: 26.6 years, 12 male) who responded to a bulletin calling for volunteers for the study. All participants were native German speakers, and they all reported to have normal hearing without permanent or current temporary impairments and with no known history of neurological disorders. Each participant filled out a questionnaire, detailing basic demographic information, as well as music preference, proficiency and education. As detailed in Hanke et al. (2014) "Participants were fully instructed about the nature of the study, and gave their informed consent for participation in the study as well as for publicly sharing all obtained data in anonymized form. They were paid 100 EUR for their participation. The study was approved by the ethics committee of the Otto-von-Guericke-University of Magdeburg, Germany" (approval reference 37/13).

2.2. Stimuli and Procedure

Stimuli used in this study were identical to those used in three previous studies: Casey et al. (2012), Guntupalli (2013), and Hanke et al. (2015), and are made publicly available in the OpenFMRI *Study Forrest* dataset (Hanke et al., 2014). Twenty five stereo, high-quality naturalistic music stimuli (6 s duration; 44.1 kHz sampling rate) were acquired, with five stimuli in each of five different music genres: (1) Ambient, (2) Country (3) Heavy Metal, (4) RocknRoll, and (5) Symphonic, see **Table 1**. Each stimulus consisted of a six-second excerpt from the middle of a distinct music recording captured from a high-quality Internet streaming service that was seeded by a representative artist for each genre. Clips were manually aligned to the nearest metrical down beat, and they were energy balanced so that the root-mean-square power value was equal across clips. A 50 ms quarter-sine ramp was applied at the start and end of each excerpt to suppress transients. The most prominent differences between the music clips were the presence or absence of vocals and percussion.

Procedures and stimulation setup were as previously reported in Hanke et al. (2014). Participants listened to the audio using custom-built in-ear headphones. After an initial sound calibration, eight scanning runs were performed with each run started by the participant with a key-press ready signal. There were 25 trials, with five different stimuli for each of the five genres per run. Stimulus genre ordering was 2nd-order sequence counter-balanced using De Bruijn cycles. Scanning was continuous, with a delay of 4 s, 6 s, or 8 s between trials. The order of delays was also randomized within each run. Five times per run, once per genre, participants were presented with a question asking for a Yes/No response to a particular feature of the stimulus: e.g., "Was there a female singer?" "Did the song have a happy melody?" The questions were designed to keep subjects' attention on the listening task. Participants were given inter-run breaks, with most resting for under a minute between runs. Stimulus presentation and response logging were implemented using PsychoPy running on a computer with the (Neuro)Debian operating system.

2.2.1. Schematic and Acoustic Features Extraction

In addition to genre labels, the following musical features were extracted from each stimulus: melody schema (absolute pitch), melody schema (relative pitch), and acoustic chromagram features (absolute pitch). The melodies for each of the twenty-five 6-second stimuli were annotated manually by two music undergraduate students and one music graduate student, using the ABC symbolic music standard (Oppenheim, 2010) with discreet pitch-classes aligned to a tempo-invariant metrical grid quantized by 16th-notes. The three sets of annotations were subsequently compared to achieve maximal agreement. These human transcriptions were automatically converted to schematic observation matrices consisting of 12-dimensional absolute pitch-class binary indicator vectors both in the original key (absolute pitch), and transposed to the key of C (relative pitch). Annotations were automatically re-sampled from tempo-normalized 16th-note metrical locations to an absolute time-scale of regular 0.1 s sample intervals, using stimulus tempo information, yielding a 60×12 observation matrix per stimulus. **Figure 1** shows the absolute-pitch melody binary indicator matrix and the corresponding chromagram feature matrix for "Theme from 'Creation"' by Brian Eno, which is the second stimulus in the Ambient category.

Schematic features are invariant to non pitch-class variation in the stimulus, such as loudness fluctuations, timbre, frequency content, articulation, and spatial information. To test whether such variation would be a confounding factor, we also extracted acoustic chromagram features–continuous-valued energies of equal-temperament pitch-class profiles extracted via the Essentia audio MIR toolkit (Bogdanov et al., 2013). Among the

TABLE 1 | List of stimuli used in experiments showing details of music genres, (seed artist), title, artist, and musical key for each clip.

Style/(Seed artist)	Title	Artist	Key (Clip)
Ambient			
(Brian Eno)	A Clearing	Brian Eno	F
	Theme from "Creation"	Brian Eno	C
	Old Land	Eno Moebius Roedelius	C
	Horizons Lointains	Galerie Stratique	Cm
	IO - Moon of Jupiter	Anugarma	B
Country			
(Waylon Jennings)	Are You Sure...?	Waylon Jennings	C
	Me and Paul	Willie Nelson	A
	Pancho and Lefty	Merle Haggard	D
	Whiskey Bent and Hell Bound	Hank Williams Jr.	G
	Welfare Line	Willie Nelson	D
Heavy Metal			
(Ozzy Osbourne)	Fire in the Sky	Ozzy Osbourne	D♭
	You've Got Another Thing Coming	Judas Priest	F♯
	Of Wolf & Man	Metallica	E
	You Shook Me All Night Long	AC-DC	G
	Rock You Like A Hurricane	Scorpions	Em
Rock & Roll			
(Eddie Cochran)	Jailhouse Rock	Elvis Presley	E♭
	Shake Rattle and Roll	Bill Haley	F
	Bama Lama Bama Loo	Little Richard	F
	Come On Let's Go	Ritchie Valens	A
	Money Honey	Eddie Cochran	E
Symphonic			
(Beethoven)	Symphony No. 9 Mvt. 2	Beethoven	F
	Symphony No. 4 Mvt. 4	Tchaikovsky	B♭m
	Symphony No. 2 Mvt. 4	Sibelius	D
	Symphony No. 5 Mvt. 1	Schubert	F
	Symphony No. 6 Mvt. 1	Beethoven	F

All clips were 6 s duration, acquired from 44.1 kHz stereo 192 kbps streams.

numerous differences between acoustic chromagrams and binary-chord schema are the presence of continuous energy values, amplitude modulation (due to loudness and dynamics), spectral envelope modulation (due to timbre), energy (from melody and bass notes and their harmonics), mis-aligned frequency channels (tuning), harmonic energy, room acoustics, and additive noise–to enumerate just a few differences. All features, schematic and acoustic, were further processed by singular-value decomposition, preserving the coefficients that explained at least 95% of the feature variance across the training stimulus set. As with the EPI features, feature matrices were flattened into vectors by stacking the 60 observation vectors (60×0.1 s samples) for each stimulus, thereby preserving their temporal sequence information, prior to subsequent data analysis.

2.3. fMRI Data Acquisition and Pre-processing

The high-resolution 7-Tesla fMRI data was previously released via the *OpenFMRI* initiative (Hanke et al., 2015); the stimuli and experiment design used in the music perception phase of the data release (scanning session III) reproduce the original 3T experiment of Casey et al. (2012). To our knowledge, the current study is the first feature-based analysis of the music representational spaces revealed by the published high-resolution data set.

Functional MRI data was recorded during auditory stimulation. Anatomical T1-weighted scans were performed at 3 Tesla, and T2*-weighted functional scans were performed at 7 Tesla for slabs with partial brain coverage (MNI152 $z \approx -30\,\text{mm}\ldots 40\,\text{mm}$). Subjects were given the cognitive task of listening attentively to the twenty five music clips in five genres, as shown in **Table 1**, and answering a dual-choice question, e.g., "did the clip have a happy melody?" Subjects responded "yes" or "no" to these questions via a response button box. These questions helped to ensure that subjects attended to the music across trials. Data from these catch trials were discarded from the analyses. The process was repeated eight times for each participant, using a unique quasi-randomized second-order balanced stimulus sequence for each subject and for each of the eight acquisition runs. Data consisted of 153 volumes per run, with a repetition time (TR) of 2.0 s each volume. Following is a summary of details of scanning, motion correction, and distortion processing as described in Hanke et al. (2014). T2*-weighted echo-planar images were acquired during stimulation using a 7-Tesla Siemens MAGNETOM magnetic resonance scanner. Thirty six axial slices (thickness 1.4 mm, 1.4×1.4 mm in-plane resolution, 224 mm field-of-view (FoV), anterior-to-posterior phase encoding direction) with a 10% inter-slice gap were recorded in ascending order. This configuration was chosen to achieve a balance between spatial resolution, volume coverage and volume acquisition time. Slices were oriented to include the ventral portions of frontal and occipital cortex while minimizing intersection with the eyeballs. The field-of-view was centered on the approximate location of Heschl's gyrus. Head-movement correction utilized reference scans at the start of the recording session and was performed on-line within the scanner in conjunction with a high-field distortion correction procedure.

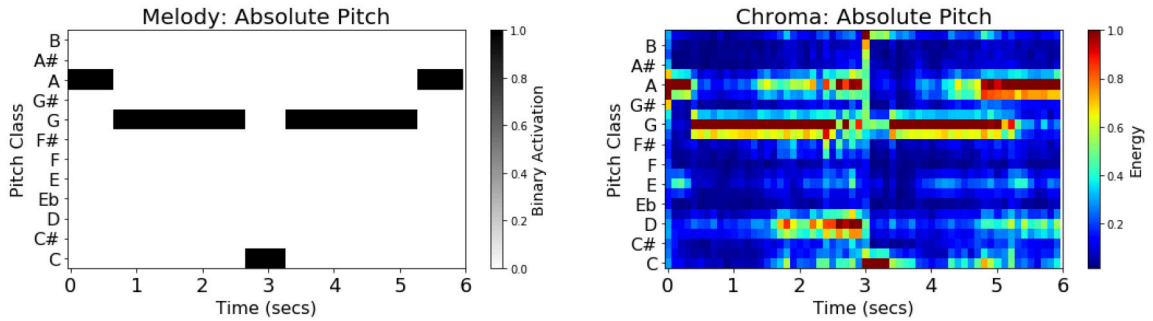


FIGURE 1 | (Left) Schematic melody features (absolute pitch) for stimulus Ambient002. The binary-valued indicator matrices were obtained by three independent human expert transcriptions, followed by machine encoding to absolute time-scale and relative-pitch representation. **(Right)** Audio chromagram features (absolute pitch) for the same stimulus automatically extracted using the *Essentia* audio feature extraction toolkit in Python. Visible in the diagram is the trait that chromagram features are polyphonic, encoding all pitches present in the music clip, such as those corresponding to bass and chords, in addition to the melody.

EPI images were co-registered to a common group template using FSL's FLIRT, MCFLIRT, and FNIRT software. A group-specific template volume for EPI images was derived in order to aid anatomical alignment across brains. Subject's functional images were aligned to their respective reference-scan images, acquired at the start of the session, via a rigid body transformation using MCFLIRT. Each subject's reference-aligned images were averaged to create a template image for each brain. Subsequently, all subjects' template images were aligned by means of an affine transformation using FLIRT. The affine transformation was determined using the subject's brain with the least root mean square difference to the average image across all brains prior to alignment. The resulting average template volume was masked to produce the maximal intersection of individual brains to create the group EPI template volume (Hanke et al., 2014).

EPI data were then projected to voxel features using a per-voxel General Linear Model (GLM) for each stimulus in each run. The GLM was fitted for the EPI voxel time series in each run using the PyMVPA software framework (Hanke et al., 2009). The model fitting algorithm used the event-related design matrix (e.g., 3×2 s TRs per 6-s stimulus condition) with a double-gamma hemodynamic response function (HRF) regressor.

2.4. Analysis

2.4.1. Analysis 1: Mulivoxel Classification by Song and by Music Genre

Within-subject classifiers were trained on two tasks: song (stimulus) classification and genre (category) classification. After feature selection using a held-out portion of the dataset, song classifiers were cross-validated by run, and genre classifiers were cross-validated by stimulus–with category balancing achieved by holding out all runs of one stimulus from each of the five categories per cross-validation fold. We used linear-kernel support vector machines (SVM) with margin-parameter, C, scaled according to the norm of the data.

2.4.1.1. Region of interest specification

Three bilateral regions in temporal cortex were selected from the Harvard-Oxford Cortical Structural Atlas, using FSLVIEW's Atlas Tools, and then warped to each subject's brain via the common group template. Regions of interest (ROIs) were selected spanning primary and secondary auditory cortex due to their implication in prior music classification studies (Casey et al., 2012; Guntupalli, 2013); these were: Heschl's gyrus (HG), anterior superior temporal gyrus (aSTG), and posterior superior temporal gyrus (pSTG).

To reduce the impact of noisy voxels on classifier performance, sensitivity-based feature selection retained only the top 5,000 voxels in each ROI. One-way analysis of variance (ANOVA), with individual stimulus factors, was applied followed by sensitivity-based feature selection, keeping only 5,000 voxels with the highest F-scores. To address possible circularity bias between feature selection and model training and testing, e.g., see Kriegeskorte et al. (2009), runs 1 and 4 were held out for feature selection and the remaining six runs were used for model training and cross-validation. Z-score mapping of the fMRI data was folded into the cross validation. Analysis scripts were implemented in Python 2.7.12 using the Anaconda distribution and the PyMVPA 2.6.0 framework (Hanke et al., 2009).

2.4.2. Analysis 2: Stimulus-Encoding Model Searchlight

The anatomical distribution of cognitive music representations was analyzed using a searchlight algorithm (Kriegeskorte et al., 2006; Haxby et al., 2014). This procedure yielded an anatomical map of stimulus-model-based prediction accuracies in spherical subsets ("searchlights") centered on every voxel; the map value for each voxel thus derives from the information present in each searchlight volume, and not each voxel individually. Stimulus encoding models were trained and tested for each of $\approx 6,250$ searchlight volumes–varied by subject anatomy–over a large volume of cortex–all Harvard-Oxford Cortical Atlas regions within the field of view, including pre-frontal, frontal, parietal, occipital, para-hippocampal, and cerebellar regions–using ridge

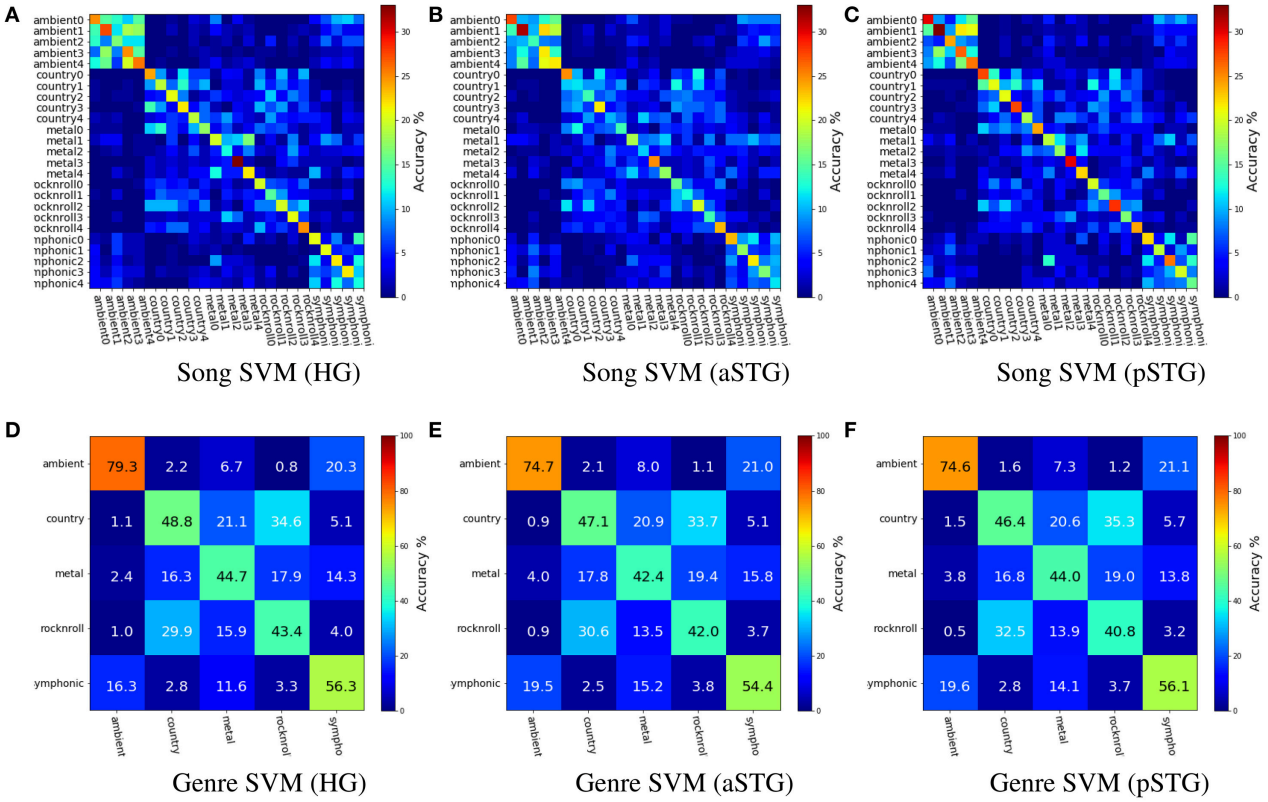


FIGURE 2 | Group-level mean accuracies for song (upper) and genre (lower) SVM classifiers for voxels in Heschl's gyrus (HG), anterior superior temporal gyrus (aSTG), and posterior superior temporal gyrus (pSTG). The figures show that distinct anatomical ROIs yield similar relational information about music. The patterns of misclassifications show that when songs are misclassified, they are more likely to be confused with items from the same genre, or a similar sounding genre: e.g., Ambient and Symphonic; and Rock-n-Roll and Country. **(A)** Song SVM (HG). **(B)** Song SVM (aSTG). **(C)** Song SVM (pSTG). **(D)** Genre SVM (HG). **(E)** Genre SVM (aSTG). **(F)** Genre SVM (pSTG).

regression with music stimulus features as input variables and voxel pattern responses as the dependent variables. The models were used to predict the voxel pattern response vector for new stimuli on the basis of their extracted musical features. Ridge regression was chosen due to its use of Tikhonov regularization, to counter possible deleterious effects of overdetermined models and other numerical instabilities.

A sphere-radius of 3 voxels was used and the accuracy of the predictions was defined as the correlation-error probability $(1 - p)$ between model predictions and voxels in each searchlight volume. The correlation-error probability yielded a measure in the range $[0 \ldots 1]$, with perfect predictions scoring 1. The searchlight creates ROIs by exhaustive subset selection, therefore we did not need to hold runs out for feature selection as we did in Analysis 1. For testing on novel data, balanced cross-validation held out all 8 runs of a randomly-selected stimulus in each of the five genre categories. Cross-validation was repeated 10 times in each searchlight, yielding 5-stimuli $\times$ 8 runs $\times$ 10 repetitions $=$ 400 tests per searchlight, which were averaged to give a single correlation-error probability score per searchlight. Due to the large computational demand of searchlight analysis, we used randomized scattering by 3 voxels, and averaged results over the multiple cross-validation folds, which sped-up the computation by a factor of 27 relative to a searchlight sphere spacing of 1 voxel. The searchlight analysis, and permutation computations for bootstrapping the null distribution, took approximately 15,000 h of CPU time using scattering, so the speed-up factor was critical to the computational feasibility of the results. The searchlight with radius 3 voxels yielded spheres containing a maximum of 123 voxels for each center location. Following the methods of Stelzer et al. (2013), group-level statistical evaluation of the searchlight analysis was implemented using 100,000 bootstrap samples drawn pair-wise by subjects from 100 randomized-target null models in each searchlight, and then estimating a voxel-wise threshold with probability $p < 0.001$ with respect to the bootstrap null distribution.

3. RESULTS

3.1. Analysis 1

Figure 2 shows the group-averaged cross-validated results of within-subject SVM classification for the three bilateral temporal-region ROIs used for Analysis 1. Song classification results, with balanced cross-validation by run (Chance $=$ 4%), were: HG (Mean $=$ 21.1%, SE $=$ 0.9%), aSTG (Mean $=$ 18.5%, SE $=$ 1.1%), and pSTG (Mean $=$ 23.2%, SE $=$ 1.0%). Results for 5-way genre classification, with balanced cross-validation by stimulus (Chance $=$ 20%), were: HG (Mean $=$ 54.5%, SE $=$ 5.9%),

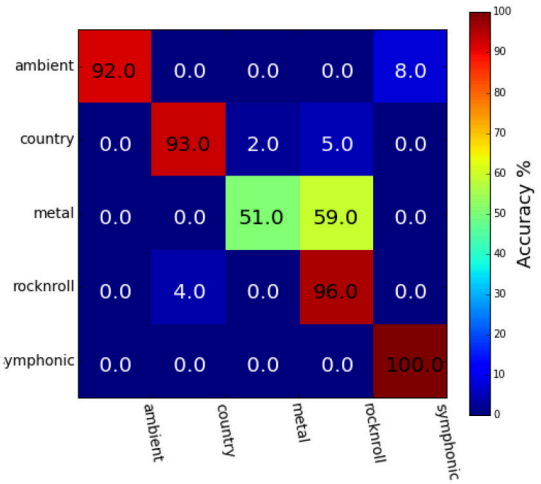


FIGURE 3 | Group-averaged behavioral genre category-assignment confusions ($n = 20$; Mean = 86.4%, $SE = 8.0\%$, Chance = 20%). Category confusions overlap those shown in **Figure 2**, with predominant confusions occurring between the Ambient and Symphonic genres, and also between the Country, Heavy Metal, and Rock&Roll genres.

aSTG (Mean = 52.1%, SE = 5.4%), and pSTG (Mean = 52.4%, SE = 5.5%).

Figure 3 shows the results of behavioral genre categorization ($n = 20$) for the 25 stimuli used in the genre classification task. Accuracies in the behavioral task (Mean = 86.4%, SE = 8.0%, Chance = 20%) were higher than the SVM classifier reported above. The Spearman rank-order correlation scores, r, between the group-averaged confusion matrix of the behavioral task and the group-averaged confusion matrix for the genre classifier for each ROI were HG ($r = 0.76, p < 0.01$), aSTG ($r = 0.79, p < 0.01$), and pSTG ($r = 0.79, p < 0.01$). The spearman rank-order correlation was calculated using the values above the main diagonal of the confusion matrices only, so as to remove positive correlation bias due to the diagonal structure of confusion matrices (Guntupalli, 2013).

3.2. Analysis 2

Figure 4 and **Table 2** show MNI-space group-level FWE-corrected clusters ($p < 0.05$) based on stimulus-model-based encoding prediction accuracies (correlation-error probabilities). Significant clusters were identified for all three feature representations–melody relative pitch, melody absolute pitch, and acoustic chromagram features–in multiple sites spanning the searchlight regions of interest (ROIs). Acoustic chromagram features yielded the greatest number of significant clusters, 97 (43 left, 47 right, 7 both hemispheres), followed by absolute-pitch melody features (15 left, 12 right, 2 both), then relative-pitch melody features (1 left, 1 right, 1 both). Significant chromagram (Chrom) feature clusters occupied a total volume 10,595 voxels, spanning sites in most of the bilateral searchlight ROI volume: namely, temporal primary and secondary auditory cortex (A1, A2)–including Heschl's gyrus (HG), planum temporale (PT), superior temporal gyrus (STG), supramarginal gyrus (SMG), middle temporal gyrus (MTG) all lateralized marginally to the right hemisphere; Rolandic operculum (ROL); inferior frontal gyrus (IFG); temporal, frontal, and occipital poles (TP, FP, OP); middle frontal gyrus (MFG)/Broca's area; frontal orbital cortex (FO); intracalcarine cortex (CAL); insular cortex (IC); lingual gyrus (LING); parahippocampal gyrus (PHG); cerebellum; and multiple visual areas (V1, V2, V3, V4).

Clusters due to absolute-pitch melody features occupied a total volume of 3,276 voxels and were concentrated in temporal and frontal areas largely overlapping those of chromagram features, but with fewer and smaller significant clusters. Notable differences in the distribution of clusters compared with chromagram features were the inclusion of clusters in the putamen; a greater presence of clusters in right MTG and STG; and left-lateralized clusters in multiple visual areas (V1, V2, V3, V4). Finally, relative-pitch melody features exhibited clusters that occupied a total volume of 317 voxels which were lateralized and concentrated in three clusters: the junction of the right cerebellum (c-VI) and temporal-occipital fusiform gyrus (FFG), left planum polare (PP), and right PT (A2). We observed overlapping representations of all three feature representations in the left PP. Outside of this area, relative-pitch and absolute-pitch melody features had no further overlapping clusters. Chromagram and relative-pitch melody clusters overlapped in the right cerebellum and in the right PT extending through the parietal operculum (PO) area of A1. Chromagram and absolute-pitch clusters overlapped in numerous sites that were mostly lateralized to the right: pMTG, HG, PT, FFG, CAL, FO, SMG, FP, IFG.

4. DISCUSSION

4.1. Analysis 1

The within-subject song classification results show significantly higher accuracies than previously reported in Guntupalli (2013) (Mean = 15.95%, SE = 1.62%, Chance = 4%) for the same stimuli using different subjects with different (3T) fMRI data. One reason for the greater accuracies in the current study may be the use of high-field (7T) fMRI data, which doubles the spatial resolution of voxels in each dimension thus affording greater detail for pattern discrimination. Differences in voxel selection strategies are enumerated below.

The within-subject genre classification accuracies are slightly lower than those reported in Casey et al. (2012) (Mean = 60.0%) for the same stimuli, but with more stringent cross validation in the current study, and ≈25% lower than those reported in Guntupalli (2013) for the same stimuli. Apart from the use of high-field fMRI in the current study, differences between the current and the two former studies include 5,000-voxel feature selection by ROI in the current study, no sensitivity based selection in Casey et al. (2012), and 1,000-voxel feature-selection from whole brain voxels in Guntupalli (2013). The latter study also employed a different cross-validation scheme, which also accounts for some of the difference in accuracy. In the case of genre classification, selection of voxels from the whole brain 3T data yielded greater classifier accuracies

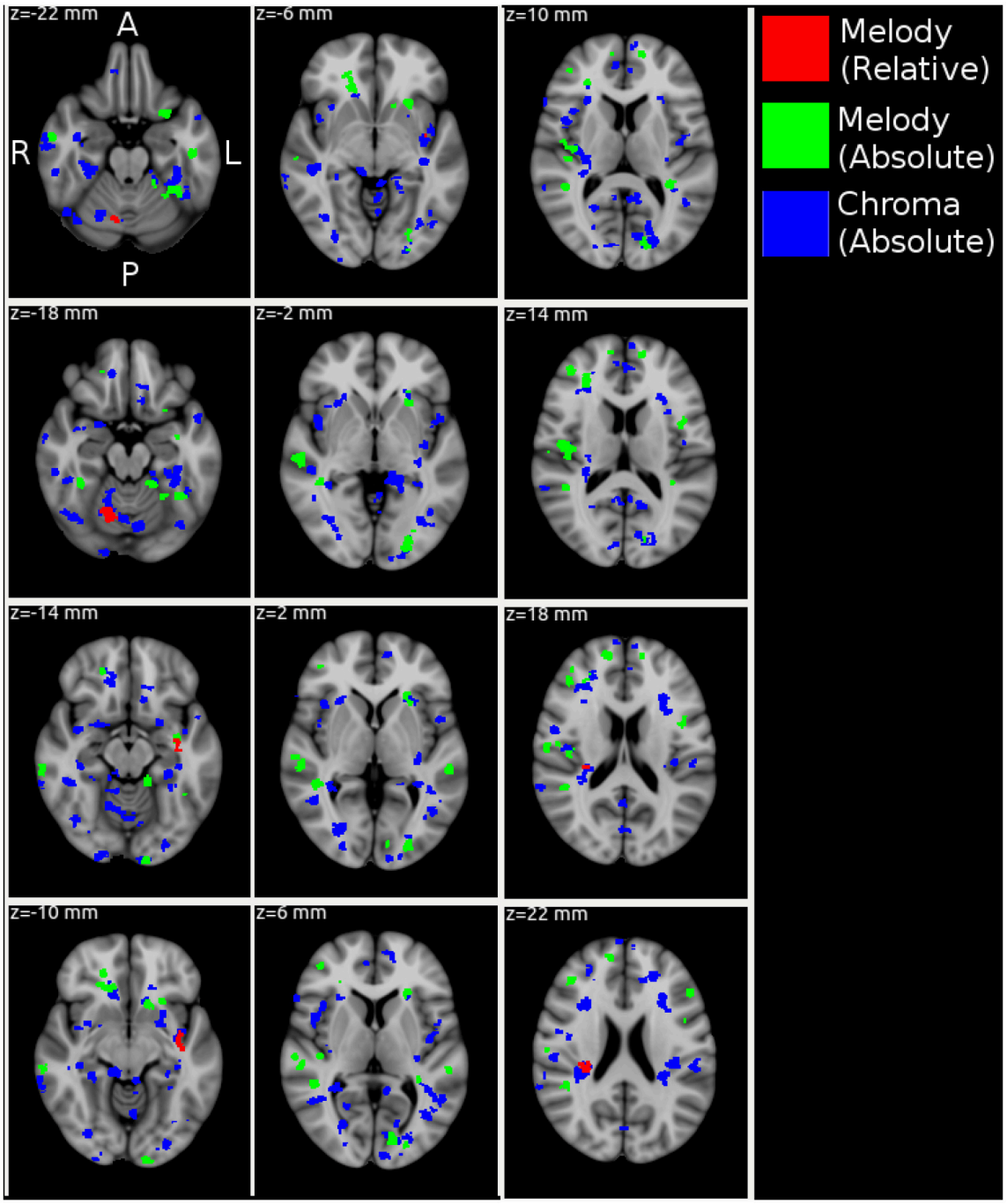


FIGURE 4 | MNI-space group-level FWE-corrected clusters ($p < 0.05$) organized in 4 mm-spaced axial columns. In this multivariate analysis, the map value for each voxel derives from the information present in a 3-voxel-radius searchlight volume (max 123 voxels) and not each voxel individually. Acoustic chromagram features yielded the greatest number of significant clusters, 97 (43 left, 47 right, 7 both hemispheres), followed by absolute-pitch melody features (15 left, 12 right, 2 both), then relative-pitch melody features (1 left, 1 right, 1 both).

than restricting voxel selection to temporal cortex with 7T data. Overall, these results show that distinct anatomical ROIs yield similar pattern-space information about song identity and genre, thus they hierarchically encode multiple levels of music information.

The high correlation score between behavioral and classifier confusion matrices is due to both exhibiting the same pattern of confusions between Ambient and Symphonic categories, and between Country and Rock&Roll categories. The most prominent difference between these two groups of confusions

TABLE 2 | Average group results: searchlight-based (sphere radius = 3 voxels, max. size 123 voxels) cross-validated within-subject stimulus encoding ($n = 20$; ridge regression).

					Center of mass (MNI)				
#	**Voxels**	**Max**	**Mean**	**Std**	***X***	***Y***	***Z***	$p_{clus.}$	**Structure**
Melody (Relative)									
1	122	0.84	0.67	0.21	73.8	57.0	54.9	0.0110	Occipital Fusiform Gyrus
2	100	0.85	0.76	0.15	130.0	111.1	62.1	0.0174	Planum Polare
3	95	0.83	0.66	0.21	58.4	97.3	90.7	0.0174	Planum Temporale
Melody (Absolute)									
1	282	0.84	0.69	0.20	119.8	80.4	62.8	0.0001	Temporal Occip. Fusiform
2	281	0.86	0.78	0.13	58.5	95.9	86.9	0.0001	Planum Temporale
3	230	0.86	0.84	0.01	49.9	145.1	73.2	0.0001	Frontal Operculum Cortex
4	209	0.84	0.63	0.25	104.7	54.2	76.1	0.0001	Intracalcarine Cortex
5	206	0.83	0.69	0.21	68.7	53.5	75.7	0.0001	Intracalcarine Cortex
6	164	0.85	0.78	0.14	121.8	147.2	87.9	0.0004	Frontal Operculum Cortex
7	161	0.83	0.76	0.14	75.1	66.7	59.4	0.0004	Lingual Gyrus
8	136	0.84	0.69	0.21	54.4	144.0	95.2	0.0013	Inferior Frontal Gyrus
9	125	0.86	0.78	0.15	99.4	85.9	68.7	0.0019	Parahippocampal Gyrus
10	112	0.85	0.64	0.26	60.1	90.3	52.5	0.0036	Temporal Fusiform Cortex
11	101	0.82	0.65	0.23	147.1	90.5	99.8	0.0064	Parietal Operculum Cortex
12	90	0.86	0.77	0.18	52.8	111.1	80.7	0.0117	Insular Cortex
13	85	0.86	0.79	0.13	103.2	88.8	59.5	0.0149	Parahippocampal Gyrus
14	80	0.87	0.86	0.00	139.9	154.4	94.1	0.0191	Inferior Frontal Gyrus
15	78	0.85	0.81	0.07	134.7	131.8	93.0	0.0204	Precentral Gyrus
Chroma (Absolute)									
1	503	0.83	0.70	0.19	119.8	80.4	62.8	0.0001	Temporal Occip. Fusiform
2	422	0.83	0.68	0.22	58.5	95.9	86.9	0.0001	Planum Temporale
3	373	0.85	0.74	0.18	49.9	145.1	73.2	0.0001	Frontal Operculum Cortex
4	304	0.84	0.66	0.22	104.7	54.2	76.1	0.0001	Intracalcarine Cortex
5	280	0.84	0.67	0.23	68.7	53.5	75.7	0.0001	Intracalcarine Cortex
6	276	0.83	0.64	0.24	121.8	147.2	87.9	0.0001	Frontal Operculum Cortex
7	273	0.84	0.66	0.22	75.1	66.7	59.4	0.0001	Lingual Gyrus
8	270	0.83	0.69	0.18	54.4	144.0	95.2	0.0001	Inferior Frontal Gyrus
9	257	0.86	0.77	0.14	99.4	85.9	68.7	0.0001	Parahippocampal Gyrus
10	182	0.82	0.66	0.20	60.1	90.3	52.5	0.0002	Temporal Fusiform Cortex
11	181	0.85	0.67	0.23	147.1	90.5	99.8	0.0002	Parietal Operculum Cortex
12	167	0.84	0.56	0.28	52.8	111.1	80.7	0.0003	Insular Cortex
13	159	0.83	0.70	0.17	48.6	120.0	50.4	0.0003	No label found!
14	159	0.85	0.69	0.18	139.9	75.1	78.8	0.0003	Middle Temporal Gyrus
15	157	0.83	0.72	0.17	72.1	160.3	55.6	0.0003	Frontal Orbital Cortex

The table lists statistics (size, max/mean/std accuracy) as well as localization information (coordinates in mm MNI152) for clusters with above-chance classification performance in the group (FWE-corrected, cluster-level probability $p < 0.05$).

is that the confusion between Rock'n Roll and Heavy Metal in the behavioral task is much greater than it is with the classifier. These classification results show that songs that are misclassified at either the song level or the genre level are more likely to be confused with items from the same genre, or with items from a *similar-sounding* genre: e.g., Ambient and Symphonic, and Rock-n-Roll and Country. The latter implies that there is a super-ordinate category above the level of genre, one possibility for which is the presence or absence of vocals and/or percussion.

4.2. Analysis 2

In the realm of schematic feature representations, Janata et al. (2002) showed how dynamic attributes of tonal music, namely key changes, can be mapped onto a consistent cortical topography in prefrontal areas. Furthermore, they showed that the "tonality surface" representation was invariant to changes in the starting reference key, when the study was repeated with the same subjects over multiple scanning sessions. Hence, they demonstrated a direct cognitive representation of relative pitch

encoding. In our work, we also found group-level representations of relative pitch, but for melodic encoding, rather than the slowly varying key surface of the previous work. Foster and Zatorre (2010) implicated IPS in the manipulation of auditory representations, such as used in a melodic transposition memory task. Whilst we found no significant clusters in the vicinity of IPS for relative-pitch melody features, we surmise that the naturalistic listening condition of the current study–i.e., attentive listening without an explicit memory task–elicited a differing view of voxel response patterns to relative pitch encoding of melodies than did the earlier work. Our relative-pitch results do however overlap with Janata et al. (2002) who also found in their tonality study with key that relative-pitch representations were present in the cerebellum and hippocampus, as well as in pre-frontal areas, both of which are present in our results.

Alluri et al. (2013) used an aggregate stimulus encoding model to perform voxel-wise response predictions to novel stimuli. Since the features were aggregated, they were not able to map responses to individual musical attributes. However, their aggregate model prediction results anatomically overlap with the current study, in that they found significant model-prediction accuracies in primary and secondary auditory areas (STG, HG, MTG), as well as pre-frontal and frontal areas (SFG), Rolandic operculum, putamen, and insula. In their earlier work, Casey et al. (2012) demonstrated stimulus-encoding-model-based decoding for low-level audio features corresponding to chromagram, spectral, and cepstral audio features. Chromagram features performed significantly above chance level in predicting the brain response for superior-temporal regions. In the current study, we have found wide activation of acoustic chromagram features across the all cortical and subcortical ROIs of the searchlight analysis. However, we note that the acoustic feature has folded within it the acoustic confounds described in Section 2, so components of the chromagram feature for acoustic mixtures, as in naturalistic music stimuli, may elicit sensitivities across many ROIs because the feature encodes substantial additional information beyond the intended representation of polyphonic pitch content of the stimulus.

5. CONCLUSION AND FUTURE DIRECTIONS

We have demonstrated parallel, distributed, and overlapping representations of musical features using machine learning models, high-field fMRI, and naturalistic music stimuli. The results from Analysis 1 show that decoding models can identify songs significantly above chance levels by their voxel pattern responses for held-out runs, and that categorical models accurately decode music genre categories for voxel pattern responses to novel stimuli in five genres. Furthermore, the pattern of confusions exhibited by the classifiers was significantly correlated with confusions in a behavioral categorization task. These results support our hypothesis that music cognition is neurally represented by multivoxel pattern spaces whose geometric properties, such as distance between response vectors, underlie observed human musical behavior.

Results from Analysis 2 demonstrate that stimulus-model-based-encoding accurately predicts voxel responses to music stimuli yielding significant clusters in multiple sites across the cortical volume. As we expected to see, distinct musical features are differentially encoded in distributed and anatomically overlapping sites. The current study extends prior work in stimulus-model-based encoding of music representational spaces by providing maps, not only of audio-based feature encoding, but also of schematic music features. Mapping parallel features of the information content in music content reveals wide networks of overlapping representational spaces for music. Future work will explore how well different pitch and rhythm representational space hypotheses, such as the tonnetz and the simplex models, can predict multivoxel responses in areas known to be implicated in the processing of these musical attributes, which will allow us to select the most likely neurally encoded representation among competing representational hypotheses for specific musical attributes.

We note, however, that care must be taken when extracting acoustic features to avoid confounding within the feature multiple unintended attributes of the stimulus along with the intended musical attribute, as we observed with the chromagram feature. This highlights a potentially important advantage of symbolic music features for mapping music cognition, and it also throws into question the utility of mixed low-level audio features for mapping music representations across cortical volumes. Audio source separation methods, which are the subject of much current music informatics research, may proove useful for increasing the representaitonal specificity of automatic acoustic feature extraction.

Our future work will include regrouping our analyses, separating results by genre, to test the hypothesis that music is cortically organized by high-level categories–such as genre, emotion, and semantic categories–with lower-level schematic and acoustic features repeatedly embedded within these superordinate representational spaces. The current study modeled stimulus-synchronous imaging. A further refinement to our work would be to introduce of models for predictive stimulus encoding, in which features of current and past time steps predict future voxel responses. Such models would be necessary to illuminate the neural representation of prediction-driven mechanisms that are widely understood to be implicated in the anticipation and reward mechanisms of musical enjoyment.

AUTHOR CONTRIBUTIONS

MC designed the experiments, prepared the stimuli, prepared data and software for analysis of the 3T and 7T images, analyzed the data to produce the results in this paper, and wrote the paper.

FUNDING

Funding for this research was provided by the Neukom Institute for Computational Science and the Dean of Faculty at Dartmouth College. Open access article fees were provided by the Dean of the Faculty at Dartmouth College.

ACKNOWLEDGMENTS

The following people at Dartmouth College provided assistance. Professor Thalia Wheatley helped with the genre experiment design and performed the original 3T scans with Olivia Kang. Kristen Colwell, Jianyu Fan, and Deborah Johnson assisted with preparation of stimuli and features. Alison Mattek conducted the behavioral genre-categorization experiment. Jefferey Mentch performed the anatomical parcellations using FSLVIEW. John Hudson assisted with preparing the Dartmouth Discovery cluster for data analysis.

REFERENCES

Aertsen, A., and Johannesma, P. (1981). The spectro-temporal receptive field. *Biol. Cybern.* 42, 133–143. doi: 10.1007/BF00336731

Alluri, V., Toiviainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., and Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage* 59, 3677–3689. doi: 10.1016/j.neuroimage.2011.11.019

Alluri, V., Toiviainen, P., Lund, T. E., Wallentin, M., Vuust, P., Nandi, A. K., et al. (2013). From vivaldi to beatles and back: predicting lateralized brain responses to music. *Neuroimage* 83, 627–636. doi: 10.1016/j.neuroimage.2013.06.064

Bartsch, M. A., and Wakefield, G. H. (2001). "To catch a chorus: using chroma-based representations for audio thumbnailing," in *2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY: IEEE), 15–18. doi: 10.1109/ASPAA.2001.969531

Bendor, D., and Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex. *Nature* 436, 1161–1165. doi: 10.1038/nature03867

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., et al. (2013). "Essentia: an audio analysis library for music information retrieval," in *Proceedings ISMIR (Citeseer)* (Curitiba), 493–498.

Cariani, P. A., and Delgutte, B. (1996). Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. *J. Neurophysiol.* 76, 1698–1716.

Casey, M., Thompson, J., Kang, O., Raizada, R., and Wheatley, T. (2012). "Population codes representing musical timbre for high-level fmri categorization of music genres," in *Machine Learning and Interpretation in Neuroimaging* (Sierra Nevada: Springer), 34–41.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. (2008). Content-based music information retrieval: current directions and future challenges. *Proc. IEEE* 96, 668–696. doi: 10.1109/JPROC.2008.916370

Eggermont, J., Aertsen, A., Hermes, D., and Johannesma, P. (1981). Spectro-temporal characterization of auditory neurons: redundant or necessary? *Hear. Res.* 5, 109–121. doi: 10.1016/0378-5955(81)90030-7

Foster, N. E., and Zatorre, R. J. (2010). A role for the intraparietal sulcus in transforming musical pitch information. *Cereb. Cortex* 20, 1350–1359. doi: 10.1093/cercor/bhp199

Guntupalli, J. S. (2013). *Whole Brain Hyperalignment: Intersubject Hyperalignment of Local Representational Spaces*. Dartmouth College, Ph.D. Thesis.

Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., et al. (2014). A high-resolution 7-tesla fMRI dataset from complex natural stimulation with an audio movie. *Sci. Data* 1:140003. doi: 10.1038/sdata.2014.3

Hanke, M., Dinga, R., Häusler, C., Guntupalli, J. S., Casey, M., Kaule, F. R., et al. (2015). High-resolution 7-tesla fmri data on the perception of musical genres–an extension to the studyforrest dataset. *F1000Research* 4:174. doi: 10.12688/f1000research.6679.1

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., and Pollmann, S. (2009). Pymvpa: a python toolbox for multivariate pattern analysis of fmri data. *Neuroinformatics* 7, 37–53. doi: 10.1007/s12021-008-9041-y

Haxby, J. V., Connolly, A. C., and Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Ann. Rev. Neurosci.* 37, 435–456. doi: 10.1146/annurev-neuro-062012-170325

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430. doi: 10.1126/science.1063736

Honing, H. (2012). Without it no music: beat induction as a fundamental musical trait. *Ann. New York Acad. Sci.* 1252, 85–91. doi: 10.1111/j.1749-6632.2011.06402.x

Janata, P., Birk, J. L., Van Horn, J. D., Leman, M., Tillmann, B., and Bharucha, J. J. (2002). The cortical topography of tonal structures underlying western music. *Science* 298, 2167–2170. doi: 10.1126/science.1076262

Kriegeskorte, N. (2011). Pattern-information analysis: from stimulus decoding to computational-model testing. *Neuroimage* 56, 411–421. doi: 10.1016/j.neuroimage.2011.01.061

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. *Proc. Natl. Acad. Sci. U.S.A.* 103, 3863–3868. doi: 10.1073/pnas.0600244103

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* 12, 535–540. doi: 10.1038/nn.2303

Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch*. New York, NY: Oxford University Press.

Lee, Y.-S., Janata, P., Frost, C., Hanke, M., and Granger, R. (2011). Investigation of melodic contour processing in the brain using multivariate pattern-based fMRI. *NeuroImage* 57, 293–300. doi: 10.1016/j.neuroimage.2011.02.006

McDermott, J. H., and Oxenham, A. J. (2008). Music perception, pitch, and the auditory system. *Curr. Opin. Neurobiol.* 18, 452–463. doi: 10.1016/j.conb.2008.09.005

Oppenheim, I. (2010). *The abc Music Standard 2.0 (December 2010), 2010*. Available online at: http://abcnotation.com/wiki/abc:standard:v2.0 (Accessed 27 March, 2017).

Shepard, R. N. (1964). Circularity in judgments of relative pitch. *J. Acoust. Soc. Amer.* 36, 2346–2353. doi: 10.1121/1.1919362

Stelzer, J., Chen, Y., and Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (mvpa): random permutations and cluster size control. *Neuroimage* 65, 69–82. doi: 10.1016/j.neuroimage.2012.09.063

Toiviainen, P., Alluri, V., Brattico, E., Wallentin, M., and Vuust, P. (2014). Capturing the musical brain with lasso: dynamic decoding of musical features from fmri data. *Neuroimage* 88, 170–180. doi: 10.1016/j.neuroimage.2013.11.017

Tymoczko, D. (2012). The generalized tonnetz. *J. Music Theory* 56, 1–52. doi: 10.1215/00222909-1546958

Tzanetakis, G., Member, S., and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process.* 10, 293–302. doi: 10.1109/TSA.2002.800560

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 25 April 2017
doi: 10.3389/fpsyg.2017.00621

Predicting Variation of Folk Songs: A Corpus Analysis Study on the Memorability of Melodies

Berit Janssen[1,2]*, *John A. Burgoyne*[2] *and Henkjan Honing*[2]

[1] Meertens Institute, Royal Netherlands Academy of Arts and Sciences, Amsterdam, Netherlands, [2] Music Cognition Group, Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, Netherlands

We present a hypothesis-driven study on the variation of melody phrases in a collection of Dutch folk songs. We investigate the variation of phrases within the folk songs through a pattern matching method which detects occurrences of these phrases within folk song variants, and ask the question: do the phrases which show less variation have different properties than those which do? We hypothesize that theories on melody recall may predict variation, and as such, investigate phrase length, the position and number of repetitions of a given phrase in the melody in which it occurs, as well as expectancy and motif repetivity. We show that all of these predictors account for the observed variation to a moderate degree, and that, as hypothesized, those phrases vary less which are rather short, contain highly expected melodic material, occur relatively early in the melody, and contain small pitch intervals. A large portion of the variance is left unexplained by the current model, however, which leads us to a discussion of future approaches to study memorability of melodies.

Keywords: music information retrieval, music cognition, recall, memorability, stability, folk songs, corpus analysis

OPEN ACCESS

Edited by:
Geraint A. Wiggins,
Queen Mary University of London, UK

Reviewed by:
Esther Adi-Japha,
Bar-Ilan University, Israel
Andrei Radu Teodorescu,
Tel Aviv University, Israel

***Correspondence:**
Berit Janssen
berit.janssen@gmail.com

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *24 October 2017*
Accepted: *04 April 2017*
Published: *25 April 2017*

Citation:
Janssen B, Burgoyne JA and Honing H (2017) Predicting Variation of Folk Songs: A Corpus Analysis Study on the Memorability of Melodies. Front. Psychol. 8:621. doi: 10.3389/fpsyg.2017.00621

1. INTRODUCTION

Songs and instrumental pieces in a musical tradition are subject to change: as they are adopted by a new generation of listeners and musicians, they evolve into something new while retaining some of their original characteristics. The current article investigates to what extent this change of melodies may be explained by hypotheses on the memorability of melodies.

To address this question, we investigate a corpus of folk songs collected in the second half of the twentieth century, in which we can identify groups of variants. The variants are results of real-life melody transmission, something which would be difficult to study in an experimental setting, but for which the present folk song collection possesses high ecological validity. In folk song research, there is a long-standing interest in those melodic segments which resist change during melody transmission. This resistance to change is also referred to as *stability* (Bronson, 1951).

According to models of cultural evolution, the relative frequency of cultural artifacts can be explained based on *drift* alone: certain phrases might have been copied more frequently than others purely based on chance, and the relative stability of a given phrase in a collection of folk songs would be random (Henrich and Boyd, 2002). We hypothesize, instead, that stability can be predicted through the memorability of melodies.

To quantify stability, or the amount of variation a folk song segment undergoes through oral transmission, we follow Bronson's notion that "there is probably no more objective test of stability than frequency of occurrence." (Bronson, 1951, p. 51). We formalize the relative stability of a

melodic segment as its frequency of occurrence across variants of the same folk song. We focus on melodic phrases from the folk songs and employ a novel pattern matching method to determine whether or not a match for a given phrase may be found in a given folk song variant, based on similarity measures tested in Music Information Retrieval, and evaluated on a subset of folk songs in previous work (Janssen et al., in press). We then test whether there is a statistical relationship between a given phrase's matches in variants, and the same phrase's memorability, i.e., properties which might facilitate its recall.

Part of our predictions for the memorability of melodies are drawn from serial recall experiments, which typically test how well participants in studies remember word lists—presented visually or auditorily—or purely visual or spatial cues. Based on this research, we can expect that the length of a phrase might influence its memorability: a phrase with many notes contains more items that need to be correctly reproduced, and will therefore be harder to remember than a phrase with few notes. This does not take into account effects of chunking, which might reduce the memory load of phrases with many notes (Miller, 1956). Recall experiments with lists of different lengths have shown that increasing the length of a memorized list decreases the proportion of correctly recalled items (Ward, 2002). Moreover, rehearsal in the form of phrase repetitions might play a role: a phrase that is repeated several times within a melody might be memorized more faithfully than a phrase that only occurs once in each verse. The repetition can be considered rehearsal, which has been shown to increase retention of items (Murdock and Metcalfe, 1978).

Besides, the position of a melodic phrase within a piece might influence its memorability: in serial recall experiments, these effects are known as *serial position effects* (Deese and Kaufman, 1957). When the start of lists is remembered better, this is considered a *primacy effect* (Murdock, 1962). When words were presented auditorily, Crowder and Morton (1969) found that the end of lists were remembered better, which might lead one to expect that melodies, also auditory in nature, exhibit a *recency effect*. However, in Rubin's (1977) experiments on long-term retention of well-known spoken word passages (the Preamble to the constitution of the United States, Psalm 23, and Hamlet's monolog from the eponymous Shakespeare play), words at the start of such a passage are recalled better than items in the middle or at the end. As this situation is maybe closest to singing a folk song from memory, we assume that phrases at the start of melodies may also be more stable. Of course, serial position effects may be caused by an individual's more frequent exposure to items early or late in a melody (Ward, 2002), in which case we would expect that rehearsal is more important than serial position to explain the stability of melodic segments.

Next to these general hypotheses on recall, we test hypotheses based on melody recall research. Firstly, a significant body of research links melody recall to expectancy. According to Kleeman (1985), only music which can be processed by listeners based on their musical expectations, will be selected for transmission in a musical tradition (p. 17). Supporting this, Schmuckler (1997) found a relationship between expectancy ratings and melody recall in an experimental study on folk song melodies. To this end, 16 participants were instructed to rate how well artificial variants of 14 folk songs confirmed their expectancy. The variants of the folk songs were generated by scrambling the notes at the end of each song, maintaining the rhythmical structure and the end note. Afterwards, participants had to identify the melodies they had encountered in the first part of the experiment, presented along with previously unheard melodies. The hit rates were positively correlated with the expectancy rating, indicating that those melodies which conform best to melodic expectations of listeners are also most reliably recalled.

An alternative prediction would be that it is actually more unexpected items which are easier remembered. This is corroborated by evidence from free recall, where items which are unusual are usually better remembered (von Restorff, 1933). For music, Müllensiefen and Halpern (2014) found that memorability of melodies was increased if they contained a large amount of unique motifs, i.e., melodic material which is unusual and therefore unexpected. This means that expectancy may influence variation of melodies in opposing ways, which we both adopt as hypotheses (see hypotheses 4a and 4b in the list of hypotheses below).

Different formalizations of melodic expectancy exist, among which the influential implication-realization theory by Narmour (1990) predicts that the direction and distance, or *pitch interval*, between two ensuing musical tones implies the direction and size of the next pitch interval. Schellenberg (1996) quantified the principles that Narmour defined, such that for a given implicative pitch interval, there is a measurable expectancy of which note is likely to ensue. He performed three listening experiments in which listeners rated how well the last note fulfilled their expectations after listening to excerpts from British and Chinese folk songs, and from atonal music, and reanalyzed data from Unyk and Carlsen (1987). His experiments showed that the quantified implication-realization principles were highly correlated with listeners' expectancies.

Schellenberg found that Narmour's model can be reduced to two factors, *pitch proximity* and *pitch reversal*, without significant loss in explanatory power (Schellenberg, 1997). Hence, Schellenberg's simplified model can be considered a quantification of expectancy, which may predict how well a given melody is retained in a musical tradition.

Inspired by an article by Meyer (1957), Conklin and Witten (1995) approach expectancy with information-theoretical measures: according to Meyer's theory, expectancies are generated by learned probabilities of given events. A listener expects musical events she has heard frequently before, and will be surprised by musical events she hears for the first time. Conklin and Witten assume that this learning, and hence expectancy, can be based on different musical aspects, such as pitches, pitch intervals or durations, among others. For this, they developed a predictive model based on various musical aspects, which they refer to as *viewpoints*.

Conklin and Witten's model applies Prediction by Partial Matching (Cleary and Witten, 1984) to a given note event, expressed by one or several viewpoints. Prediction by Partial Matching (PPM) is a statistical model that is trained on the frequencies of *n-grams*, or sequences of *n* symbols, in a collection

of documents, and which can then be used to predict a symbol in an unseen document given its context. In music prediction, the symbols are musical notes, described by various viewpoints, e.g., pitch, duration, pitch interval, or accentuation. If the model encounters a note sequence it has not seen in the learning phase, it will backtrack to the next shorter note sequence which it did encounter, and use the frequency of the shorter sequence to predict the following note.

Pearce and Wiggins (2004) extended Conklin and Witten's model such that the length of the musical sequence, or the order of the *n-gram*, is variable. Pearce and Wiggins confirmed that statistical information as modeled by their system, dubbed IDyOM (Information Dynamics of Music)[1], predicts listener's expectancy ratings from various listening experiments on folk songs, hymns and single intervals to a great extent (Pearce and Wiggins, 2012).

Some recent corpus studies of popular music have indicated that the presence of repeating motifs in a melody or phrase may enhance its memorability. As such, Müllensiefen and Halpern (2014) investigated a large number of musical features derived from music notation of 80 Western pop songs, to see which of them would best predict the memorability of 80 pop song excerpts. The memorability was determined in a recall experiment with 34 participants, who listened to 40 excerpts and later were presented with all of the excerpts, having to indicate whether they had heard the song before, and how pleasant they considered the excerpt in question. The researchers considered responses on the pleasantness to represent *implicit* memory for the music, through the mere exposure effect (Zajonc, 1980). Müllensiefen and Halpern's results indicate that a melody is more easily remembered explicitly if it consists of motifs which are repeated within the melody. For the implicit memory of melodies, however, it was found that motifs should not repeat too much.

Van Balen et al. (2015) measure the memorability of pop songs that participants are likely to have heard through radio and other media. They register this memorability through reaction times in a game. The goal of the game is to indicate whether or not the player recognizes a given song segment (cf. Burgoyne et al., 2013). If the player's response is fast, Van Balen and colleagues surmise that the song segment in question is very memorable, or catchy. They use a range of features to predict the memorability of the melodies, among which the features used by Müllensiefen and Halpern (2014).

One of Balen and colleagues' strongest predictors of memorability turned out to be motif repetivity, which is in line with Müllensiefen and Halpern's findings on explicit melody recall. As our study focusses on melodies which were explicitly remembered by their singers, rather than pleasantness ratings of these melodies, we therefore adopt the prediction that motif repetivity will increase a phrase's stability. Motif repetivity can also be seen as related to chunking, as repeating motifs would provide meaningful subdivisions within a phrase. Chunking has been shown to facilitate learning in various domains (Gobet et al., 2001).

Based on the above observations, in the current paper we investigate the following five hypotheses of how variation of folk songs may be predicted through theories on melody recall:

1. Shorter phrases show less variation.
2. Phrases which repeat within their source melody show less variation.
3. Phrases which occur early in their source melody show less variation.
4. A phrases' expectancy is related to its variation.
   (a) Phrases which contain highly expected melodic material show less variation.
   (b) Phrases which contain highly surprising melodic material show less variation.
5. Phrases composed of repeating motifs show less variation.

2. MATERIALS AND METHODS

Our research was carried out using the folk song corpus (FS) from the Meertens Tune Collections[2]. This corpus comprises 4,125 digitized transcriptions of monophonic songs, of which the largest part has been recorded in field work between 1950 and 1980 (Grijp, 2008). 1,245 transcriptions originate from song books of the nineteenth and twentieth century known to contain variants to the recorded songs.

The corpus has been categorized into *tune families*, or groups of variants, by domain experts (c.f. Volk and van Kranenburg, 2012), and we use these pre-defined groups to investigate stability between song variants. We compare variants from the same tune family. Each variant is considered to represent the variation imposed by a particular singer or song book editor to a given melody. Consequently, we analyze which phrases of the songs belonging to a tune family vary more, or vary less between different variants: if a phrase occurs in many variants, this means that this phrase is less subject to change, or more stable.

To this end, we separate the FS corpus into three sub-corpora: (1) a training corpus of 360 melodies for which annotations of phrase occurrences were available; (2) a test corpus of 1,695 melodies with tune families comprising at least five variants, but excluding tune families from the training corpus; (3) a background corpus of 1,000 melodies with tune families comprising very few variants. All melodies which could potentially be related to melodies from the test corpus—because they might be hitherto unrecognized variants of a tune family in the test corpus (tune family membership undefined), or because they were subtypes of a tune family in the test corpus—were excluded from the background corpus.

The training corpus was used to train the computational method to find phrase occurrences; the background corpus was used to train information theoretical models; the test corpus was used to test the relationship between variation of the folk song phrases and their hypothesized memorability.

[1] https://code.soundsoftware.ac.uk/projects/idyom-project.

[2] www.liederenbank.nl/mtc.

2.1. Detecting Phrase Occurrences

To quantify the amount of variation, or stability of a given melodic phrase (the query phrase), we detect its occurrences in melodies belonging to the tune family from which it was taken (its source tune family): the more variants of the source tune family the query phrase occurs in, the higher the stability of the phrase.

We detect occurrences through pattern matching, or the computational comparison of the query phrase to all melodies in its source tune family. The extent to which any segment in a given melody resembles the query phrase can be detected through various similarity measures. Earlier research on the above-mentioned training corpus with phrase occurrence annotations has shown that a combined measure of the similarity measures city-block distance (Steinbeck, 1982), local alignment (Smith and Waterman, 1981) and structure induction (Meredith, 2006) reproduces human annotations of phrase occurrences best. The similarity measures, as well as the way in which they were combined, are described in the Supplementary Material.

Research on the training corpus also showed which similarity score should be used as a threshold to separate between relevant occurrences (i.e., detected matches which were also annotated as instances of the query phrase) and irrelevant occurrences (i.e., detected matches which were not annotated as instances of the query phrase) for each of the three measures (Janssen et al., in press). This optimal similarity threshold results in the best trade-off between missing as few relevant occurrences as possible, while producing as few as possible irrelevant occurrences.

Our previous research indicated that the combined measure produces errors in comparison to human annotators, i.e., it misses about 30% of the relevant occurrences, and detects about 8% irrelevant occurrences. The percentage of produced errors differs depending on the analyzed tune family. Using the pattern matching procedure, for the 9,639 phrases from 147 tune families in the test corpus, we receive 170,803 computational judgements on the occurrences of these phrases in their respective source tune families.

2.2. Formalizing Hypotheses

This section describes the formalization of hypotheses on memorability of melodies[3]. For illustration purposes, we present a running example in **Figure 1**, a folk song melody from the tune family *Van Peer en Lijn (1)*, part of the test corpus. This melody has ten phrases and shows how under the current formalizations, different hypotheses arrive at different predictions of stability for each phrase. Throughout this section, we refer to a query phrase as q, which is taken from its source melody, s. The source melody's notes are referred to as s_j. The query phrase starts at index $j = a$ and has a length of n notes.

2.2.1. Influence of Phrase Length

We test whether the length of the phrases has influence on their stability by defining the *phrase length* as the number of notes n of which a given phrase is composed.

$$Len(q) = n \tag{1}$$

[3]The implementations of the hypotheses can be found at https://github.com/BeritJanssen/Stability.

In the example melody, the shortest phrases (phrase 2 and 4) have a length of seven notes, the longest phrase (phrase 9) has 16 notes. According to the prediction of the list length effect, we would expect the second and fourth phrases to be more stable than the ninth phrase. Over the whole dataset, phrase length takes values in the range $[3, 26]$ in the dataset, with a mean of $\overline{Len} = 9.11$ and a standard deviation of $\mathbf{SD}(Len) = 2.23$.

2.2.2. Influence of Rehearsal

Rehearsal is modeled based on phrase repetitions: if a phrase is repeated multiple times within a melody, it is subject to more rehearsal, hence it may be expected to be more stable. The resulting predictor, *phrase repetition*, is measured by counting the occurrences of a phrase in its source melody. All phrases in a melody s are defined as sets of notes P_{id}. id refers to the sequential index of the phrase P in the melody. Each phrase's notes are represented by their onset from the start of the phrase and their pitch. The query phrase is a set of notes Q with the same representation. For every phrase P_{id} we determine its equality score to Q as follows:

$$Eq(P_{id}, Q) = \begin{cases} 1 & \text{if} \quad P_{id} = Q \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Then we measure the number of phrase repetitions Rep of the query phrase q by summing the equality scores of all f phrases P_{id} in the melody.

$$Rep(q) = \sum_{id=1}^{f} Eq(P_{id}, Q) \tag{3}$$

In the example melody, the first and second phrase repeat exactly as the third and fourth phrase, respectively. The other phrases do not repeat anywhere in the melody. This means that phrase repetition is $Rep = 2$ for the first four phrases, $Rep = 1$ for the other six phrases. This would lead to the prediction that the first four phrases are more stable than the last six phrases. Phrase repetition takes values in the range of $[1, 4]$ in the dataset, with a mean of $\overline{Rep} = 1.17$ and a standard deviation of $\mathbf{SD}(Rep) = 0.39$.

2.2.3. Influence of the Primacy Effect

We test the primacy effect based on the position of a phrase in its source melody. We formalize the *phrase position* as a given phrase's sequential index, q_{id}, from $q_{id} = 1$ to $q_{id} = g$ for all g phrases in the source melody. For the example melody of **Figure 1**, $g = 10$.

$$Pos(q) = q_{id} \tag{4}$$

In the example melody, the first phrase has a value of $Pos = 1$, and the last phrase a value of $Pos = 10$. Phrase position takes values in the range of $[1, 22]$ in the dataset, with a mean of $\overline{Pos} = 3.44$ and a standard deviation of $\mathbf{SD}(Pos) = 2.06$.

2.2.4. Influence of Expectancy

To quantify expectancy, we make use of two formalizations: one by Schellenberg (1997), which is based on observations from

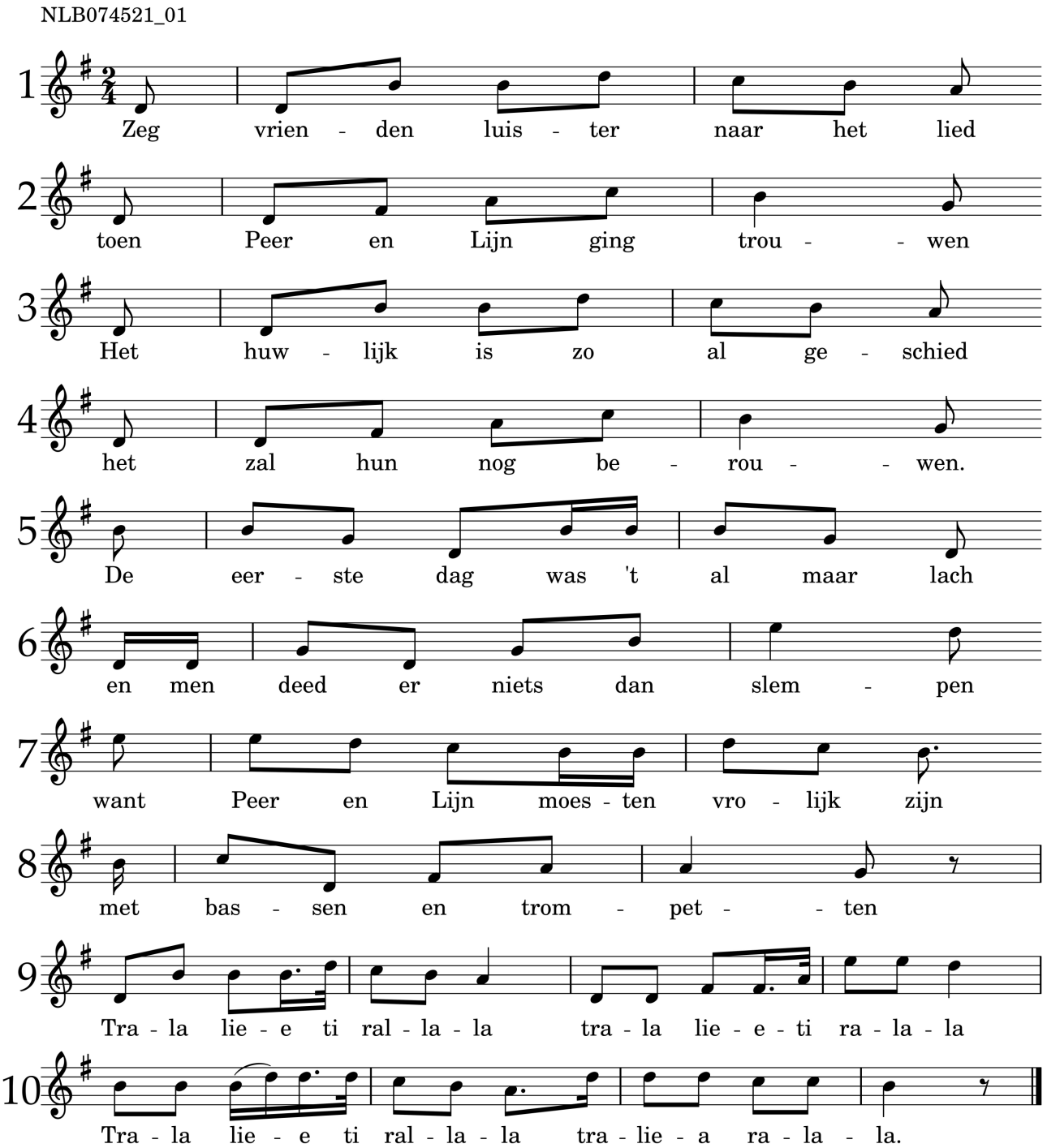


FIGURE 1 | An example melody from the test corpus, belonging to the tune family *Van Peer en Lijn (1)*, which comprises six variants. This melody is used to illustrate the formalizations of the hypotheses. The number on top of the sheet music shows the record number in the Dutch folk song database, the numbers left of the staves show the sequential phrase indices. A recording can be found at http://www.liederenbank.nl/sound.php?recordid=74521&lan=en.

music theory, and one by Pearce and Wiggins (2004), which is based on statistical analysis of a background corpus.

We base both models on pitch intervals between consecutive notes. The pitch of a given note $pitch(s_j)$, or its height in the human hearing range, is represented by its MIDI note number. This entails that pitches are integers, in which a semitone difference between two pitches is indicated by a difference of one. The pitch interval between a note s_j and its predecessor s_{j-1} is

defined by $pInt(s_j) = pitch(s_j) - pitch(s_{j-1})$, where a positive sign indicates that the preceding note is lower, and a negative sign that the preceding note is higher. Both models make predictions for single *notes*, rather than whole phrases. We derive predictions for whole phrases through averaging the note values over the length of the phrase.

2.2.4.1. Expectancy: music theory

The first component of Schellenberg's model, *pitch proximity*, states that listeners expect small steps between melody tones. The further a given note is away from its predecessor, the more unexpected it is. The model does not make any predictions for pitch intervals equal to or larger than an octave.

$$PitchProx(s_j) = \begin{cases} \left| pInt(s_j) \right| & \text{if} \quad \left| pInt(s_j) \right| < 12 \\ \text{undefined} & \text{otherwise} \end{cases} \tag{5}$$

In **Figure 2A** we show the first phrase of the example melody, with the pitch proximity values printed underneath each note, referring to the pitch interval to its preceding note. Note that the pitch interval, and therefore pitch proximity, is not defined for the first note of a melody, as there is no previous pitch from which a pitch interval could be measured.

To calculate the pitch proximity of a phrase, the pitch proximity values of the notes s_j belonging to a given phrase are averaged over the whole phrase, and the negative value of this average is used for easier interpretation, such that if a phrase has a high value of pitch proximity, its pitches are close to each other, while lower values indicate larger pitch intervals. Notes for which pitch proximity is not defined are discarded from the averaging procedure.

$$Prox(q) = -\frac{1}{n}\sum_{j=a}^{a+n} PitchProx(s_j) \tag{6}$$

We show the pitch proximity values for the seventh and eighth phrase of the example melody in **Figure 2A**. The average proximity of the two phrases amounts to $Prox = -13/9 = -1.44$ and $Prox = -20/7 = -2.85$, respectively, which means that we would expect the seventh phrase to be more stable than the eighth phrase. Pitch proximity takes values in the range of $[-6.0, 0.0]$ in the whole data set, with a mean of $\overline{Prox} = -2.01$ and a standard deviation of $\mathbf{SD}(Prox) = 0.69$.

The other factor in Schellenberg's model is *pitch reversal*, which summarizes the long-standing observation from music theory that if leaps between melody notes do occur, they tend to be followed by stepwise motion in the opposite direction (Meyer, 1956). See the Supplementary Material for the quantification of this principle, which for a given melody note results in values ranging from $PitchRev(s_j) = -1$, or least expected, to $PitchRev(s_j) = 2.5$, or most expected. As with pitch proximity, we calculate the average reversal of a phrase through calculating the arithmetic mean of its constituent notes. As pitch reversal makes predictions based on two pitch intervals, it is not defined for the first two notes of a melody. Notes for which pitch reversal is not defined are discarded from the averaging procedure.

$$Rev(q) = \frac{1}{n}\sum_{j=a}^{a+n} PitchRev(s_j) \tag{7}$$

We show the pitch reversal values for the seventh and eighth phrase of the example melody in **Figure 2B**. The average reversal of the two example phrases amounts to $Rev = 3/9 = 0.33$ and $Rev = 1/7 = 0.14$, respectively, which means that we would expect the seventh phrase to be more stable than the eighth phrase. Pitch reversal takes values in the range of $[-0.5, 1.5]$ in the whole data set, with a mean of $\overline{Rev} = 0.30$ and a standard deviation of $\mathbf{SD}(Rev) = 0.24$.

2.2.4.2. Expectancy: information theory

The IDyOM (Information Dynamics of Music) model by Pearce analyzes the frequencies of *n-grams* in a music collection, and based on these observed frequencies, assigns probabilities to notes in unseen melodies, given the notes preceding it. The preceding notes are also referred to as *context*. The length of the context can be set by the user. If the model cannot find a relevant *n-gram* of the context length specified by the user, it backtracks to shorter melodic contexts, and uses those frequencies to return the probability of a given note.

We let the model analyze our background corpus, with the melodies represented as pitch intervals. As we are interested in contexts of phrase length, we limit the *n-gram* length to the average phrase length of nine. We use IDyOM's long-term model, i.e., the model does not update itself while observing the query phrases, and we apply the interpolation weighting scheme *C*, which balances longer and shorter melodic contexts evenly. This was proven to be the best performing weighting scheme in experiments by Pearce (2005).

We express the expectancy of a given melodic segment through its average information content. Information content is the natural logarithm of the inverse probability $\mathbb{P}(s_j)$ of a note to occur given the previous melodic context, based on the probabilities of the background corpus. We choose information content rather than probability, as the logarithmic representation makes it possible to compare the typically small probability values in a more meaningful way. Information content is often also referred to as *Surprisal*, as its values increases as events get *less* expected.

We average the information content of all notes in a query phrase by their arithmetic mean, which is equivalent to a geometric mean of the probabilities. We call this average information content surprisal in the following, to indicate that higher values denote less expected phrases.

$$Sur(q) = \frac{1}{n}\sum_{j=a}^{a+n} \log\left(\frac{1}{\mathbb{P}(s_j)}\right) \tag{8}$$

We show the information content for the seventh and eighth phrase of the example melody in **Figure 2C**. The context used

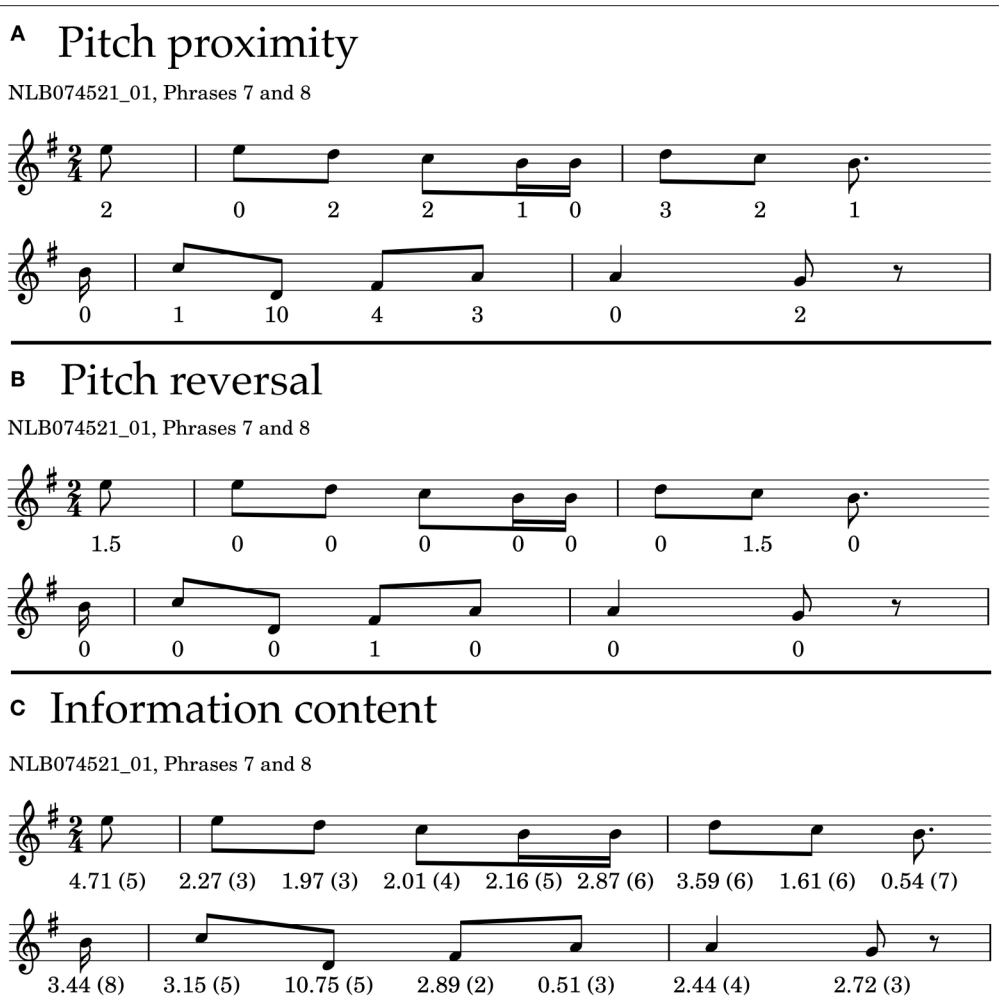


FIGURE 2 | Phrase 7 and 8 of the example melody, showing the values for each note resulting from different theories. (A) Values according to Schellenberg's pitch-proximity principle. **(B)** Values according to Schellenberg's pitch-reversal principle. **(C)** Information Content, calculated with IDyOM, based on a background corpus. The numbers in brackets indicate how much context is considered to calculate information content, which in this case ranges from 2 (two previous notes considered) to 8 in the second phrase (eight previous notes considered).

to generate the information content is shown in brackets. The surprisal of the two example phrases amounts to $Sur = 21.74/9 = 2.42$ and $Sur = 25.88/7 = 3.7$, respectively, which means that we would expect the seventh phrase to be more stable than the eighth phrase. Surprisal takes values in the range of $[1.15, 6.86]$ in the whole data set, with a mean of $\overline{Sur} = 2.68$ and a standard deviation of $\mathbf{SD}(Sur) = 0.53$.

2.2.5. The Influence of Repeating Motifs

As Müllensiefen and Halpern (2014) and Van Balen et al. (2015) found a relationship between repetitiveness of short motifs and the recall of a melody, we follow their procedure and use the FANTASTIC toolbox (Müllensiefen, 2009) to compute a frequency table of such short motifs t for each phrase. FANTASTIC uses a music representation which codes the relative pitches and durations of consecutive notes, see the Supplementary Material for a detailed description.

We follow Müllensiefen (2009) in their formalization to measure repeating motifs through entropy. The motifs are *n-grams* of character sequences representing the pitch and duration relationships between notes. The lengths of motifs to be investigated can be determined by the user. For each investigated motif length l, the frequency of unique motifs $v_{z,l}$ is counted, and compared to the total number of motifs of that length $N_{t,l}$ covering the phrase. The entropy H_l is then calculated from each unique motif's relative frequency $f_{z,l}$, i.e., how often a given motif $v_{z,l}$ occurs in a phrase with relation to all motifs of that length in the phrase.

The relative frequencies of all unique motifs are multiplied with their binary logarithm, summed, and divided by the binary logarithm of the number of all motifs of that length in the phrase ($N_{u,l}$) for normalization. A value of $H = 1.0$ then indicates maximal entropy, and minimal repetitiveness: there are no repeating motifs of length l at all in the phrase; a lower value indicates that there are some

repeating motifs.

$$H(l) = -\frac{\sum_{z=1}^{N} t,l f_{z,l} \cdot \log_2 f_{z,l}}{\log_2 N_{u,l}} \quad (9)$$

The mean entropy of the motifs is then the average over all possible motif lengths. We analyze, in accordance with earlier work, motifs from two notes to six notes in length. We take the negative value of this average to define motif repetivity: the higher the average entropy, or the more distinct motifs in the phrase, the lower the repetivity.

$$MR(q) = -\frac{\sum_{l=2}^{6} H(l)}{5} \quad (10)$$

To illustrate the concept, refer again to **Figure 1**, in which the second and fourth phrase, consist of repeated steps up by a third. This sequence can be subdivided into three identical sequences of two notes each (as the representation of the FANTASTIC toolbox does not distinguish between minor and major intervals): this would mean that this phrase has higher motif repetivity than, for instance, the sixth phrase. See the Supplementary Material for an example calculation of the motif repetivity of the second/fourth and the sixth phrase. The motif repetivity of the second/fourth phrase amounts to $MR = -0.90$, and of the sixth phrase to $MR = -0.98$, so we would expect the second and fourth phrase to be more stable than the sixth phrase. Motif repetivity takes values in the range of $[-1.0, 0.0]$ in the whole data set, with a mean of $\overline{MR} = -0.92$ and a standard deviation of $\mathbf{SD}(MR) = 0.09$.

2.3. Measuring Statistical Relationships

Since our outcome variables are binary, i.e., a given query phrase occurs or does not occur in a given melody, we model the statistical relationship between the likelihood that a given query phrase occurs and its properties through logistic regression. In logistic regression, the odds that an event happens are predicted as a function of one or multiple independent variables. The logarithm of the odds is also known as the *logit* function, where $\mathbb{P}$ stands for the probability that an event happens:

$$logit(\mathbb{P}) = \log(\frac{\mathbb{P}}{1-\mathbb{P}}) \quad (11)$$

The goal of logistic regression is to find a curve that best separates the true events from the false events. In our case, this means that we want to predict the probability $\mathbb{P}$ that a given query phrase q has a match in a given melody s, based on the vector $\mathbf{F}$ of the independent variables hypothesized to contribute to long-term memorability of melodies.

$$logit(\mathbb{P}) = \beta\, \mathbf{F} + \epsilon \quad (12)$$

In this equation, β represents the slope of the prediction function, ϵ represents the random effects of the model, i.e., the random error for each melodic segment, assumed to be normally distributed. If the prediction of the probability of occurrence (i.e., the inverse logit of the prediction function) were perfect, this would lead to a curve separating the occurrences clearly from the non-occurrences.

However, the tune family dependent error of the computational method detecting occurrences needs to be taken into account. This could be done by separate logistic regression models for each tune family; yet this would mean that we could not globally estimate how well a specific hypothesis accounts for probability of occurrence. We therefore choose another solution to model the relationship between phrase properties and occurrence: a generalized linear mixed model (GLMM) which can model the variation of all data at the same time.

Generalized linear models are a framework in which relationships between independent variables and dependent variables of binomial, multinomial, ordinal and continuous distributions can be investigated. A special case of this framework are mixed models, in which not only a general random effect (ϵ), but also random effects for subgroups of the data can be taken into account. This way, we can model the tune family dependent error of the computational method. We assume that every tune family has a different intercept term in the model, i.e., the height at which the logistic regression curve crosses the y axis. Hence, the decision function between occurrence vs. non-occurrence of the model is shifted, depending on the tune family.

We again assume $\mathbf{F}$ as the vector representing the independent variables of the query phrases, β as the slope of the prediction function, ϵ as the random error, but now also take into account the random effect μ, based on the individual error of each tune family, summarized in the vector $\mathbf{tf}$. Then the model can be formalized as follows:

$$logit(\mathbb{P}) = \beta\, \mathbf{F} + \mu\, \mathbf{tf} + \epsilon \quad (13)$$

One could also think of the fixed effects, expressed by μ $\mathbf{tf}$ as the between-tune-family variance, and the random effects, expressed by ϵ, as the within-tune-family variance. Using this model, we test our hypotheses on possible correlates of long-term melody recall.

To be able to compare the independent variables derived from our hypotheses, we standardize all variables x of the predictor vector by subtracting the arithmetic mean $\overline{x}$, and dividing by the standard deviation $\mathbf{SD}(x)$ of a given variable.

$$\mathbf{F}_x = \frac{x - \overline{x}}{\mathbf{SD}(x)} \quad (14)$$

This leads to the overall model for all phrase occurrences, in which units can be compared against each other. We apply a Generalized Linear Mixed Model with fixed slopes and random intercepts for each tune family through the R package LME4[4] to the test corpus of the dataset containing 9,639 phrases from 147 tune families.

2.4. Model Selection

We select the independent variables contributing to the strongest model predicting long-term memorability of folk song phrases,

[4]https://CRAN.R-project.org/package=lme4.

TABLE 1 | The best models for different degrees of freedom, from 3 df with one parameter, to 9 df with seven parameters.

Parameter estimate	3 df	4 df	5 df	6 df	7 df	8 df	9 df
Surprisal	−0.27	−0.29	−0.30	−0.30	−0.29	−0.24	−0.24
Phrase length		−0.32	−0.32	−0.33	−0.30	−0.31	−0.30
Phrase position			−0.10	−0.12	−0.12	−0.10	−0.10
Phrase repetition				0.08	0.09	0.09	0.09
Motif repetivity					0.08	0.08	0.08
Average proximity						0.09	0.10
Average reversal							0.05
AIC_c	209159.8	206889.7	206584.4	206355.5	206157.6	206012.1	205941.5

For each model, the second order Akaike information criterion (AIC_c) is shown, with lower values indicating better model fit. Surprisal is the parameter which leads to the best model with only one predictor; the other parameters are listed in the order by which they are added, leading to the best model fit when all parameters are used.

using the R library MuMIn[5]. This model selection compares all possible combinations of independent variables and ranks them based on their second-order Akaike information criterion (AIC_c) (Hurvich and Tsai, 1989). The second-order Akaike information criterion penalizes the addition of extra parameters to a model, such that it strikes a balance between model fit and parsimony (Burnham and Anderson, 2004). Furthermore, we estimate the effect size of the best model with a technique to determine R^2 of mixed models by Nakagawa and Schielzeth (2013).

3. RESULTS

We show the best models selected from three degrees of freedom (3 df), with one model parameter, to nine degrees of freedom (9 df), with seven model parameters, in **Table 1**. The models' second-order Akaike information criteria decrease as more parameters get added, indicating better model fit. Our results show that the strongest model for the stability of melodic phrases is the full model with all independent variables: phrase length, phrase repetition, phrase position, pitch proximity, pitch reversal, surprisal and motif repetivity. This model yields an AIC_c lower by 70.65 than the second best model. **Table 2** shows the estimated prediction coefficients, the variances of the tune family dependent error and the residual error for the full model, as well as the model fit in R^2. The fixed effects alone, marginalized, explain $R^2_{marginal} = 0.05$, or about 5% of the variance, which is a mid-sized effect for mixed models (Cohen, 1992; Kirk, 1996). When the tune family dependent random effects are considered along with the fixed effects ($R^2_{conditional}$), 22% of the variation in the data is explained.

The prediction coefficients show that phrase length and surprisal possess most predictive power: with increase of a given query phrase's length, its stability decreases. Higher expectancy leads to increased stability. Furthermore, the coefficients also indicate that earlier phrases tend to be more stable, as with an increase in the phrase index, the odds that a query phrase occurs in a given melody are decreased. Moreover, an increase in pitch proximity, or a decrease in the average size of the pitch intervals in a phrase, leads to a higher chance of an occurrence. More repetitions of a query phrase also result in the increased odds of occurrence. Pitch reversal and motif repetivity contribute least strongly to the model, but the signs of the parameters are as expected: if a phrase confirms expectations of pitch reversal, its odds of occurrence are increased, and likewise, if a phrase contains many repeating motifs, its odds of occurrence are increased.

We also tested the model for multicollinearity, confirming that the approximate correlations of parameter estimates do not exceed 0.6, which justifies our treatment of the model parameters as independent predictors.

To illustrate the predictions of the model, we show the predicted as well as the observed frequency of occurrence for the ten phrases of the example melody in **Figure 3**. According to the model, the first four phrases have the highest probability of occurrence, and indeed these phrases also have the highest observed frequency of occurrence (i.e., stability). The predictions do differ from many of the observed values, as for instance the higher stability of phrase 1 and 3 as compared to phrase 2 and 4 is not captured by the model.

4. DISCUSSION

The current research shows that folk song collections are a valuable resource for studying the relationship between melody variation and memorability. All proposed hypotheses relating to recall in general and music recall in particular contribute to prediction of folk song variation, as model selection among all combinations of parameters leads to a model with all hypotheses as predictors.

Of course, the variation that is explained with the current model is still rather low at $R^2 = 0.05$. This might mean that there are potentially more, and stronger predictors for melody variation that have not been tested in this study. It is also good to keep in mind that the phrase occurrences in folk songs do not represent "clean" experimental data in which all aspects but melody recall are controlled. The ecological validity comes at the cost of potential noise. Some aspects that might deteriorate the observed variation are (a) the computational method to detect occurrences; (b) the inherent ambiguity of phrase occurrences,

[5]https://CRAN.R-project.org/package=MuMIn.

i.e., humans do not agree on occurrences perfectly (Janssen et al., in press); (c) a bias in the corpus toward specific regions and demographic groups (Grijp, 2008).

Alternatively, one could assume that a large proportion of melody variation is a result of drift, and therefore random (Henrich and Boyd, 2002). Therefore, it is enlightening that the hypotheses *do* contribute to explaining variation in the dataset, in spite of potential noise in the data. Memorability predicts the amount of melodic variation, or stability, as follows: phrases which resist change should be short (list length effect, hypothesis 1) and contain little surprising melodic material (i.e., low surprisal, a formalization of expectancy, hypothesis 4a). Moreover, it is beneficial if a phrase occurs relatively early in a melody (primacy effect, hypothesis 3), and has mostly small pitch intervals (i.e., high average proximity, a formalization of expectancy, hypothesis 4a). The repetition of a phrase in its source melody also contributes to its memorability (rehearsal effect, hypothesis 2), even though this effect is somewhat weaker in our analysis than other predictors. Average reversal, or the tendency for a melody to adhere to the gap fill principle, i.e., following a leap by stepwise motion in the opposite direction (expectancy, hypothesis 4a) and motif repetivity within the phrase (hypothesis 5) seem to account for long-term memorability to a more limited extent. All predictors related to expectancy indicate that more expected melodic material increases stability, leading us to reject hypothesis 4b.

As for possible drawbacks of the presented study, the three predictors related to expectancy (average proximity, average reversal and surprisal) share the disadvantage that for the first few notes of a melody, no or little information on expectancy is available. This means that there is a potential imbalance between the initial and later phrases of a melody, as the predictor values of initial phrases are based on less information. The alternative, treating every phrase as isolated, so that no context from previous phrases is used for creating expectancy values, seemed unrealistic, however, as the recall of phrases is cued by previous melodic material (cf. Rubin, 1995, p. 190). For the current folk song collection, in which the same melody is sung multiple times with different verses, it may be interesting to investigate in how far considering the end of a given melody as the melodic context for the start of this melody influences expectancy predictions.

The expectancy predictors defined by Schellenberg (1997), average proximity and average reversal, may be comparatively unsuccessful model parameters as they were not necessarily designed to be averaged for a longer melodic context: they were defined to quantify the fulfillment of listener expectations at a given note. However, these predictors still contribute to a better model, which shows that they capture some information on memorability which may predict variation of melodies in this corpus.

The relatively low contribution of motif repetivity as a predictor for melodic variation may be partly ascribed to the fact that the phrases are very short melodic material, and as such rarely contain repeated motifs. It would be interesting to investigate if motif repetivity increases stability of longer melodic contexts, e.g., full folk song melodies. For the current analysis of phrases with an average length of nine notes, which are unlikely to contain repeated motifs longer than four notes, it may be sufficient to limit the maximal *n-gram* length to four notes for future research on motif repetivity in phrases. To hold our use of the method comparable to earlier research, we decided to analyze motifs of the same lengths as previous authors. Moreover, there is no disadvantage to considering longer *n-grams* other than longer computation time, as the FANTASTIC toolbox automatically disregards *n-grams* which are longer than the length of a phrase.

TABLE 2 | The parameters of the best model of the model selection: estimated regression coefficient $\hat{\beta}$ and 95% confidence interval for *phrase length*, *phrase repetitions* within the source melody, *phrase position* in the source melody, *pitch proximity* and *pitch reversal* as defined by Schellenberg (1997), *expectancy*, as defined by IDyOM (Pearce and Wiggins, 2004), and *motif repetivity*, as defined by Müllensiefen (2009).

Parameter	$\hat{\beta}$	95% CI
Intercept	−0.22	[−0.35, −0.08]
Surprisal	−0.24	[−0.25, −0.22]
Phrase length	−0.30	[−0.32, −0.29]
Phrase position	−0.10	[−0.11, −0.09]
Phrase repetition	0.09	[0.08, 0.10]
Motif repetivity	0.08	[0.07, 0.09]
Average proximity	0.10	[0.08, 0.11]
Average reversal	0.05	[0.04, 0.06]
σ_{tf}	0.84	[0.74, 0.95]
$R^2_{marginal}$		0.05
$R^2_{conditional}$		0.22

At the bottom of the table we report the standard deviation of the random effect (tune family), as well as the marginalized and conditional R^2 calculated according to Nakagawa and Schielzeth (2013).

With the current approach, we cannot address the influence of other memory effects on melody variation, such as fill-in effects, spacing effects or confusion errors. Fill-in effects (Conrad and Hull, 1964), which lead to the later inclusion of an item that was skipped earlier in serial recall, may also play a role in melody recall. This might be observed, for instance, if melodic material within a phrase or melody is rearranged, such that a motif which usually starts a melody appears later instead. With the current method, these effects would be missed, as only the amount of melodic variation, but not the kind of melodic variation, is investigated. In the same vein, the spacing effect from free recall (c.f. Hintzman, 1969; Madigan, 1969), which relates to the space between rehearsals of items, cannot be studied on the basis of phrases, which do not necessarily repeat within a melody, and if they do, usually are not spaced far apart. Instead, shorter melodic contexts might be interesting to study to this effect.

Furthermore, confusion errors (Page and Norris, 1998), which in serial recall of words lead to the erroneous recall of acoustically similar words, might also be interesting to study for melody variation. This might occur if instead of a melodic phrase in a given folk song, a similar phrase from another folk song is recalled. As our study analyzes variation per tune family and not across different tune families, melodic material that might correspond between different folk songs is not identified as such.

As our analysis of an existing folk song corpus highlighted some mechanisms behind melodic variation which may be tied

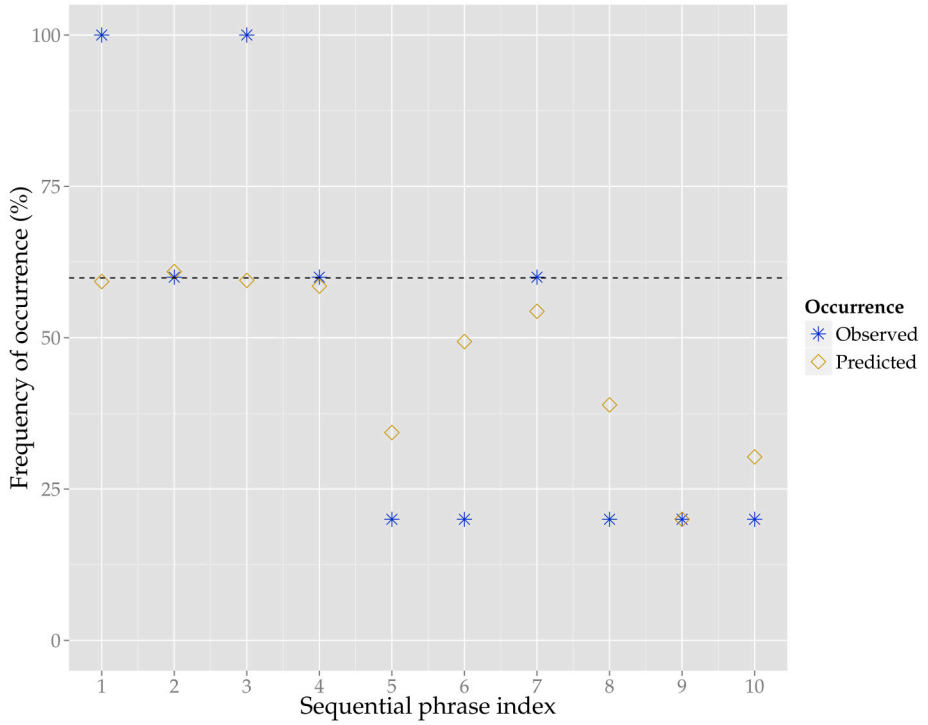


FIGURE 3 | The predicted (yellow diamonds) and observed (blue stars) frequency of occurrence, in percent, for the ten phrases of the example melody. The predictions are generated by the generalized linear mixed model, for the model parameters see **Table 2**. The observed frequency of occurrence is based on how many of the five variants, other than the example melody, contain a given phrase from the example melody. The dashed line shows the model's intercept for frequency of occurrence for this tune family, which is at 58%, meaning that is slightly more likely for the phrases of this tune family to occur in the respective variants than not.

to memorability of melodies, this shows that it would certainly be fruitful to perform more studies based on computational music analysis: such research could be performed on the present folk song corpus to investigate other potential effects of recall (cf. Olthof et al., 2015), or our methods could be applied to other music collections, to see whether our findings can be replicated with respect to melodic variation in other musical traditions.

Next to further computational studies, it would certainly also be an important future contribution to test the predictions on melodic variation in an experiment with human participants. Could the amount of variation when melodies are learned in an experimental setting also be predicted through important parameters of our corpus analysis, e.g., through surprisal, phrase length and phrase position?

As the melodies in the Meertens Tune Collections were recorded or notated long after the singers or editors had learned the melodies, it would also be interesting to investigate whether immediate recall of melodies in a laboratory setting leads to different kinds of variation than if melodies are recalled weeks or months later. As such, the present collection, and other folk song collections, might be an overlooked resource to study recall and long term memory for melodies.

AUTHOR CONTRIBUTIONS

BJ performed the analyses of musical and statistical data and wrote the manuscript. JB advised on the statistical analysis and edited the manuscript. HH advised on the analysis of musical data and edited the manuscript.

FUNDING

BJ is supported by the Computational Humanities Programme of the Royal Netherlands Academy of Arts and Sciences, under the auspices of the Tunes & Tales project. For further information, see http://ehumanities.nl. JB is supported by the Amsterdam Brain and Cognition Talent grant. HH is supported by a Horizon grant (317-70-10) of the Netherlands Organization for Scientific Research (NWO).

ACKNOWLEDGMENTS

We thank our colleagues from the Music Cognition Group for feedback at various stages of the research, Andrei Teodorescu and Esther Adi-Japha for their invaluable comments on the manuscript, the folk song experts at the Meertens Institute for their annotations and data curation, and Peter van Kranenburg for advice and feedback on the current research, as well as the publication of the Meertens Tune Collections.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2017.00621/full#supplementary-material

REFERENCES

Bronson, B. H. (1951). Melodic Stability in Oral Transmission. *J. Int. Folk Music Council* 3, 50–55. doi: 10.2307/835773

Burgoyne, J. A., Bountouridis, D., Van Balen, J., and Honing, H. (2013). "Hooked: a game for discovering what makes music catchy," in *Proceedings of the 14th International Society for Music Information Retrieval Conference* (Porto), 245–250.

Burnham, K. P., and Anderson, R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociol. Methods Res.* 33, 261–304. doi: 10.1177/0049124104268644

Cleary, J. G., and Witten, I. H. (1984). Data compression using adaptive coding and partial string matching. *IEEE Trans. Commun.* COM-32, 396–402. doi: 10.1109/TCOM.1984.1096090

Cohen, J. (1992). A power primer. *Psychol. Bull.* 112:155. doi: 10.1037/0033-2909.112.1.155

Conklin, D., and Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *J. New Music Res.* 24, 51–73. doi: 10.1080/09298219508570672

Conrad, R. and Hull, A. (1964). Information, acoustic confusion and memory span. *Br. J. Psychol.* 55, 429–432. doi: 10.1111/j.2044-8295.1964.tb00928.x

Crowder, R. G., and Morton, J. (1969). Precategorical acoustic storage (PAS). *Percept. Psychophys.* 5, 365–373. doi: 10.3758/BF03210660

Deese, J., and Kaufman, R. A. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *J. Exp. Psychol.* 54, 180–187. doi: 10.1037/h0040536

Gobet, F., Lane, P., Croker, S., Cheng, P., Jones, G., Oliver, I., et al. (2001). Chunking mechanisms in human learning. *Trends Cogn. Sci.* 5, 236–243. doi: 10.1016/S1364-6613(00)01662-4

Grijp, L. P. (2008). "Introduction," in *Under the Green Linden. 163 Dutch Ballads from the Oral Tradition*, eds L. P. Grijp and I. van Beersum (Hilversum: Music & Words), 18–27.

Henrich, J., and Boyd, R. (2002). On modeling cognition and culture representations. *J. Cogn. Cult.* 2, 87–112. doi: 10.1163/156853702320281836

Hintzman, D. L. (1969). Apparent frequency as a function of frequency and the spacing of repetitions. *J. Exp. Psychol.* 80, 139–145. doi: 10.1037/h0027133

Hurvich, C. M., and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* 76, 297–307. doi: 10.1093/biomet/76.2.297

Janssen, B., van Kranenburg, P., and Volk, A. (in press). Finding occurrences of melodic segments in folk songs: a comparison of symbolic similarity measures. *J. New Music Res.* 46. doi: 10.1080/09298215.2017.1316292

Kirk, R. E. (1996). Practical significance: a concept whose time has come. *Educ. Psychol. Meas.* 56, 746–759. doi: 10.1177/0013164496056005002

Kleeman, J. E. (1985). The parameters of musical transmission. *J. Musicol.* 4, 1–22. doi: 10.2307/763720

Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *J. Verb. Learn. Verb. Behav.* 8, 828–835. doi: 10.1016/S0022-5371(69)80050-2

Meredith, D. (2006). "Point-set algorithms for pattern discovery and pattern matching in music," in *Content-Based Retrieval. Dagstuhl Seminar Proceedings 06171*, eds T. Crawford and R. C. Veltkamp (Dagstuhl).

Meyer, L. B. (1956). *Emotion and Meaning in Music*. Chicago, IL: The University of Chicago Press.

Meyer, L. B. (1957). Meaning in music and information theory. *J Aesthet. Art Crit.* 15, 412–424. doi: 10.2307/427154

Miller, G. E. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63, 81–97. doi: 10.1037/h0043158

Müllensiefen, D. (2009). *FANTASTIC: Feature ANalysis Technology Accessing STatistics (In a Corpus): Technical Report*. Technical Report, Goldsmiths University of London.

Müllensiefen, D., and Halpern, A. R. (2014). The role of features and context in recognition of novel melodies. *Music Percept.* 31, 418–435. doi: 10.1525/mp.2014.31.5.418

Murdock, B., and Metcalfe, J. (1978). Controlled rehearsal in single-trial free recall. *J. Verb. Learn. Verb. Behav.* 17, 309–324. doi: 10.1016/S0022-5371(78)90201-3

Murdock, B. J. (1962). The serial position effect of free recall. *J. Exp. Psychol.* 64, 482–488. doi: 10.1037/h0045106

Nakagawa, S., and Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* 4, 133–142. doi: 10.1111/j.2041-210x.2012.00261.x

Narmour, E. (1990). *The Analysis and Cognition of Basic Melodic Structures. The Implication-Realization Model.* Chicago, IL: University of Chicago Press.

Olthof, M., Janssen, B., and Honing, H. (2015). The role of absolute pitch memory in the oral transmission of folksongs. *Empir. Musicology Rev.* 10, 161–174. doi: 10.18061/emr.v10i3.4435

Page, M. P. A., and Norris, D. (1998). The primacy model: a new model of immediate serial recall. *Psychol. Rev.* 105, 761–781. doi: 10.1037/0033-295X.105.4.761-781

Pearce, M., and Wiggins, G. A. (2004). Improved methods for statistical modelling of monophonic music. *J. New Music Res.* 33, 367–385. doi: 10.1080/0929821052000343840

Pearce, M. T. (2005). *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. Ph.D. thesis, City University, London.

Pearce, M. T., and Wiggins, G. A. (2012). Auditory expectation: the information dynamics of music perception and cognition. *Top. Cogn. Sci.* 4, 625–652. doi: 10.1111/j.1756-8765.2012.01214.x

Rubin, D. C. (1977). Very long-term memory for prose and verse. *J. Verb. Learn. Verb. Behav.* 16, 611–621. doi: 10.1016/S0022-5371(77)80023-6

Rubin, D. C. (1995). *Memory in Oral Traditions. The Cognitive Psychology of Epic, Ballads, and Counting-out Rhymes*. New York, NY: Oxford University Press.

Schellenberg, E. G. (1996). Expectancy in melody: tests of the implication-realization model. *Cognition* 58, 75–125. doi: 10.1016/0010-0277(95)00665-6

Schellenberg, E. G. (1997). Simplifying the implication-realization melodic expectancy. *Music Percept. Interdiscip. J.* 14, 295–318. doi: 10.2307/40285723

Schmuckler, M. A. (1997). Expectancy effects in memory for melodies. *Can. J. Exp. Psychol.* 51, 292–306. doi: 10.1037/1196-1961.51.4.292

Smith, T., and Waterman, M. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197. doi: 10.1016/0022-2836(81)90087-5

Steinbeck, W. (1982). *Struktur und Ähnlichkeit. Methoden Automatisierter Melodienanalyse*. Bärenreiter, Kassel.

Unyk, A. M., and Carlsen, J. C. (1987). The influence of expectancy on melodic perception. *Psychomusicology* 7, 3–23. doi: 10.1037/h0094189

Van Balen, J., Burgoyne, J. A., Bountouridis, D., Müllensiefen, D., and Veltkamp, R. (2015). "Corpus analysis tools for computational hook discovery," in *Proceedings of the 16th International Society for Music Information Retrieval Conference* (Malaga), 227–233.

Volk, A., and van Kranenburg, P. (2012). Melodic similarity among folk songs: an annotation study on similarity-based categorization in music. *Music. Sci.* 16, 317–339. doi: 10.1177/1029864912448329

von Restorff, H. (1933). Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychol. Forsch.* 18, 299–342. doi: 10.1007/BF02409636

Ward, G. (2002). A recency-based account of the list length effect in free recall. *Mem. Cogn.* 30, 885–892. doi: 10.3758/BF03195774

Zajonc, R. B. (1980). Feeling and thinking: preferences need no inferences. *Am. Psychol.* 35, 151–175. doi: 10.1037/0003-066X.35.2.151

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 04 July 2017
doi: 10.3389/fpsyg.2017.00931

Acoustic Features Influence Musical Choices Across Multiple Genres

Michael D. Barone[1], *Jotthi Bansal*[1] *and Matthew H. Woolhouse*[2]*

[1] *Psychology, Neuroscience and Behaviour, McMaster University, Hamilton, ON, Canada,* [2] *School of the Arts, McMaster University, Hamilton, ON, Canada*

Based on a large behavioral dataset of music downloads, two analyses investigate whether the acoustic features of listeners' preferred musical genres influence their choice of tracks within non-preferred, secondary musical styles. Analysis 1 identifies feature distributions for pairs of genre-defined subgroups that are distinct. Using correlation analysis, these distributions are used to test the degree of similarity between subgroups' main genres and the other music within their download collections. Analysis 2 explores the issue of main-to-secondary genre influence through the production of 10 feature-influence matrices, one per acoustic feature, in which cell values indicate the percentage change in features for genres and subgroups compared to overall population averages. In total, 10 acoustic features and 10 genre-defined subgroups are explored within the two analyses. Results strongly indicate that the acoustic features of people's main genres influence the tracks they download within non-preferred, secondary musical styles. The nature of this influence and its possible actuating mechanisms are discussed with respect to research on musical preference, personality, and statistical learning.

Keywords: Nokia DB, acoustic features, musical preference, musical genre, music downloads, musical influence, music information retrieval

OPEN ACCESS

Edited by:
Frank A. Russo,
Ryerson University, Canada

Reviewed by:
Daniel Mullensiefen,
Goldsmiths, University of London, United Kingdom
John Ashley Burgoyne,
University of Amsterdam, Netherlands

***Correspondence:**
Matthew H. Woolhouse
woolhouse@mcmaster.ca

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *01 January 2017*
Accepted: *22 May 2017*
Published: *04 July 2017*

Citation:
Barone MD, Bansal J and Woolhouse MH (2017) Acoustic Features Influence Musical Choices Across Multiple Genres. Front. Psychol. 8:931. doi: 10.3389/fpsyg.2017.00931

1. INTRODUCTION

This paper concerns the degree to which the acoustic features of a person's preferred musical genre influence their choice of songs or tracks within non-preferred, secondary musical styles. For example, do people who favor Dance music, which is typically faster in tempo than other styles, listen to up-tempo Jazz and/or Reggae tracks, rather than slower examples of these genres? Similarly, might someone whose preference is for Metal also gravitate toward relatively dynamic, "high-octane" Country or Blues music (assuming, of course, that those genres are not mutually exclusive; Bansal and Woolhouse, 2015). Although conceptually straightforward, this question addresses active research areas within the fields of music cognition and Music Information Retrieval (MIR), and, to some extent, highlights current limitations within both. Firstly, the phenomenon of features within a preferred genre influencing song selection within secondary musical styles falls under the general topic of "cognitive leakiness," a notion explored in depth in the area of consumer choice and commerce (e.g., Rieskamp et al., 2006), but less so in music perception. Secondly, topics involving musical features, in this case extracted from audio, by necessity utilize MIR techniques (e.g., Lartillot and Toiviainen, 2007). The psychological reality of extracted acoustic features is an open question (Friberg and Schoonderwaldt, 2014), and therefore research demonstrating their influence upon musical preference, may, in part, help to legitimize their perceptual existence.

In attempting to investigate song selection and acoustic-feature influence, our study brought music cognition and MIR together within the context of "big data" (Russom, 2011). The data in question consisted of ca 1.3 billion music downloads made by approximately 17 million users in multiple countries between 2007 and 2014. Somewhat anticipating our results, the following analyses demonstrate significant effects with respect to 10 extracted acoustic features, and 10 subgroups of users defined by preferred musical genre. Before describing the data, methodology, and reporting our results, we first review literature that addresses factors responsible for, and that influence, device usage and decision making, including song selection.

Similarly to the devices used in our study, which were mobile phones (see Section 2), Butt and Phillips (2008) sought to predict amounts and types of mobile-phone use from 120 participants rated for extraversion, agreeableness, conscientiousness, neuroticism, and self-esteem (using the Coopersmith self-esteem inventory; Coopersmith, 1959). Individuals assessed as being neurotic, disagreeable, unconscientious, and/or extraverted tended to spend more time messaging using SMS; disagreeable extraverts changed cellphone backgrounds and ringtones more frequently, indicating phone use as a means of stimulation and/or diversion; individuals who scored highly in neuroticism had relatively greater internet use, according to the authors, perhaps in an attempt to overcome loneliness. In sum, Butt and Phillips (2008) concluded that psychological profiling with respect to established personality dimensions could robustly explain how people chose to use their mobile phones.

While successfully modeling human behavior, some researchers (e.g., Ross et al., 2009) have argued that the personality traits referred to above may be too general to model online behavior, including cellphone usage. For example, Hughes et al. (2012) investigated whether a lower-order, relatively narrow personality facet such as Need for Cognition (NFC) was able to predict online social and information-seeking behaviors. NFC is an individual's predisposition to engage with and enjoy information and cognitive endeavors, e.g., news content, crossword puzzles, Sudoku (Haugtvedt et al., 1992; Verplanken, 1993). Despite its specificity, as opposed to a broader dimension such as openness, NFC in the study conducted by Hughes et al. (2012) was found to correlate positively with Twitter usage, presumably due to this social-networking service's relatively high information content. Those with high ratings for sociability and extraversion appeared to prefer Facebook. For additional research concerning social media and personality, see Moore and McElroy (2012).

In addition to device- and personality-specific research, studies exploring the interconnectedness of various forms of media and the consumption of culture, including music, have been undertaken. Finn (1997), in a diary study of over 200 university students, correlated radio listening, TV watching, pleasure reading, and moviegoing with openness, conscientiousness, extraversion, agreeableness, and neuroticism (referred to as the Big Five personality traits; see Costa and MacCrae, 1992). The strongest relationship for mass-media use was between openness and pleasure reading; extraversion was negatively associated with pleasure reading, as was openness and watching TV. Rentfrow and Gosling (2003) assert that the perception of a musical genre depends in part upon the social setting in which it is heard and, by extension, the medium through which it is accessed; in other words, that people's preferences for certain media over others may influence musical categorization. With respect to music listening, where radio continues to play a major role (Peoples, 2015), personality studies have uncovered multiple associations: openness with Blues and Jazz; conscientiousness with Soul and Funk; extraversion with Pop and Rap, and so on (Zweigenhaft, 2008; see also Rentfrow and Gosling, 2003). Moreover, personality appears not only to influence the extent to which individual genres are chosen, but also the overall heterogeneity of our musical tastes, i.e., whether we possess narrow or wide-ranging music-listening habits (Rawlings and Ciancarelli, 1997). In sum, personality research provides evidence for the existence of an overarching psychological framework in which effects akin to cognitive leakiness may occur (Rieskamp et al., 2006). As the research outlined above indicates, personality is a potent phenomenon, suffusing, guiding, and shaping our decisions, including the seemingly inconsequential behavior of choosing music.

In contrast to personality, which is assumed to be relatively stable over time (Leon et al., 1979), mood can undergo rapid affective swings (McFarlane et al., 1988). Moreover, while research has tended to concentrate on how music influences or induces mood, particularly with respect to consumer choice (e.g., Kim and Areni, 1993; North et al., 1999), the converse is also true: mood influences musical choice (Friedman et al., 2012). Which is to say, assuming environmental factors and personal histories to be equal, a person's musical preferences do not depend solely on their personality, but, in addition, are subject to spur-of-the-moment choices influenced by mood.

Amongst the theoretical models advanced to elucidate the role of mood in decision-making, perhaps the most influential is the Affect Infusion Model (AIM), developed by Joseph Forgas in the early 1990s (Forgas, 1995). In brief, the AIM seeks to explain how mood determines a person's capacity to process information—the importance of mood tends to increase in situations involving heavy cognitive load. As information complexity rises, and redundancy falls, the influence of mood on an individual's evaluations and responses increases, resulting in "intuitive" decision-making. Presumably, therefore, when faced with a plethora of diverse musical artists, tracks, and genres, people tend to rely more upon their current mood, in which case the influence of personality may be temporarily reduced or suspended. To the authors' knowledge, within the domain of music-preference research, this hypothesis has yet to be tested.

Despite this possible lacuna within the experimental literature, paradigms employing music-induced moods have produced results that are consistent with aspects of the AIM model. For example, risk-taking varies when mood is induced through listening to preferred vs. disliked music. In a real-money gambling study, in which participants placed bets during either music-liked and disliked trial blocks, Halko and Kaustia (2015) found that people's appetites for risk-taking significantly increased when listening to preferred tracks. They conjectured that listening to preferred types of music increases the "marginal

utility" of money (i.e., the additional satisfaction someone gains from consuming a good or service; Kauder, 2015), which, in turn, increases the likelihood of participating in gambling. Furthermore, Halko and Kaustia (2015) argued that their results are supported by recent studies in neuroscience. Berns et al. (2010) have found levels of activation in reward areas of the brain (e.g., nucleus accumbens) to be proportional to the degree to which music is liked. The behavioral effect of music on risk-taking also co-varies with brain activation in the amygdala and the dorsal striatum (Halko et al., 2015), key brain regions associated with the calculation and assessment of value. In short, in addition to its mood-inducing properties, music listening gives rise to distributed neurological operations in which functionally differentiated networks are simultaneously activated. For a review of research relating to the induction of mood through music, see Västfjäll (2002).

While mood and personality pertain, to some degree, to an individual, shared demographic factors, including culture, education, sex, and age appear to affect people's musical choices (Christenson and Peterson, 1988; Roberts and Henriksen, 1990; Peterson and Kern, 1996; LeBlanc et al., 1999; Schäfer and Sedlmeier, 2009; North and Davidson, 2013). Of these, age is the strongest predictor of musical preference (Christenson and Peterson, 1988). Older adolescents prefer 'lighter' qualities in music compared to younger adolescents (Roberts and Henriksen, 1990). General enjoyment of music from Grade 1 to college drops for a time until rising around the age of puberty, following a U-shaped curve across development (LeBlanc et al., 1996). Supported by cross-cultural studies, sociological research suggests that preferences for eclectic artists rise as national education values improve (Peterson and Kern, 1996). With respect to sex, a music-choice study suggested that males prefer music with themes of dominance and independence, whereas females preferred music with relationship and emotion themes (Christenson and Peterson, 1988). However, the extent to which this research is generalizable is open to debate: almost 30 years has elapsed since Christenson and Peterson's study, which was based on low-sample surveys with relatively little demographic variation. Furthermore, LeBlanc et al. (1999) and North and Davidson (2013) found that demographic information did not conclusively determine music preferences; two- and three-way interactions were found between age, sex and country, and controlling for these factors significantly reduced the strength of the relationships.

Although the foregoing covers aspects of decision-making involving music, none of the research and experimental scenarios referred to above necessarily replicate, or are fully applicable to, the particular issue at hand; namely, the degree to which the features of a person's preferred musical genre influence their choice of tracks within non-preferred, secondary musical styles. A primary motivation for undertaking this research was because, to our knowledge, musical-feature influence has yet to be investigated using large behavioral data sets. Despite not containing user-personality information *per se*, our database of global music consumption afforded us the opportunity to undertake research in this hitherto underexplored area, and, in the process, develop a series of relatively novel analytical methods. The study is divided into two main analyses. Using correlation, Analysis 1 identified differences in feature-dispersion patterns of genre-defined subgroups of users. Analysis 2 involved the exhaustive calculation of feature-influence matrices, which, in combination with central-tendency statistics, were used to detect the influence of main-genre features on those of secondary genres. The methods and results of each analysis follow a description of the data.

2. DATA

2.1. Database

This study utilized a global music-download database, consisting primarily of music metadata made by people downloading tracks and albums onto Nokia mobile phones. The data became accessible through a data-sharing agreement between McMaster University and the Nokia Corporation, begun in 2012; the aim of the agreement was to facilitate sociocultural and musicological research relating to global music consumption. In January 2015, the Nokia division responsible for online music became a separate entity under the name MixRadio; MixRadio ceased commercial operations in February 2016. Henceforth, we refer to the data as pertaining to the Nokia DB[1].

Nokia DB contains downloads from 32 countries, representing every major continent, made between November 2007 and September 2014. In total there are over 1.36 billion track-downloads, relating to a subset of ca 17 million user accounts, and approximately 36 million tracks, written and/or performed by over one million artists. Following the purchase of a mobile device, users could explore artists and tracks without further cost constraints via online stores. Each download's metadata includes information such as track name, artist, album, anonymous user identifier (ID), date, local time, country, and artist-level genre. Supplied by record labels, in total there are 63 genre tags, ranging from mainstream (Country, Pop, Rock) to relatively obscure (Ambient, Flamenco, Khaleeji). For additional information and research concerning Nokia DB, see Woolhouse and Bansal (2013), Woolhouse et al. (2014), Woolhouse and Renwick (2016).

2.2. Data Enrichment

Prior to embarking upon this study, we enriched the download metadata with acoustic features from open-source databases, including The Echo Nest (Bertin-Mahieux et al., 2011). As of May 2016, The Echo Nest application programming interface (API) was subsumed by Spotify; henceforth, for the sake of simplicity, we refer to all extracted acoustic features in our analyses as relating to Spotify. Examples of acoustic features accessed from the Spotify Web API[2] include Acousticness, Danceability, Duration, Energy, Instrumentalness, Liveness, Loudness, Speechiness, Tempo, and Valence. The data are

[1]Nokia DB represents a portion of Nokia's total commercial activity, and is therefore not indicative of market share.

[2]https://developer.spotify.com/web-api/

arranged in a relational database management system and queried using the open-source MySQL 5.1 implementation of SQL (Groff and Weinberg, 2002). In addition, the Python Database API (Lemburg, 2008) enabled more extensive, iterative analyses to be undertaken.

2.3. Acoustic Features

Of the 36 million songs available in Nokia DB, 9 million have been linked to the 10 high- and low-level acoustic features (McKay, 2004, p. 10) listed below. A brief description of each feature now follows; see Jehan and DesRoches (2011) for further information[3].

Acousticness. Value representing the probability that a track was created using acoustic instruments, including voice. Float; range, 0–1.

Danceability. A track's "foot-tapping" quality, based on tempo, rhythm stability, beat strength, and isochrony. Float; range, 0–1.

Duration. The duration of a track in seconds as calculated by the Spotify analyzer. Float; maximum value, 6,060 s.

Energy. A perceptual estimation of frenetic activity throughout a track. High-Energy tracks have increased entropy, and tend to feel fast, loud, and noisy (e.g., Death Metal). Float; range, 0–1.

Instrumentalness. Value representing the probability that a track was created using only instrumental sounds, as opposed to speech and/or singing. Float; range, 0–1.

Liveness. Value representing the probability that a track was recorded in the presence of an audience rather than in a studio. Float; range, 0–1.

Loudness. The average loudness of a track in decibels. Loudness is the psychological correlate of signal amplitude.

Speechiness. Value representing the presence of spoken words in a track, e.g., talk show, audio book, poetry, rap. Float; range, 0–1.

Tempo. The estimated tempo of a track in beats per minute. Float; range, 0–294.

Valence. A perceptual estimation of a tracks positive/negative affect, e.g., happy and cheerful, or sad and depressed. Float; range, 0–1.

2.4. X-heads

The behavioral aspects of our analyses were based on the categorization of users into "X-head" subgroups, where X was the most numerous genre. For example, a user with a majority of Metal downloads was classified as a Metal-head; most Classical downloads, a Classical-head, and so on. This enabled us to identify groups of users that were more accustomed, so we assume, to one particular genre than another, and, thus, attuned to the acoustic features prevailing within that genre. In rare instances where no genre had an absolute majority, the genre of the chronologically earliest download determined a user's categorization.

Our intention was for the definition of an X-head to be as straightforward as possible, and hence our simple criterion of a majority of downloads of a particular genre. In order to keep our study within manageable parameters, 10 X-head subgroups were selected for investigation: Bollywood, Classical, Dance, Jazz, Mandarin Pop, Metal, Pop, Rap/Hip Hop, Reggae, and Rock. Two primary reasons determined this choice: (1) these are amongst the most heavily downloaded genres within Nokia DB; and (2) they include culturally distinct genres, some of which are perhaps less well represented in music-psychology research, e.g., Mandarin Pop. **Table 1** shows the total number of users and tracks per X-head subgroup entered into the analyses.

3. ANALYSIS 1: FEATURE DISTRIBUTIONS

3.1. Method

The initial task in Analysis 1 was to identify feature distributions for pairs of X-head subgroups that were distinct. The reason for this is illustrated in **Figures 1**, **2**. **Figure 1** shows the Energy distributions of tracks belonging to two X-head subgroups: the solid-orange line, D_M, shows the distribution for Dance tracks downloaded by Dance-heads ($Dance_{Main} = D_M$); the solid-blue line, J_M, shows the distribution for Jazz tracks downloaded by Jazz-heads ($Jazz_{Main} = J_M$). Notice that the peak of D_M is to the right, while the peak of J_M is to the left. The two peaks' relative positions indicate that, in general, Dance tracks listened to by Dance-heads have higher Energy than Jazz tracks listened to by Jazz-heads, as calculated by the Spotify analyzer.

Also present within **Figure 1** are lines that show Energy distributions belonging to Dance- and Jazz-heads, but for tracks other than their predominant genres: the dotted-orange line, D_O, shows the distribution of non-Dance tracks downloaded by Dance-heads ($Dance_{Other} = D_O$); the dotted-blue line, J_O, shows the distribution of non-Jazz tracks downloaded by Jazz-heads

TABLE 1 | Descriptive statistics for X-head subgroups: number of users; number of downloads (average downloads per user).

X-head subgroup	Users	Downloads
Pop	3,215,135	74,421,204 (23.14)
Bollywood	2,134,919	30,340,368 (14.21)
Mandarin Pop	1,944,975	15,165,454 (7.80)
Rock	349,205	23,489,799 (67.27)
Dance	163,388	1,676,634 (10.26)
Rap/Hip Hop	156,384	2,122,862 (13.57)
Metal	94,015	4,513,560 (48.01)
Classical	45,903	2,260,496 (49.25)
Jazz	22,644	1,131,941 (50.00)
Reggae	13,530	240,615 (17.63)

[3]Online information can be accessed at the following webpage: https://web.archive.org/web/20150112031805/http://developer.echonest.com/acoustic-attributes.html

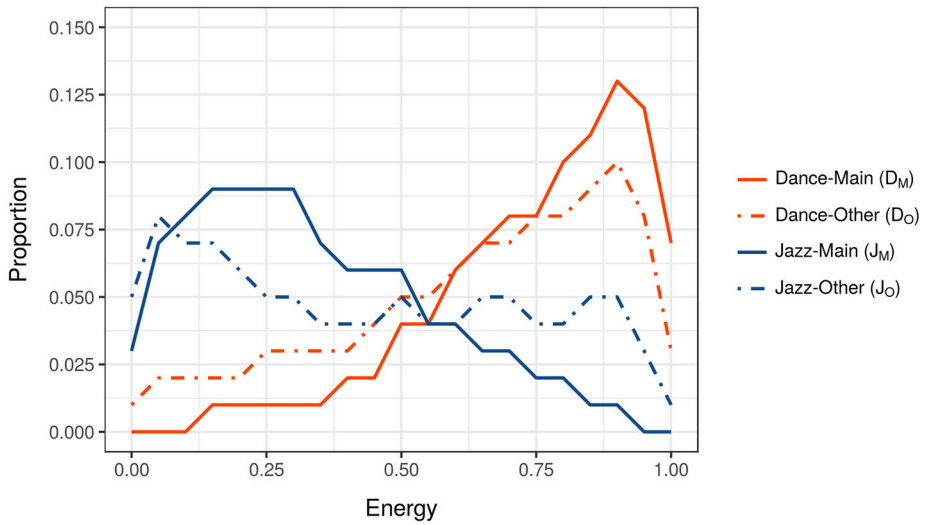


FIGURE 1 | Energy distributions of tracks belonging to Dance-head and Jazz-head subgroups. The orange lines, D_M and D_O, show the distributions of tracks downloaded by Dance-heads; the blue lines, J_M and J_O, show the distributions of tracks downloaded by Jazz-heads.

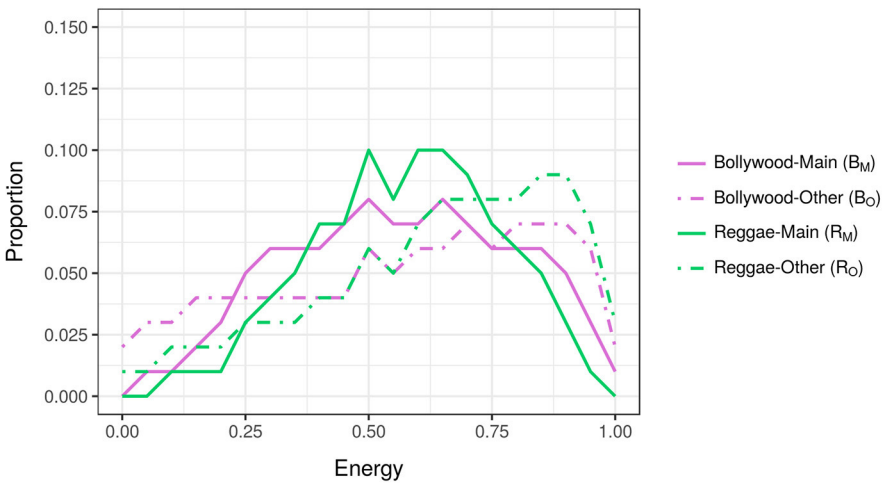


FIGURE 2 | Energy distributions of tracks belonging to Bollywood-head and Reggae-head subgroups. The purple lines, B_M and B_O, show the distributions of tracks downloaded by Bollywood-heads; the green lines, R_M and R_O, show the distributions of tracks downloaded by Reggae-heads.

($Jazz_{Other} = J_O$). Two things are important to note: (1) the Energy distribution of Dance-heads' non-Dance tracks mirrors the distribution of their Dance tracks, e.g., both D_M and D_O peak on the right; and (2) the Energy distribution of Jazz-heads' non-Jazz tracks mirrors the distribution of their Jazz tracks, e.g., both J_M and J_O peak on the left. Which is to say, when Dance-heads download non-Dance tracks, there is a tendency for these tracks to be similar in terms of Energy to Dance tracks. Alternatively put, the generally high Energy of Dance tracks influences the choices Dance-heads make with respect to non-Dance music, while the generally low Energy of Jazz tracks influences the choices Jazz-heads make with respect to non-Jazz music.

The observation above relies upon X-head pairs having dissimilar feature distributions (i.e., lines D_M and J_M), and, in the case of **Figure 1**, the distribution of D_M being closer to D_O than J_O, and J_M being closer to J_O than D_O. If, however, the distributions of the X-heads' main genres are homologous, as is the case for Bollywood- and Reggae-heads in **Figure 2** (solid green and purple lines), then no such pattern of similar/different distributions is possible. Which is to say, distributions where X-heads' main genres are more-or-less similar, are less able to demonstrate acoustic-feature influence.

The distributions for all possible X-heads' main genres were correlated with each other in order to identify pairs

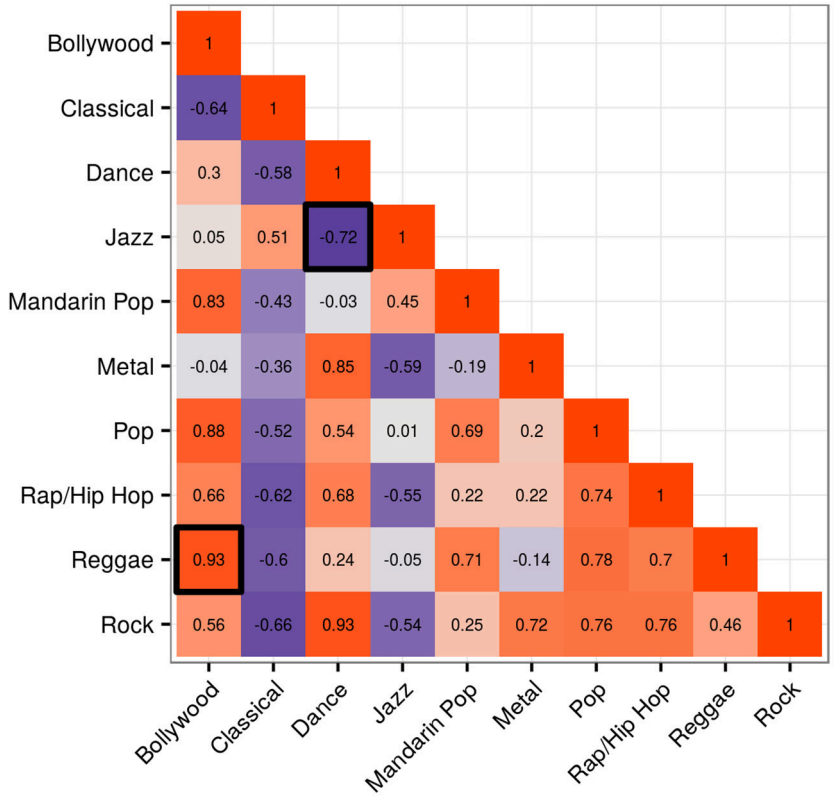


FIGURE 3 | Correlation matrix for Energy showing coefficients between pairs of X-head subgroups. The two highlighted cells relate to the solid lines in **Figures 1**, **2**, Dance and Jazz, and Bollywood and Reggae respectively.

with dissimilar distributions. This was conducted for all 10 features. Using Pearson product-moment correlation, five features yielded no negative coefficients, and were thereby eliminated from the analysis. The five remaining features yielding negative coefficients, usable in the analysis, included Acousticness, Danceability, Energy, Loudness, and Valence. **Figure 3** shows the correlation matrix for feature Energy. The two highlighted cells within the matrix relate to the solid lines in **Figures 1**, **2**, Dance and Jazz, and Bollywood and Reggae respectively. The Dance-Jazz coefficient is negative (reflected in the dissimilar distributions in **Figure 1**); the Bollywood-Reggae coefficient is positive (reflected in the similar distributions in **Figure 2**). In the case of Energy, this process yielded 19 X-head pairs suitable for analysis, i.e., 19 cells with negative coefficients.

Following this, for each X-head pair AB, the distribution of A's main genre (e.g., D_M, **Figure 1**) was correlated with the distribution of A's other music (e.g., D_O). Next, the distribution of A's main genre (e.g., D_M) was correlated with the distribution of B's other music (e.g., J_O). This produced two coefficients. This process was then repeated for B: B's main genre (e.g., J_M) was correlated with the distribution of their other music (e.g., J_O), and the distribution of B's main genre (e.g., J_M) was correlated with the distribution of A's other music (e.g., D_O). A and B together, therefore, produced four coefficients. For each feature, this was repeated for all X-head pairs with negatively correlated distributions, and the resulting coefficients entered into a paired sample *t*-tests in which "within-group" coefficients (e.g., D_M correlated with D_O) were paired with "between-group" coefficients (e.g., D_M correlated with J_O).

Figure 4 illustrates this process for Energy with respect to Dance- and Jazz-heads. In total, the 19 X-head pairs identified in the Energy correlation matrix in **Figure 3** gave rise to a *t*-test into which 38 pairs were entered. This enabled us to observe whether there was a closer relationship between the features of A's main genre and their other music (**Figure 4**, red column; e.g., D_M and D_O) than with the features of B's other music (**Figure 4**, blue column; e.g., D_M and J_O) and vice versa, i.e., whether there was a significant influence of the main genre on music of secondary importance within people's downloads. If there had been no influence, then the distributions of either A or B's other music (e.g., D_O or J_O) would not be expected to show a consistently closer relationship to their respective main genre distributions (e.g., D_M or J_M). The results of this analysis for the five viable features referred to above are now presented.

3.2. Results

As previously described, the presence of negatively correlated distributions, shown in **Figure 3** with respect to Energy,

	Paired sample t-test		
Within X-head subgroup correlation	**Coefficient**	**Coefficient**	**Between X-head subgroup correlation**
Dance-head main (D_M) to Dance-head other (D_O)	.96	-.44	Dance-head main (D_M) to Jazz-head other (J_O)
Jazz-head main (J_M) to Jazz-head other (J_O)	.56	-.72	Jazz-head main (J_M) to Dance-head other (D_O)
.	.	.	.
.	.	.	.
.	.	.	.

FIGURE 4 | Example of paired sample *t*-test with respect to Energy in which within-group coefficients were paired with between-group coefficients. Only the first two from 38 pairs are shown.

enabled the influence of five features to be studied using the present methodology. In sum, Acousticness had 12 negatively correlated distributions, Danceability 10, Energy 19, Loudness 5, and Valence 9 (see **Supplementary Figure 1**). **Figure 5** shows boxplots of the five features within the analysis. The red boxes on the left of each graph represent the within X-head coefficients; blue boxes on the right are the between X-head coefficients. Paired-sample *t*-tests, conducted to compare the within X-head coefficients and between X-head coefficients, showed the following results (sig. 2-tailed):

Acousticness. Significant difference for within ($M = 0.749$, $SD = 0.170$) and between ($M = 0.250$, $SD = 0.302$) X-head coefficients; $t(23) = 11.887$, $p < 0.0001$.

Danceabilty. Significant difference for within ($M = 0.608$, $SD = 0.225$) and between ($M = 0.313$, $SD = 0.288$) X-head coefficients; $t(19) = 8.046$, $p < 0.0001$.

Energy. Significant difference for within ($M = 0.557$, $SD = 0.233$) and between ($M = 0.174$, $SD = 0.414$) X-head coefficients; $t(37) = 9.110$, $p < 0.0001$.

Loudness. Significant difference for within ($M = 0.636$, $SD = 0.270$) and between ($M = 0.315$, $SD = 0.301$) X-head coefficients; $t(9) = 10.656$, $p < 0.0001$.

Valence. Significant difference for within ($M = 0.653$, $SD = 0.113$) and between ($M = 0.223$, $SD = 0.199$) X-head coefficients; $t(17) = 11.887$, $p < 0.0001$.

3.3. Discussion

The statistics above confirm what is clearly evident in the boxplots in **Figure 5**: there is a significant difference in the two sets of coefficients for each feature; in general, coefficients for the within condition are greater than the between condition. This is also true for the feature Loudness, which had only five X-head pairs with negatively correlated distributions (producing 10 pairs of coefficients). In other words, even with a relatively low *n*, there is a statistically closer relationship between the features of an X-head's main genre and their other music than with the features of a different X-head's other music, i.e., a significant influence of the main genre on music of secondary importance within people's downloads.

Despite the results having a clear direction, the current method was not able to test the influence of five of the 10 features within the analysis: Duration, Instrumentalness, Liveness, Speechiness, and Tempo. Although the observed pattern of influence may well extend to these features, this is by no means certain—for cognitive and neurological reasons, this phenomenon may be limited to particular acoustic features; for example, perhaps those that are more closely tied in some way to personality (e.g., McCown et al., 1997). Moreover, Analysis 1 was not able to address whether some X-head subgroups exhibited more influence, or if specific genres were more susceptible to being influenced by other genres more dominant within people's download collections. For example, is it the case that Classical is more prone to the influence of negatively valenced or sad music than, say, Metal? Similarly, might Jazz be more immune to the influence of up-tempo music than Bollywood, and what might be the interaction of X-head subgroup on these processes? Our aim was not necessarily to explain such patterns, which may well involve a combination of personality and sociocultural factors, but rather to observe the degree to which they existed within Nokia DB. To this end, we undertook a second analysis in which detailed information relating to each X-head subgroup and our selected 10 genres was brought to light.

4. ANALYSIS 2: FEATURE-INFLUENCE MATRICES

4.1. Method

Each feature-influence matrix, referred to as C, was calculated from two submatrices, A and B. A, a 10 (X-heads) × 10 (genres) submatrix, contained the average feature values of all songs within a genre downloaded by an X-head subgroup (for example, the average value of Valence for all Reggae tracks downloaded

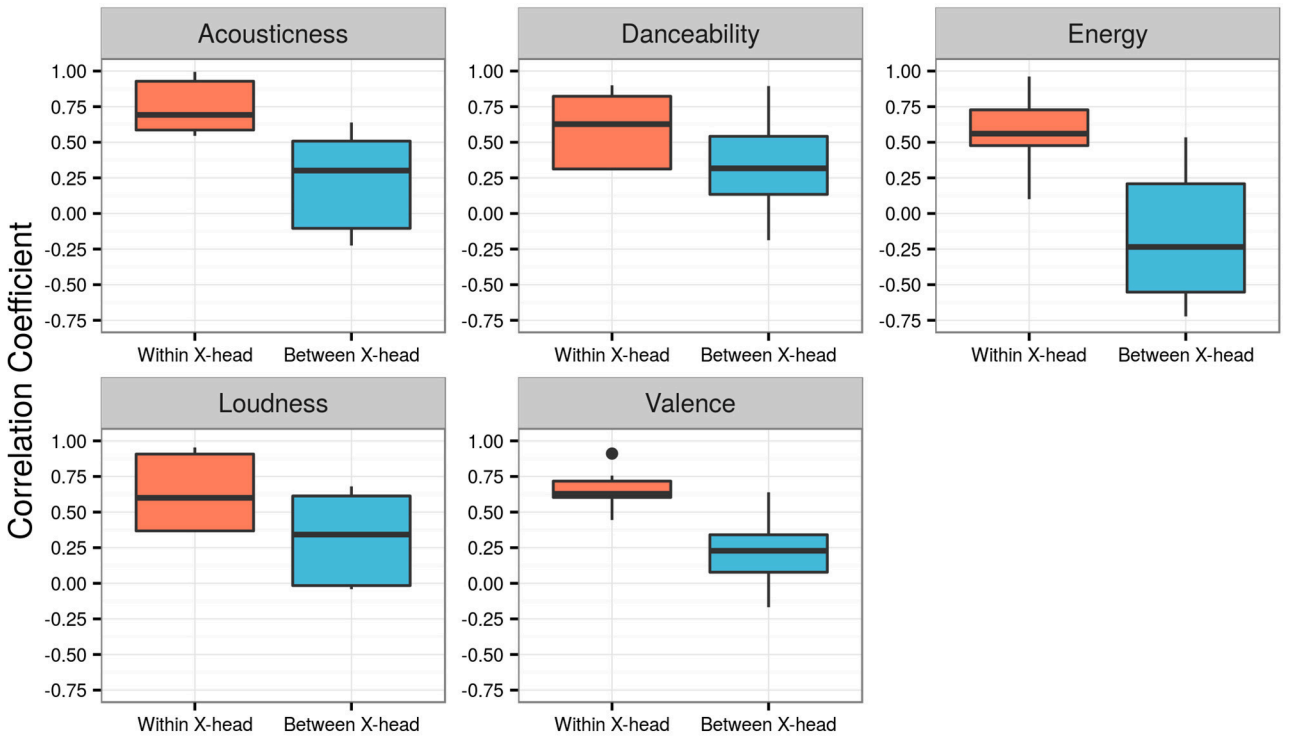


FIGURE 5 | Boxplots of the five features within Analysis 1. Red boxes represent the within X-head coefficients; blue boxes represent the between X-head coefficients.

by Classical-heads). B, a 10 (genres) × 1 (averages) submatrix, contained the average feature value of each genre downloaded by all users, excluding those made by the main X-head subgroup (for example, the average Valence of Metal tracks downloaded by everyone except Metal-heads). C, the 10 (X-heads) × 10 (genres) feature-influence matrix, was calculated by subtracting the row values in A (subtrahends) from those in B (minuends), and converting the resulting differences into percentage changes from the averages in B. Formally, the above is given by:

$$F = \{\forall x_g \in F_{xg}, x_g = \{\frac{F_{xg} - P_g}{P_g} * 100\}\} \quad (1)$$

Where:

F = Feature-influence matrix (Matrix C)

x = X-head subgroup

F_{xg} = Average feature value for genre (g) in X-head (x) subgroup (Submatrix A)

P_g = Average feature value for genre (g) for entire population (Submatrix B)

x_g = Average feature value for genre (g) listened to by X-head subgroup (x)

We illustrated this process with reference to Submatrices A and B, Matrix C (the feature-influence matrix), and feature Valence. For clarity, the calculation is simplified to include only three X-head subgroups and genres: Dance, Metal, and Pop.

TABLE 2 | Example of Submatrix A showing the average Valence of three genres downloaded by three X-head subgroups.

	Classical	Dance	Metal
Classical-heads	0.27	0.43	0.38
Dance-heads	0.28	0.41	0.37
Metal-heads	0.28	0.44	0.35

Rows represent X-heads; columns represent genres.

4.1.1. Submatrix A

Table 2 shows Matrix A: rows (i) represent X-head subgroups; columns (j) represent genres downloaded by each X-head subgroup. For example, average Classical, Dance, and Metal Valence values for Dance-heads ($i = 2, j = (1, 2, 3)$) are $(2, 1) = 0.28, (2, 2) = 0.41$, and $(2, 3) = 0.37$ respectively.

4.1.2. Submatrix B

Table 3 shows Submatrix B: the columns are genres; the row is the average Valence of each genre, excluding members of that particular X-head subgroup. For example, the average Valence for Metal music downloaded by non-Metal-heads $(1, 3) = 0.39$.

4.1.3. Matrix C (Feature-Influence Matrix)

Table 4 shows Matrix C, generated by subtracting cell i, j in Submatrix A from cell i, j in Submatrix B. We take the percentage change for that feature using the population average for a particular genre in Submatrix B (similar results were obtained

TABLE 3 | Example of Submatrix B for feature Valence with three genres.

	Classical	Dance	Metal
Population average	0.28	0.47	0.39

Average feature values of genres, excluding members of each particular X-head subgroup.

TABLE 4 | Example of Matrix C, the feature-influence matrix, showing the percentage Valence change of three genres downloaded by three X-head subgroups.

	Classical (%)	Dance (%)	Metal (%)
Classical-heads	−3.57	−8.51	−2.56
Dance-heads	0.0	−12.8	−5.12
Metal-heads	0.0	−6.1	−10.26

Rows represent X-heads; columns represent genres.

using population medians as opposed to averages). For example, to calculate cell (2, 2) of Matrix C:

$$C_{i,j} = \left(\frac{A(i,j) - B(j)}{B(j)}\right) * 100$$

$$C_{2,2} = \left(\frac{0.41 - 0.47}{0.47}\right) * 100 = -12.8$$

This example indicates that Dance-heads downloaded Dance music that was 12.8% more negatively valenced than the rest of the population downloading Dance.

4.2. Results

Figure 6 shows the feature-influence matrix for Acousticness. Cell values indicate the percentage change in Acousticness of genres (columns) downloaded by X-head subgroups (rows), compared to the average Acousticness of genres downloaded by the overall population. The highlighted diagonal cells (running top left to bottom right) show X-heads with respect to their main genres. The highlighted column on the right shows the median value of each row, excluding diagonally highlighted cells.

Of the 100 possible diagonal-to-median cell pairings (10 features × 10 X-heads), the signs of 64 were in agreement; 36 were in disagreement (see **Supplementary Figure 2**). These pairings are shown in the scatterplot in **Figure 7**. Light-green quadrants indicate sign agreement between the row medians and X-heads with respect to their main genres, either positive or negative; pink quadrants indicate sign disagreement. The adjusted r^2-value, 0.1039, gives rise to the following statistic: $r = 0.34$, $n = 100$, $p < 0.0001$.

The 10 feature-influence matrices enabled two further, complementary questions to be explored. First, across all X-heads, which feature of their main genres most strongly influenced their other genres? For example, is the relationship between X-heads' main and other genres stronger for Energy than Danceability? This question was assessed by correlating X-heads' main genres with the nine other genres in each of their download collections. This produced one overall coefficient per feature-influence matrix; the resulting 10 coefficients were then ranked in order. The second question asked which X-head subgroup across all features had the closest relationship between their main genre and other genres. For example, do the features of Mandarin Pop-heads' main genre more strongly influence the corresponding features of their other genres than is the case for Reggae-heads? This question was investigated by correlating each X-head's main genre with the nine other genres in each feature-influence matrix. This produced one overall coefficient per X-head subgroup, and, as before, the resulting 10 coefficients were ranked in order. The results of these analyses are shown in **Tables 5**, **6**. Rows represent ranks, either of features or X-heads. Also shown are associated Pearson product-moment correlation coefficients.

4.3. Discussion

The finding in the feature-influence matrices that the signs of 64 diagonal-to-median cell pairings were in agreement, with 36 in disagreement, strongly suggests that there is a directional relationship, either positive or negative, between the features of X-heads' main genres and those of their other genres. This is confirmed in the scatterplot in **Figure 7**, and associated correlation statistic ($r = 0.34$), in which there was a significant, positive relationship between the variables. Of course, our assumption is that the direction of influence is from the main to the other genres in each X-head subgroup: intuitively, at least, it seems to make sense that most individuals have a preferred musical style that influences their choices in other genres. However, the converse could be true: the features of X-heads' secondary genres influence the choices they make in their main genre, although this is perhaps less likely.

In the foregoing diagonal-to-median cell analysis, two additional analyses sought to establish ranked orders showing: (1) which feature of X-heads' main genres most strongly influenced their other genres, and (2) which X-head subgroup's main genre most strongly influenced their other genres across all features. In **Table 5**, the top-ranked feature was Speechiness ($r = 0.52$)—the presence or absence of spoken words in tracks belonging to main genres appears to have created a preference for similarly "speechy" tracks in X-heads' other genres. Similarly, the Danceability, Loudness, and Energy of users' predominant tracks appear to have heavily influenced tracks of secondary importance. At the other end of the spectrum, there was little-to-no relationship between X-heads' main and other genres in terms of Liveness (whether a track was recorded at a live event) and Instrumentalness (whether a track was created using only instrumental sounds).

The top-ranked X-head subgroups were Metal ($r = 0.56$) and Jazz ($r = 0.49$). The dynamic nature of much Metal music seems to have created a musical 'fingerprint' that strongly expressed itself in the other genres Metal-heads downloaded. Likewise, Jazz-heads seem compelled to seek out music containing Jazz-like qualities when exploring non-Jazz music. Conversely, Mandarin

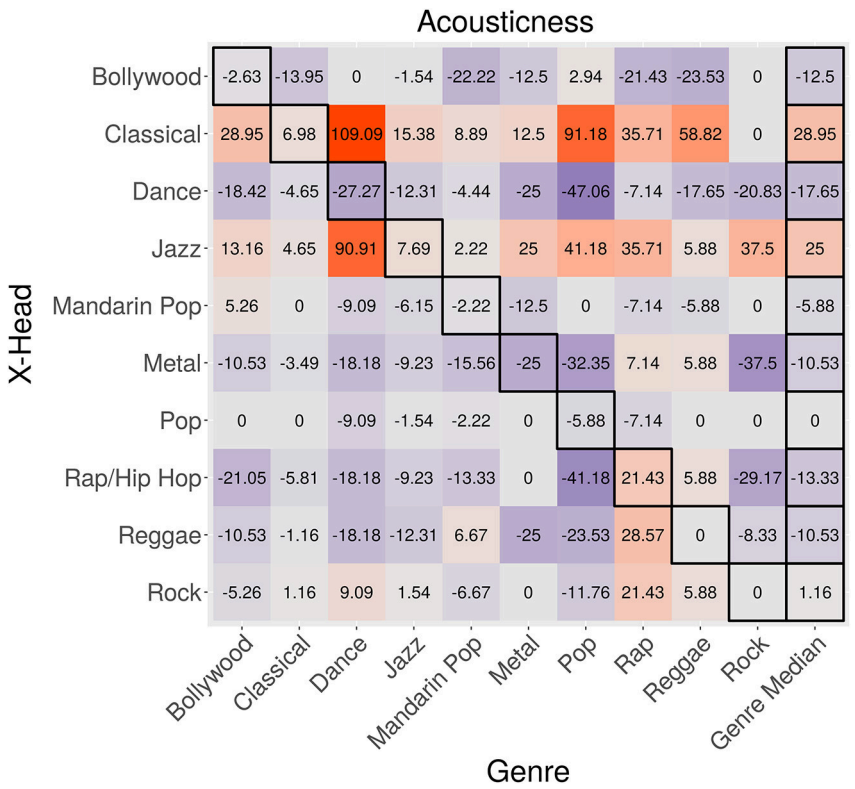


FIGURE 6 | Feature-influence matrix for Acousticness. The highlighted diagonal cells (running top-left to bottom right) show X-heads with respect to their main genres. The highlighted column on the right shows the median value of each row, excluding diagonally highlighted cells.

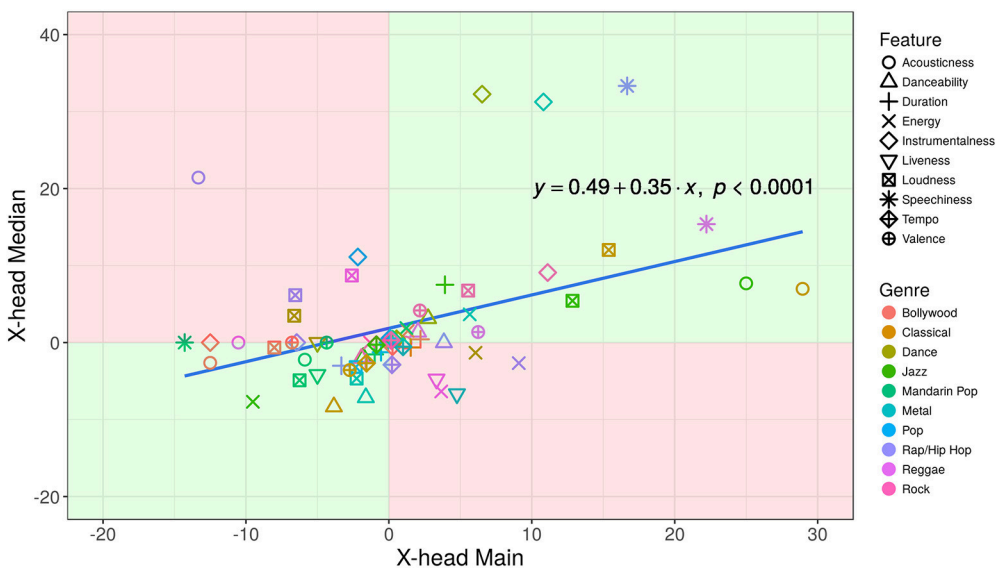


FIGURE 7 | Scatterplot showing the 100 diagonal-to-median cell pairings of the 10 feature-influence matrices. Light-green quadrants indicate sign agreement between the row medians and X-heads with respect to their main genres, either positive or negative; pink quadrants indicate sign disagreement.

Pop and Pop's musical features did not significantly influence the other music downloaded by these X-head subgroups, perhaps because the features of these genre are relatively indistinct ($r = 0.09$ and $r = 0.08$ respectively). These findings, and those relating to Analysis 1, are now discussed in the broader context of the paper.

TABLE 5 | Table showing the ranked order of features.

Rank	Feature ($n = 90$)
1	Speechiness; $r = 0.52^{***}$
2	Danceability; $r = 0.48^{***}$
3	Loudness; $r = 0.45^{***}$
4	Energy; $r = 0.44^{***}$
5	Acousticness; $r = 0.28^{**}$
6	Tempo; $r = 0.19$
7	Duration; $r = 0.17$
8	Valence; $r = 0.14$
9	Liveness; $r = 0.06$
10	Instrumentalness; $r = 0.04$

The feature column shows which feature of X-heads' main genres is closest to their other genres ($p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$).

TABLE 6 | Table showing the ranked order of X-heads.

Rank	X-head ($n = 90$)
1	Metal; $r = 0.56^{***}$
2	Jazz; $r = 0.49^{***}$
3	Dance; $r = 0.45^{***}$
4	Classical; $r = 0.41^{***}$
5	Rap; $r = 0.28^{**}$
6	Rock; $r = 0.26^{*}$
7	Bollywood; $r = 0.14$
8	Reggae; $r = 0.11$
9	Mandarin Pop; $r = 0.09$
10	Pop; $r = 0.08$

The X-head column shows the subgroup with the closest relationship between their main genre and other genres, across all features ($p < 0.1$; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$).

5. GENERAL DISCUSSION

Using pairs of X-head subgroups whose feature distributions negatively correlated, Analysis 1 found that there was a consistent relationship between X-heads' main and other genres; the methodology enabled five acoustic features to be investigated. Analysis 2 added detail to this finding through the production of 10 feature-influence matrices; there was a significant positive correlation between the diagonal-to-median cell pairings across the matrices, again strongly indicating that there is a relationship between the features of X-heads' main genres and those of their other genres. Therefore, with respect to the question posed at the outset, the core findings of this paper support the proposition that the acoustic features of a person's main musical genre influence their choices within non-preferred, secondary styles. Which is to say, attributes of the tracks within preferred genres influence the other genres of people's music-download collections. The nature of this influence, and its possible actuating mechanisms, form the major part of the following discussion.

Although, as outlined in Section 1, substantial research has been undertaken in relation to musical preference and personality, for a variety of reasons relatively few studies have explored this issue using large music-consumption databases, such as Nokia DB. First, in terms of usual research timescales (i.e., years, not months), APIs, through which large volumes of data become accessible to external researchers, are relative newcomers to the academic landscape. Second, API rate-limits typically restrict the amount of data that is available; similarly, database limits may constrain the type of information that a researcher is able to export. And third, for sound methodological reasons relating to data integrity, psychologists have tended to rely on relatively small subject pools to whom individual personality or self-image tests can be administered (e.g., Zweigenhaft, 2008; Krause and Hargreaves, 2013).

In seeking to corroborate the findings of preexisting music-personality studies, Bansal and Woolhouse (2015), using Nokia DB, investigated (1) whether X-head subgroups showed distinct patterns of genre exclusivity, and, if so, (2) whether genre exclusivity related to the Big Five personality factors (Costa and MacCrae, 1992). X-heads ranked from exclusive to inclusive were as follows: Pop, Dance, Rap, Metal, Rock, Classical, Country, Folk, Jazz, and Indie. Interestingly, this aligned with previous literature showing that individuals who prefer Jazz and Folk score highly in the Big Five factor of openness (Zweigenhaft, 2008). Those high in openness were also found to avoid genres like Pop; likewise, Bansal and Woolhouse (2015) determined Pop-heads to be the most genre exclusive. In sum, genre-openness (and –agreeableness) associations from Zweigenhaft (2008) predicted genre inclusivity in Nokia DB X-heads—individuals with high openness scores (and/or agreeableness) were more likely to have a wider selection of genres within their music collections. Bansal and Woolhouse (2015) did not find conscientiousness, extraversion, or neuroticism to be predictors of genre exclusivity.

By demonstrating that personality-related behavior is discernable within big data concerning music consumption, Bansal and Woolhouse's (2015) research is highly relevant to the current study. If personality can be shown to have guided genre exclusiveness, then its involvement in other aspects of people's musical choices is not only possible, but, arguably, probable. In the present instance, a mechanism is being sought that may account for influence in terms of acoustic features and genres of secondary importance within X-heads' downloads. McCown et al. (1997) linked exaggerated bass frequencies, i.e., a specific acoustic feature, to a particular personality factor, neuroticism—it would seem self-evident that other acoustic features, including those explored within our study, will likewise be linked to aspects of personality, and therefore expressed *throughout* individuals' music collections. For example, in **Table 6** Dance-heads are the third most influence-exhibiting subgroup ($r = 0.45$), indicating that the feature values of their main genre were significantly related to those of their other genres. Similarly, Danceability, a feature strongly associated with Dance-heads (see **Figure 1**), also ranks highly in **Table 5** ($r = 0.48$). Given the work of McCown et al. (1997) linking Dance with neuroticism, it is tempting to conjecture that the increased feature influence of Dance-heads and Danceability is in someway related to heightened obsessiveness, a trait strongly associated with neurotic tendencies (Samuels et al., 2000).

However, although intriguing, this proposition is beyond the scope of the present study, and thus awaits further investigation.

Alongside our fledgling personality hypothesis, expounded above, the work of Berns et al. (2010) and Halko and Kaustia (2015), discussed briefly in Section 1, is suggestive of neurophysiological mechanisms underpinning musical-feature influence. Specifically, we address this issue with reference to Aniruddh Patel's research involving music, language, and statistical learning—the ability of humans and other animals to acquire implicit knowledge about the world through the extraction of statistical regularities within their environments (Friedman et al., 2001; for neurological evidence concerning statistical learning of language, see Cheour et al., 1998; Rivera-Gaxiola et al., 2005). In order to account for the finding that the prosodies of English and French are reflected in the rhythms and melodies of these countries' respective instrumental music, Patel proposes a "direct-route" hypothesis, in which "statistical learning of the prosodic patterns of speech creates implicit knowledge of rhythmic and melodic patterns in language, which can in turn influence the creation of rhythmic and tonal patterns in music" (Patel et al., 2006, p. 3043). In other words, statistically acquired sound-pattern knowledge "leaks" from the domain of language, resulting in the rhythmic and melodic modification of music. Typically assessed using the Normalized Pairwise Variability Index (nPVI), a technique that measures the degree of durational contrast between successive elements in a sequence, research demonstrating this phenomenon is both robust and compelling (e.g., Huron and Ollen, 2003; Patel and Daniele, 2003; Daniele and Patel, 2004).

Patel's work is highlighted here by way of analogy—the phenomenon of musical-feature influence is limited to music, and therefore is not a cross-domain effect. However, statistical learning may well be pertinent to our findings, and suggests the existence of a mechanism that is more or less independent of personality (to our knowledge, no research has linked personality factors with abilities in statistical learning). Given empirical evidence of temporal and intervallic relationships between music and language, and Patel's assertion that this is underpinned by statistical learning and hence causal in nature, it is plausible to suggest that a similar process operates with respect to musical features. That is, listeners extract the statistical regularities of musical features, which in turn influence the creation of musical preferences beyond established style boundaries and/or genre categories. Statistical regularities of features may be relatively straightforward, such as Tempo—the speed of the most salient pulse in the music, usually measured by allowing listeners to tap along to perceptually noticeable beats (McKinney and Moelants, 2006)—or complex, such as Danceability—an amalgamation of tempo, rhythm stability, beat strength, and isochrony.

In **Table 5**, the effect of Tempo on secondary genres was only marginally significant, whereas Danceability was highly significant. If statistical learning is at play, this finding suggests that its effect is bolstered through the presence of multiple, mutually reinforcing acoustic components, as is the case for Danceability. Arguably less plausible, however, is the notion that statistical learning affects X-head subgroups differentially. If this were the case, the rankings in **Table 6** would indicate that Metal-heads, who are at the top of the table, engage in statistical learning, whereas Pop-heads, at the bottom, do not. While this seems unlikely, it might be that Metal is acoustically more regular than Pop, and therefore facilitates statistical learning to a greater degree; although, given the high level of signal redundancy in much Pop music, this hypothesis would seem to be doubtful.

5.1. Limitations

In presenting our findings we have attempted to develop and adapt a range of approaches, suitable to the data at hand. And while the premise of the question motivating our research is supported by a series of cogent results, the adopted methodologies, as well as the data themselves, are limited in a variety of ways and raise a number of questions.

First, the algorithms responsible for Spotify's acoustic features are proprietary, and therefore not publically available. As a result, although our primary aim was to investigate and record the presence of musical-feature influence, we were unable to assess in detail which specific acoustic elements were responsible for our findings. Which frequency bands within an X-head's main genre, for example, have in general a greater influence on their other genres? Which components of Acousticness are present throughout an X-head's download collection, and which are specific to their main genre? Moreover, and perhaps of greater import, as mentioned at the outset, the psychological reality of acoustic features is, as yet, unquantified (Friberg and Schoonderwaldt, 2014). Although a feature like Valence may make sense to those who know and love music's emotional power, its interpretation across listeners may be highly divergent. Valence is frequently characterized with reference to mode, either major (positive/happy) or minor (negative/sad) (Kastner and Crowder, 1990). However, those familiar with works such as Elgar's "Nimrod" (*Enigma Variations, Op. 36*), which although in a major key is deeply poignant, may take a very different view of this dichotomy.

Second, no attempt has been made to address the issue of mood, referred to in Section 1. As discussed, in contrast to the stability of personality, mood is thought to change relatively rapidly (McFarlane et al., 1988). Our analyses did not take into account temporal order or download timelines, which may have revealed day-to-day effects of mood. For example, an important question might be, do downloads oscillate between negative and positive Valence, and, if so, is the influence of the upswing to positive different from the downswing? Although this question is beyond the scope of the present study, and would no doubt require very different methodologies to those used here (e.g., time-series analysis), the Nokia DB does contain detailed date/time information that would, in theory, enable this matter to be addressed.

Third, as mentioned in Section 2, our intention was to define X-heads straightforwardly, i.e., a majority of downloads in a particular genre. While this simple metric has the advantage of transparency—X-heads are not cooked up using a complicated, opaque recipe—the approach will undoubtedly have created a class of users with overlapping, ill-defined boundaries, which could have introduced undue noise into the analyses. In this respect, no attempt was made to separate "Super-heads," e.g.,

users in the upper quartile in terms of main-genre proportion, from "Weak-heads," e.g., users in the lower quartile. And, consequently, some users in different X-head subgroups may have been similar. For example, consider two users, P and Q, with the following download proportions: P = 55% Jazz, 45% Classical; Q = 55% Classical, 45% Jazz. Despite P and Q having a great deal in common, our method would group them as categorically distinct: P a Jazz-head, Q a Classical-head. The question then arises as to whether feature influence is more accentuated in Super-heads vs. Weak-heads (which we would imagine to be the case), or whether no such effect exists. While the downside of our simple X-head definition was that this issue could not be addressed, the upside is that the data within Nokia DB, with a little preprocessing, affords us the opportunity to answer this question in detail in the future.

5.2. Closing Remarks

In summary, Analyses 1 and 2 found strong evidence of influence with respect to users' consumption of multiple styles of music; clear relationships emerged between the features of X-heads' main and secondary genres. This effect was found to be stronger for some features than others, most noticeably Speechiness, Danceability, and Loudness, and more pronounced in certain subgroups, such as Metal-heads, Jazz-heads, and Dance-heads. While the reasons for differential effects within features and X-heads is unknown, two probable, independent causal mechanisms were suggested to account for main-to-secondary genre influence. First, personality creates an overarching psychological framework in which certain factors, such as openness and agreeableness, guide musical preference, irrespective of genre; some personality factors may be linked to specific acoustic features. Second, via statistical learning, listeners extract the acoustic regularities of various musical features, which in turn influence the creation of musical preferences beyond favored styles and/or genres. Of course, these mechanisms need not be mutually exclusive, but may serve to reinforce one another. Attempts, therefore, to tease apart the effects of personality and statistical learning could prove to be difficult, although paradigms in which these factors are independently manipulated might settle the issue of personality vs. statistical learning conclusively.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the Research Ethics Board of McMaster University, Canada. The protocol was approved by the McMaster Research Ethics Committee.

AUTHOR CONTRIBUTIONS

MB: Study design and execution, data analysis and interpretation, figure and graph creation, and manuscript review. JB: Study design and manuscript review. MW: Manuscript drafting, study design, data analysis and interpretation.

FUNDING

This research is generously supported by a Partnership Development Grant (#890-2014-0126) from the Social Sciences and Humanities Research Council of Canada, awarded to MW.

ACKNOWLEDGMENTS

The authors would like to thank the following people who have supported the research presented in this paper in a variety of ways: Dora Rosati, Kurt DaCosta, Nick Rogers, and Mark Hahn of SHARCNET/Compute Canada and Research & High-Performance Computing Support, McMaster University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2017.00931/full#supplementary-material

Supplementary Figure 1 | Correlation matrices for all acoustic features (Acousticness, Danceability, Duration, Energy, Instrumentalness, Liveness, Loudness, Speechiness, Tempo, Valence), as described in Section 3.1, **Figure 3**.

Supplementary Figure 2 | Feature-influence matrices for all acoustic features (Acousticness, Danceability, Duration, Energy, Instrumentalness, Liveness, Loudness, Speechiness, Tempo, Valence), as described in Section 4.1, **Figure 6**.

REFERENCES

Bansal, J., and Woolhouse, M. (2015). "Predictive power of personality on music-genre exclusivity," in *Proceedings of the International Society for Music Information Retrieval* (Málaga), 652–658.

Berns, G. S., Capra, C. M., Moore, S., and Noussair, C. (2010). Neural mechanisms of the influence of popularity on adolescent ratings of music. *Neuroimage* 49, 2687–2696. doi: 10.1016/j.neuroimage.2009.10.070

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011). "The million song dataset," in *Proceedings of the International Society for Music Information Retrieval*, *Vol. 2* (Miami, FL), 10.

Butt, S., and Phillips, J. G. (2008). Personality and self reported mobile phone use. *Comput. Hum. Behav.* 24, 346–360. doi: 10.1016/j.chb.2007.01.019

Cheour, M., Ceponiene, R., Lehtokoski, A., Luuk, A., Allik, J., Alho, K., et al. (1998). Development of language-specific phoneme representations in the infant brain. *Nat. Neurosci.* 1, 351–353. doi: 10.1038/1561

Christenson, P. G., and Peterson, J. B. (1988). Genre and gender in the structure of music preferences. *Commun. Res.* 15, 282–301. doi: 10.1177/009365088015003004

Coopersmith, S. (1959). A method for determining types of self-esteem. *J. Abnorm. Soc. Psychol.* 59:87.

Costa, P. T., and MacCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI): Professional Manual*. Psychological Assessment Resources.

Daniele, J. R., and Patel, A. D. (2004). "The interplay of linguistic and historical influences on musical rhythm in different cultures," in *Proceedings of the International Conference on Music Perception and Cognition* (Evanston, IL), 3–7.

Finn, S. (1997). Origins of media exposure: linking personality traits to tv, radio, print, and film use. *Commun. Res.* 24, 507–529. doi: 10.1177/009365097024005003

Forgas, J. P. (1995). Mood and judgment: the affect infusion model (AIM). *Psychol. Bull.* 117:39. doi: 10.1037/0033-2909.117.1.39

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*, Vol. 1. Berlin: Springer.

Friedman, R. S., Gordis, E., and Förster, J. (2012). Re-exploring the influence of sad mood on music preference. *Media Psychol.* 15, 249–266. doi: 10.1080/15213269.2012.693812

Friberg, A., and Schoonderwaldt, E. (2014). Using perceptu- ally defined music features in music information retrieval. *arXiv preprint arXiv*, 1–39.

Groff, J. R., and Weinberg, P. N. (2002). *SQL: The Complete Reference*, Vol. 2. Osborne: McGraw-Hill.

Halko, M.-L., and Kaustia, M. (2015). Risk on/risk off: risk-taking varies with subjectively preferred and disliked music. *PLoS ONE* 10:e0135436. doi: 10.1371/journal.pone.0135436

Halko, M.-L., Mäkelä, T., Nummenmaa, L., Hlushchuk, Y., and Schürmann, M. (2015). Hedonic context modulates risky choices and reward responses in amygdala and dorsal striatum. *J. Neurosci. Psychol. Econ.* 8:100. doi: 10.1037/npe0000036

Haugtvedt, C. P., Petty, R. E., and Cacioppo, J. T. (1992). Need for cognition and advertising: understanding the role of personality variables in consumer behavior. *J. Consum. Psychol.* 1, 239–260. doi: 10.1016/S1057-7408(08)80038-1

Hughes, D. J., Rowe, M., Batey, M., and Lee, A. (2012). A tale of two sites: twitter vs. Facebook and the personality predictors of social media usage. *Comput. Hum. Behav.* 28, 561–569. doi: 10.1016/j.chb.2011.11.001

Huron, D., and Ollen, J. (2003). Agogic contrast in French and English themes: further support for patel and daniele (2003). *Music Percept.* 21, 267–271. doi: 10.1525/mp.2003.21.2.267

Jehan, T., and DesRoches, D. (2011). *The Echo Nest Analyzer Documentation*. Available online at: https://web.archive.org/web/20150112031755/http://developer.echonest.com/docs/v4

Kastner, M. P., and Crowder, R. G. (1990). Perception of the major/minor distinction: IV. Emotional connotations in young children. *Music Percept.* 8, 189–201. doi: 10.2307/40285496

Kauder, E. (2015). *History of Marginal Utility Theory*. Princeton, NJ: Princeton University Press.

Kim, D., and Areni, C. S. (1993). The influence of background music on shopping behavior: Classical versus top-forty music in a wine store. *Adv. Consum. Res.* 20, 336–340.

Krause, A. E., and Hargreaves, D. J. (2013). myTunes: digital music library users and their self-images. *Psychol. Music* 41, 531–544. doi: 10.1177/0305735612440612

Lartillot, O., and Toiviainen, P. (2007). "A MATLAB toolbox for musical feature extraction from audio," in *Proceedings of the International Conference on Digital Audio Effects* (Bordeaux), 237–244.

LeBlanc, A., Jin, Y. C., Stamou, L., and McCrary, J. (1999). "Effect of age, country, and gender on music listening preferences," in *Bulletin of the Council for Research in Music Education*, eds J. Geringer, M. Kalmar, O. DeJesus, R. Walker, G. Welch, and D. Hargreaves (Magaliesberg: University of Illinois Press), 72–76.

LeBlanc, A., Sims, W. L., Siivola, C., and Obert, M. (1996). Music style preferences of different age listeners. *J. Res. Music Educ.* 44, 49–59. doi: 10.2307/3345413

Lemburg, M.-A. (2008). Python database API specification v2.0. *Python Enhanc. Propos.* 249. Available online at: https://www.python.org/dev/peps/pep-0249/

Leon, G. R., Gillum, B., Gillum, R., and Gouze, M. (1979). Personality stability and change over a 30-year period—middle age to old age. *J. Consult. Clin. Psychol.* 47, 517–524.

McCown, W., Keiser, R., Mulhearn, S., and Williamson, D. (1997). The role of personality and gender in preference for exaggerated bass in music. *Pers. Indiv. Differ.* 23, 543–547. doi: 10.1016/S0191-8869(97)00085-8

McFarlane, J., Martin, C. L., and Williams, T. M. (1988). Mood fluctuations. *Psychol. Women Q.* 12, 201–223. doi: 10.1111/j.1471-6402.1988.tb00937.x

McKay, C. (2004). *Automatic Genre Classification of MIDI Recordings*. Ph.D., thesis, McGill University.

McKinney, M. F., and Moelants, D. (2006). Ambiguity in tempo perception: what draws listeners to different metrical levels? *Music Percept.* 24, 155–166. doi: 10.1525/mp.2006.24.2.155

Moore, K., and McElroy, J. C. (2012). The influence of personality on Facebook usage, wall postings, and regret. *Comput. Hum. Behav.* 28, 267–274. doi: 10.1016/j.chb.2011.09.009

North, A. C., and Davidson, J. W. (2013). Musical taste, employment, education, and global region. *Scand. J. Psychol.* 54, 432–441. doi: 10.1111/sjop.12065

North, A. C., Hargreaves, D. J., and McKendrick, J. (1999). The influence of in-store music on wine selections. *J. Appl. Psychol.* 84:271. doi: 10.1037/0021-9010.84.2.271

Patel, A. D., and Daniele, J. R. (2003). An empirical comparison of rhythm in language and music. *Cognition* 87, B35–B45. doi: 10.1016/S0010-0277(02)00187-7

Patel, A. D., Iversen, J. R., and Rosenberg, J. C. (2006). Comparing the rhythm and melody of speech and music: the case of British English and French. *J. Acoust. Soc. Am.* 119, 3034–3047. doi: 10.1121/1.2179657

Peoples, G. (2015). *While Radio Still Reigns, Concerts Are an Important Source of Music Discovery, Says New Report*. Availabe online at: www.billboard.com/articles/business/6699699/while-radio-still-reigns-concerts-are-an-important-source-of-music

Peterson, R. A., and Kern, R. M. (1996). Changing highbrow taste: from snob to omnivore. *Am. Sociolo. Rev.* 5, 900–907.

Rawlings, D., and Ciancarelli, V. (1997). Music preference and the five-factor model of the NEO personality inventory. *Psychol. Music* 25, 120–132. doi: 10.1177/0305735697252003

Rentfrow, P. J., and Gosling, S. D. (2003). The do re mi's of everyday life: the structure and personality correlates of music preferences. *J. Pers. Soc. Psychol.* 84, 1236–1256. doi: 10.1037/0022-3514.84.6.1236

Rieskamp, J., Busemeyer, J. R., and Mellers, B. A. (2006). Extending the bounds of rationality: evidence and theories of preferential choice. *J. Econ. Liter.* 44, 631–661. doi: 10.1257/jel.44.3.631

Rivera-Gaxiola, M., Klarman, L., Garcia-Sierra, A., and Kuhl, P. K. (2005). Neural patterns to speech and vocabulary growth in american infants. *NeuroReport* 16, 495–498. doi: 10.1097/00001756-200504040-00015

Roberts, D. F., and Henriksen, L. (1990). *Music Listening vs. Television Viewing among Older Adolescents*. Dublin: International Communication Association.

Ross, C., Orr, E. S., Sisic, M., Arseneault, J. M., Simmering, M. G., and Orr, R. R. (2009). Personality and motivations associated with facebook use. *Comput. Hum. Behav.* 25, 578–586. doi: 10.1016/j.chb.2008.12.024

Russom, P. (2011). *Big Data Analytics. TDWI Best Practices Report*. Fourth Quarter, 1–35.

Samuels, J., Nestadt, G., Bienvenu, O. J., Costa, P. T., Riddle, M. A., Liang, K.-Y., et al. (2000). Personality disorders and normal personality dimensions in obsessive—compulsive disorder. *Br. J. Psychiatry* 177, 457–462. doi: 10.1192/bjp.177.5.457

Schäfer, T., and Sedlmeier, P. (2009). From the functions of music to music preference. *Psychol. Music* 37, 279–300. doi: 10.1177/0305735608097247

Västfjäll, D. (2002). Emotion induction through music: a review of the musical mood induction procedure. *Musicae Sci.* 5, 173–211. doi: 10.1177/10298649020050S107

Verplanken, B. (1993). Need for cognition and external information search: responses to time pressure during decision-making. *J. Res. Pers.* 27, 238–252. doi: 10.1006/jrpe.1993.1017

Woolhouse, M., and Bansal, J. (2013). Work, rest and (press) play: music consumption as an indicator of human economic development. *J. Interdiscip. Music Stud.* 7, 45–71. doi: 10.4407/jims.2015.05.003

Woolhouse, M., and Renwick, J. (2016). Generalizing case-based analyses in the study of global music consumption. *Digit. Stud.* Available online at: http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/312/397

Woolhouse, M., Renwick, J., and Tidhar, D. (2014). Every track you take: analysing the dynamics of song and genre reception through music downloading. *Digit. Stud.* Available online at: http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/266/321

Zweigenhaft, R. L. (2008). A do re mi encore: a closer look at the personality correlates of music preferences. *J. Indiv. Differ.* 29, 45–55. doi: 10.1027/1614-0001.29.1.45

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 March 2017
doi: 10.3389/fpsyg.2017.00416

Characterizing Listener Engagement with Popular Songs Using Large-Scale Music Discovery Data

Blair Kaneshiro[1,2]*, *Feng Ruan*[3], *Casey W. Baker*[2] and *Jonathan Berger*[1]

[1] Center for Computer Research in Music and Acoustics, Stanford University, Stanford, CA, USA, [2] Shazam Entertainment, Ltd., Redwood City, CA, USA, [3] Department of Statistics, Stanford University, Stanford, CA, USA

Music discovery in everyday situations has been facilitated in recent years by audio content recognition services such as Shazam. The widespread use of such services has produced a wealth of user data, specifying where and when a global audience takes action to learn more about music playing around them. Here, we analyze a large collection of Shazam queries of popular songs to study the relationship between the timing of queries and corresponding musical content. Our results reveal that the distribution of queries varies over the course of a song, and that salient musical events drive an increase in queries during a song. Furthermore, we find that the distribution of queries at the time of a song's release differs from the distribution following a song's peak and subsequent decline in popularity, possibly reflecting an evolution of user intent over the "life cycle" of a song. Finally, we derive insights into the data size needed to achieve consistent query distributions for individual songs. The combined findings of this study suggest that music discovery behavior, and other facets of the human experience of music, can be studied quantitatively using large-scale industrial data.

Keywords: Shazam, popular music, music discovery, multimedia search, music information retrieval, musical engagement, social media

OPEN ACCESS

Edited by:
Geraint A. Wiggins,
Queen Mary University of London, UK

Reviewed by:
Dipanjan Roy,
Allahabad University, India
Lin Guo,
University of Pennsylvania, USA

***Correspondence:**
Blair Kaneshiro
blairbo@ccrma.stanford.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *24 October 2017*
Accepted: *06 March 2017*
Published: *23 March 2017*

Citation:
Kaneshiro B, Ruan F, Baker CW and Berger J (2017) Characterizing Listener Engagement with Popular Songs Using Large-Scale Music Discovery Data. Front. Psychol. 8:416. doi: 10.3389/fpsyg.2017.00416

1. INTRODUCTION

Discovering new music is a popular pastime, and opportunities for music discovery present themselves throughout everyday life. However, relatively little is known about this behavior and what drives it. In a recent interview study, Laplante and Downie (2011) found that the active, deliberate search for music information—whether finding new music or information about music—is generally considered both useful and intrinsically enjoyable. In an earlier diary study, however, Cunningham et al. (2007) report that the majority of exposures to new music occur in passive encounters—that is, when a listener was not actively seeking to discover new music. Furthermore, while participants in that study reacted positively to over 60% of their encounters with new music, they also reported that passive music encounters were difficult to act upon in the moment. Since the publication of that study, the rise of mobile services and ubiquitous internet now facilitate instantaneous music discovery during everyday life, whether music is actively sought or passively encountered. Accompanying the widespread use of such services is an unprecedented volume of user data bearing potential insights into where and when people discover music, as

well as what music they choose to discover. These data surpass what can be collected through controlled laboratory or ethnographic studies in terms of size, scope, and ecological validity.

In recent years, industrial user data reflecting a variety of musical behaviors—including but not limited to social sharing, consumption, and information seeking—have been utilized in music informatics research. Twitter, being freely available for aggregation, currently serves as the most common source of data and has been used to explore a variety of topics including artist and music similarity (Schedl, 2010; Schedl et al., 2014), music recommendation (Zangerle et al., 2012; Pichl et al., 2014, 2015), geographical attributes of music consumption (Schedl, 2013; Moore et al., 2014), and hit prediction (Kim et al., 2014; Zangerle et al., 2016). Music consumption and sharing has also been approached using Spotify URLs shared via Twitter (Pichl et al., 2014, 2015) and music download data from the MixRadio database (Bansal and Woolhouse, 2015). A number of these studies have contributed or made use of publicly available research corpuses, including the Million Musical Tweets Dataset, containing temporal and geographical information linked to music-related tweets (Hauger et al., 2013); the continually updated #nowplaying dataset of music-related tweets (Zangerle et al., 2014); and Gracenote's GNMID14 dataset, which includes annotated music identification matches (Summers et al., 2016).

In the present study, we explore large-scale music discovery behavior using query data from the audio identification service Shazam[1]. In particular, we investigate whether the timing of audio identification queries within a song can be related back to specific musical events. We aggregate and analyze a large collection of Shazam query *offsets*—that moment in a song when a user initiates a query—over a set of massively popular songs. We first verify that the distribution of query offsets is not uniform but in fact varies over the course of a song. Next, we show that the overall shape of a query offset histogram also varies over the "life cycle" of a hit song, with more queries occurring toward the start of a song once the song has achieved widespread popularity. We then demonstrate that salient musical events—such as the start of a song, onset of vocals, and start of first chorus—are followed by a rise in query activity. We conclude with an assessment of histogram consistency as a function of data size in order to determine what constitutes a sufficient data size for this type of analysis. The findings from this study provide first insights into the types of musical events that engage listeners at a large scale, compelling them to take action to obtain more information about a piece of music. To our knowledge, this is the first time that engagement with specific musical events has been studied with an ecologically valid, large-scale dataset. Findings from this study will advance knowledge of consumption of popular music, information seeking about music, and—more broadly—how and when a large audience chooses to engage with music in their environment. Finally, to promote further research on music discovery, the dataset of over 188 million Shazam queries analyzed in this study is made publicly available.

[1]http://www.shazam.com.

2. MATERIALS AND METHODS

2.1. Audio Content Recognition with Shazam

Shazam is a service that returns the identity of a prerecorded audio excerpt—usually a song—in response to a user query. Over 20 million Shazam queries are performed each day by more than 100 million monthly users worldwide; incoming queries are matched over a deduplicated catalog comprising over 30 million audio tracks. Shazam's audio recognition algorithm is based on fast combinatorial hashing of spectrogram peaks, and was developed with real-world use cases in mind. As a result, Shazam's performance is robust to noise and distortion; provides fast performance over a large database of music; and offers a high recognition (true-positive) rate with a low false-positive rate (Wang, 2003).

Shazam queries typically involve a single button press once the application is loaded. For queries initiated from mobile devices,[2] the user loads the Shazam application and pushes a prominently displayed Shazam icon on the main screen (**Figure 1**, left). The ambient acoustical signal is recorded through the device microphone, converted to an audio fingerprint, and matched. If the query is matched successfully, the match result is then displayed on the device screen. In the most common use case of song identification, the application will return a variety of metadata (**Figure 1**, right) including song title and artist; total number of Shazam queries for the track identifier ("trackid") corresponding to the match; and options for sharing the query result (e.g., through social media or text message). Oftentimes the query result will also include links to third-party services to purchase or stream the song; links to watch the song's music video on YouTube; an option to view song lyrics; and music recommendations. The Shazam icon is displayed somewhere onscreen at all times; thus, users can easily initiate new queries without having to return to the home screen of the application. Selected platforms also offer an "Auto Shazam" feature, which prompts the application to listen and attempt audio matches continuously in the background. Users can additionally retrieve track results through text searches (**Figure 1**, center).

The audio matches, metadata, and other features listed above represent data returned to users. Each query additionally generates a collection of data stored internally to Shazam, including date and time of the query; location information if the user has agreed to share it; the returned track and other candidate tracks that were not returned; metadata associated with the returned track; device platform (e.g., iOS, Android); language used on the device; installation id of the application; and the length of time the query took to perform. Importantly, Shazam also stores the query "offset," which is the time stamp of the initiation of the query relative to the start of the returned track. In other words, the offset tells us when in a song the user performed the query. The present analysis utilizes query offsets and dates.

[2]Shazam also has a desktop application for Mac.

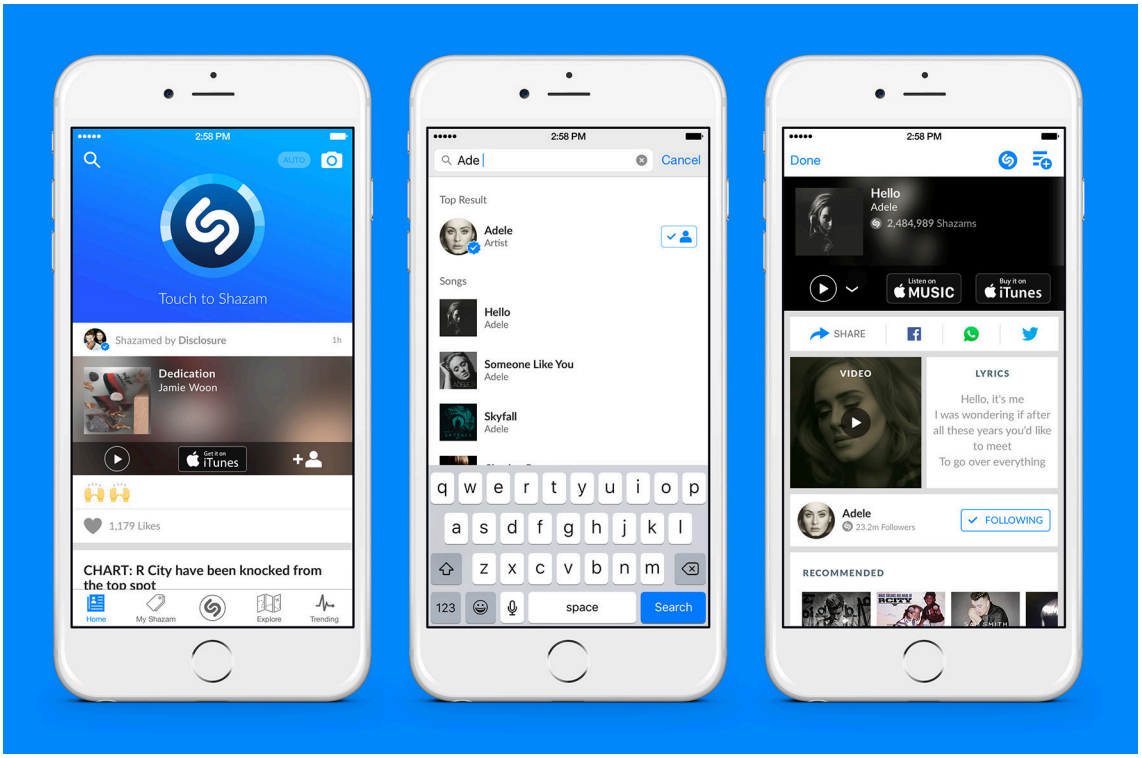


FIGURE 1 | Shazam application screenshots. Shazam audio queries are typically initiated from a mobile device. **(Left)** Upon loading the application, the Shazam icon is prominently displayed on the main screen. **(Center)** Queries can also be initiated through a text search. **(Right)** A successful audio query or selection from text query results returns the track page for the song of interest. Information returned to the user on the track page includes basic metadata about the song, as well as related media including the music video and lyrics when available. The Shazam logo is ubiquitously displayed as users navigate the application; thus, new queries can be initiated at any time. Image used with permission.

2.2. Dataset

2.2.1. Song Set

As this study is a first quantitative analysis of Shazam query offsets, we chose to limit the number of songs used for analysis, but to select songs that would each offer an abundance of Shazam queries while also reflecting a widespread listening audience. For these reasons, we chose as our song set the top 20 songs from the Billboard Year End Hot 100 chart for 2015, which lists the most popular songs across genres for the entire year, as determined by radio impressions, sales, and streaming activity[3]. An additional advantage of selecting songs from this particular chart is that the Billboard Hot 100 chart is released weekly; therefore, our analyses can probe music discovery behavior at specific stages of song popularity. Billboard charts in general are considered a standard industry measure of song popularity, and weekly Billboard Hot 100 charts in particular have been used as a benchmark of song popularity in a number of previous studies (Kim et al., 2014; Nunes and Ordanini, 2014; Nunes et al., 2015; Zangerle et al., 2016).

The set of songs is summarized in **Table 1**. The 15th-ranked song on the Billboard chart ("Bad Blood" by Taylor Swift Feat. Kendrick Lamar) was excluded from analysis due to a known problem with the query data. We therefore include the 21st-ranked song in the set in order to have a set totaling 20 songs.

2.2.1.1. Song metadata

As the selected set of songs all achieved widespread popularity, it was possible to aggregate additional information about the songs from a variety of public sources. We obtained release dates from each song's Wikipedia page. Peak Billboard chart dates were obtained from the Billboard Hot 100 weekly charts and verified against Wikipedia when possible. For songs that held their peak chart position for multiple weeks, we used the date of the first week that the peak position was reached.

To identify the most "correct" version of the audio for each song, we followed the Amazon purchase link, when it was available, from the Shazam track page corresponding to the primary trackid of the song. If the Amazon link was missing or led to a clearly incorrect destination, we located the song on Amazon manually or through an alternate Shazam trackid. We purchased digital versions of all tracks from their resolved Amazon destinations, and then verified the song lengths against primary Spotify results when possible.

[3]http://www.billboard.com/charts/year-end/2015/hot-100-songs.

TABLE 1 | Song and dataset information.

Rank	Title	Artist	Length (s)	Shazam query offsets	
				% usable	# usable
1	Uptown Funk!	Mark Ronson Feat. Bruno Mars	270	98.57	13,855,245
2	Thinking Out Loud	Ed Sheeran	282	98.97	17,142,656
3	See You Again	Wiz Khalifa Feat. Charlie Puth	230	98.73	12,522,399
4	Trap Queen	Fetty Wap	223	98.77	6,072,939
5	Sugar	Maroon 5	236	98.92	5,811,731
6	Shut Up and Dance	Walk the Moon	200	98.47	5,034,637
7	Blank Space	Taylor Swift	232	98.11	6,764,128
8	Watch Me	Silento	186	96.99	4,463,863
9	Earned It (Fifty Shades of Grey)	The Weeknd	252	98.66	7,514,440
10	The Hills	The Weeknd	243	99.08	8,657,473
11	Cheerleader (Felix Jaehn Remix)	OMI	182	96.84	17,933,224
12	Can't Feel My Face	The Weeknd	214	99.34	8,675,375
13	Love Me Like You Do	Ellie Goulding	251	99.56	9,925,090
14	Take Me to Church	Hozier	242	98.82	15,854,482
16	Lean On	Major Lazer & DJ Snake Feat. MØ	177	99.10	19,974,795
17	Want to Want Me	Jason Derulo	208	98.89	9,885,505
18	Shake It Off	Taylor Swift	220	95.90	3,162,707
19	Where Are Ü Now	Skrillex & Diplo with Justin Bieber	251	99.44	7,639,899
20	Fight Song	Rachel Platten	205	99.23	4,359,870
21	679	Fetty Wap Feat. Remy Boyz	197	98.71	3,020,785
				TOTAL	188,271,243

Shazam queries corresponding to 20 top-ranked songs from the Billboard Year End Hot 100 chart for 2015 were analyzed in the study. Song lengths are rounded up to the nearest second. The percent usable and number of usable queries reflect the cleaned datasets. Song 15 is omitted from analysis.

2.2.1.2. Coding of salient musical events

Portions of our analysis focus on the onset of vocals and onset of the first occurrence of the chorus. While the songs analyzed here broadly represent "popular music," assigning conventional pop-song labels, such as verses and choruses, to the structural elements of the songs proved somewhat challenging and subjective. Therefore, for an objective identification of chorus elements within each song, we used lyrics from the Genius website,[4] which are both fully licensed[5] and annotated with structural song-part labels such as "Verse" and "Chorus." For the first onset of vocals, we used the audio timing linked to the first occurrence of labeled (e.g., "Verse" or "Bridge") content in the lyrics, ignoring "Intro" content. For the first occurrence of the chorus, we identified the timing of the audio corresponding to the first instance of "Chorus" or "Hook" material in the lyrics. These times are not necessarily disjoint for a given song—e.g., the first entrance of vocals could be an instance of the chorus.

Additional metadata for the song set, including Shazam and Amazon track identifiers, release and peak Billboard dates, and onset times of vocals and choruses, are included in the Table S1.

2.2.2. Shazam Data Aggregation and Preprocessing

For the selected songs, we aggregated worldwide Shazam query dates and offsets from the Shazam database over the date range January 1, 2014 through May 31, 2016, inclusive. All but one song were released after January 1, 2014, and songs peaked on Billboard between September 6, 2014 and October 31, 2015. Therefore, we consider this date range representative of a song's journey through the Billboard charts. Aggregated data include audio queries only—no text queries—and do not include Auto Shazam queries or queries performed through the desktop application.

Offset values are given in seconds with sub-millisecond precision. Dates are resolved by day, based on GMT timestamps. To clean the data, we removed incomplete queries (missing date or offset values) as well as queries with offsets less than or equal to zero, or greater than the length of the corresponding audio recording. We did not exclude queries whose date preceded the release date, as listed release dates for songs as singles could come after the release date for an album on which the song was included.

The number of usable queries per song ranged from 3,020,785 to 19,974,795, with a median value of 8,148,686 queries. Between 95.90 and 99.56% of the original number of queries for each song were usable after data cleaning. In total, the dataset comprises 188,271,243 queries across the 20 songs. The cleaned datasets are publicly available for download in .csv format from the Stanford Digital Repository (Shazam Entertainment, Ltd., 2016)[6].

[4]http://genius.com.

[5]http://genius.com/static/licensing.

[6]http://purl.stanford.edu/fj396zz8014.

2.3. Analysis

All data preprocessing and analyses were performed using R software, version 3.2.2 (R Core Team, 2015).

2.3.1. Tests of Uniformity

As the present study rests on the assumption that volumes of Shazam queries are higher at some points of a song than others, our first analysis was to determine whether the volume of query offsets for a given song indeed varies over time. To address this first question, we performed two-sided Kolmogorov-Smirnov tests (Conover, 1971) on the distributions of offsets for each song, comparing each distribution of offsets to a uniform distribution over the interval [0, songLength]. Under the null hypothesis of uniformly distributed query offsets, Shazam queries would be equally likely to occur at any point during a song, precluding further exploration of musical events that drive peaks in the query offset histograms. Due to the possibility of ties with our present data size, we added a small perturbation to each offset (uniformly distributed random variables over the interval $[-0.000005, 0.000005]$) before performing the tests.

2.3.2. Assessing Changes in Histogram Shape

Our second question concerned changes in histogram shape over time. Anecdotal analyses of Shazam query offsets have suggested that once a song becomes popular, the distribution of query offsets shifts closer to the beginning of the song.

To approach this problem quantitatively required both a temporal metric of song popularity and a definition for what portion of a song constitutes its "beginning." To address the first point, we selected three points of interest in the life cycle of each song: The song's release date; the date of its peak on the Billboard Hot 100 chart; and the end dates of the dataset. Ranges of time between these three events varied by song. Songs peaked on Billboard between 19 and 463 days after release, with a median release-to-peak delay of 127 days. The time range between peaking on Billboard and the last date in the dataset ranged from 213 to 633 days, with a median value of 374 days. Dates and latencies between dates are reported in Table S1.

For the second point, instead of choosing an arbitrary, fixed duration (e.g., 30 s) to denote the beginning of each song, we devised an analysis that would compare distributions over all possible beginning durations d_b using the following procedure. For each song, we first extracted the first 100,000 queries following release and peak Billboard dates, and the final 100,000 queries, by date, in the dataset. Following that, for d_b increasing from 1 to the length of the song in seconds, we performed Chi-squared tests of proportions on Billboard peak date vs. release date, end of dataset vs. release date, and end of dataset vs. Billboard peak date. Because we were specifically interested in assessing whether queries migrated toward the beginning of the song for the later set of queries, we performed one-sided tests with the alternative hypothesis being that the proportion of queries less than d_b was greater for the set of queries corresponding to the later time point.

Due to data size, the p-values resulting from these tests were generally so small as to be uninformative. Therefore, we focus on percentile Chi-squared statistics over increasing d_b for each song, and report these results across songs. This analysis comprises a total of 13,503 multiple comparisons (three comparisons per time point per song times 4,501 total time points across all songs). Therefore, as we do not correct here for multiple comparisons, we use a conservative significance threshold of $p < 10^{-10}$, keeping us well under the statistical significance threshold of $\alpha = 0.01$, had a Bonferroni correction been performed (Bonferroni, 1936; McDonald, 2014).

2.3.3. Computing Histogram Slopes at Salient Musical Events

For our third analysis, we wished to test the hypothesis that salient musical events drive a subsequent increase in query volume. For the present analysis we chose three salient structural events that were present in every song: Beginning of song, initial onset of vocals, and initial onset of chorus/hook section.

We devised an exploratory analysis of the query offset volume around these musical events by focusing on offset histogram slopes following these events. As our previous analysis revealed a leftward shift in offset distributions for later dates, we used only the first 1,000,000 queries for each song (by date) for this computation. We first used local polynomial regression (Fan and Gijbels, 1996) to estimate histogram slopes over time for each song, with a temporal resolution of 1 s. We then converted each song's estimated histogram slopes to slope percentiles in order to bring the data to a more common scale across songs. As the timing of onset of vocals and chorus can vary from song to song, we extracted 15-s analysis windows starting from the onset of each event, and then for each event type (beginning, vocals, chorus) we aggregated the windows across songs so that the 15-s intervals were now aligned according to the onsets of the musical event of interest—similar to the approach taken by Tsai et al. (2014) in analyzing physiological responses at chorus onsets across a set of popular songs.

For each of the musical events of interest, we report the median of histogram slope percentiles over time across the songs, along with first and third quartiles. For reference, we also report results from the same analysis, using randomly selected window start times for each song.

2.3.4. Data Size and Histogram Consistency

Our final analysis examined the relationship between data size and histogram consistency. One reason for selecting massively popular songs was to have millions of queries to work with for each. But do the underlying distributions of the data require such large collections of queries, or is a smaller sample size sufficient?

To investigate this matter further, we assessed consistency of query offset distributions, computing histogram distance between disjoint data subsets of varying sample size for individual songs. For songs whose data comprised more than 8 million queries, we drew a random subsample of 8 million queries for the following analysis. On a per-song basis we randomly partitioned the collection of queries into two halves. For an increasing number of trials n_i from 1 to $nTotalTrials/2$, we normalized the cumulative histograms

of the two halves into discrete probability densities (each summing to 1), and then used the total variation distance (Levin et al., 2009) to measure the distance between these two probability distributions. This partitioning procedure was repeated over 100 randomization iterations for each song. We then computed the mean output across randomization iterations for each song. We report the median, across songs, of these results.

3. RESULTS

3.1. Distributions of Query Offsets Are Not Uniform

For our first analysis, we assessed whether query offsets for a given song are uniformly distributed over time (implying no relationship between musical events and number of queries), or whether the volume of queries varies over the course of a song. Scale-free plots of the offset histograms are shown in **Figure 2**. By visual inspection, the histograms do not reflect uniform distributions of query offsets. Additionally, the timing, height, and shape of the histogram peaks vary from song to song. Results of the Kolmogorov-Smirnov tests of uniformity provide a quantitative validation of our observations, rejecting the null hypothesis with $p < 10^{-15}$ for all songs (no correction for multiple comparisons).

3.2. Shapes of Offset Histograms Change over Time

Our second question was whether the distribution of query offsets shifts toward the beginning of a song as the song moves through its hit life cycle—that is, whether users tend to perform the Shazam query earlier in a song once the song has attained, or dropped from, popularity. Query offset histograms around release date, peak Billboard date, and end of the dataset are shown for the first four songs in our song set in **Figure 3** (plots for remaining songs are included in Figures S1–S4). Each subplot comprises 100,000 queries. The shift in the histogram toward the beginning of the song (left side of each plot) is evident for each of these songs, especially for the "End" subset of the dataset.

As a more quantitative assessment, we performed Chi-squared tests of proportions on sets of queries drawn from the time of song release, peak Billboard date, and final dates of the dataset. Chi-squared tests of proportions were performed over a beginning window of increasing duration to assess the size of the statistic when comparing pairs of life-cycle samples. Results are shown in **Figure 4**. In the top row of plots, percentile Chi-squared statistics (y-axis) as a function of beginning window length in seconds (x-axis) are plotted, with the median across songs shown in black, and first and third quartile of individual songs shown in gray. Median Chi-squared statistic percentiles are notably high at the beginnings of songs for end date vs. peak Billboard date (peaking at 13 s), and end date vs. release date (peaking at 19 s). This indicates that across songs, tests of proportions as to whether the later set of queries was distributed closer to the start of a given song returned consistently high Chi-squared statistics for the beginning portions of the songs.

More detail on individual songs is given in the bottom plots of **Figure 4**, which specifies the beginning window lengths that produced statistically significant Chi-squared statistics. Here, we see that nine of the songs in the set exhibited a constant migration of queries toward the start of the song from release date to peak Billboard date, and all 20 songs exhibited this shift when comparing queries from the peak Billboard date to those from the final dates in the dataset (recall that Song 15 was omitted from analysis). Comparing release date to end date, all but one song (Song 10) exhibit a leftward histogram shift when the first 30 s of the histogram are analyzed. Taken together, these results suggest that users do tend to perform queries earlier in a song for dates toward the end of the dataset, compared to dates around the song's release or peak on the Billboard Hot 100 chart.

3.3. Salient Musical Events Drive Increase in Queries

Our third analysis examined whether three salient musical events—the start of a song, the first onset of vocals, and the onset of the first chorus—would drive an increase in queries. This is a first step toward relating the histogram peaks, evident in **Figure 2**, to structurally salient musical events, and toward generalizing music discovery behavior across the songs, which vary in their timing and arrangement of shared musical events. The results of the histogram slope analysis by song part, summarized across songs, is shown in **Figure 5**. Each plot represents a 15-s window time-locked to the beginning, first onset of vocals, onset of first chorus, and random time point, respectively, across songs. Therefore, the x-axis of each plot is time, and the y-axis is percentile of histogram slope. The three structurally salient time points are all followed by notably high histogram slopes, representing an increase in query volume over time. As shown by the median measure across songs (black line), this behavior does generalize across the song set. The shaded quartile area suggests that this behavior is more consistent for onset of vocals than onset of chorus. In comparison, histogram slopes from randomly selected 15-s windows, shown in the bottom plot, do not reach the percentile levels of the musically salient conditions.

3.4. Sample Size for Consistent Query Offset Distributions

Our final question concerns the necessary data size to reach a "consistent" distribution of offsets. **Figure 6** shows histograms of random subsamples of varying amounts for four of the songs in our set (subsampled histograms for the remaining songs can be found in Figures S5–S8). As can be appreciated by visual inspection of the plots, main peaks in offset distributions are fairly well separated from noise with as few as 1,000 queries. Based on observation, we consider a sample of 20,000 adequate to represent the general shape of the overall distribution, with finely temporally resolved peaks emerging when 50,000 queries are used.

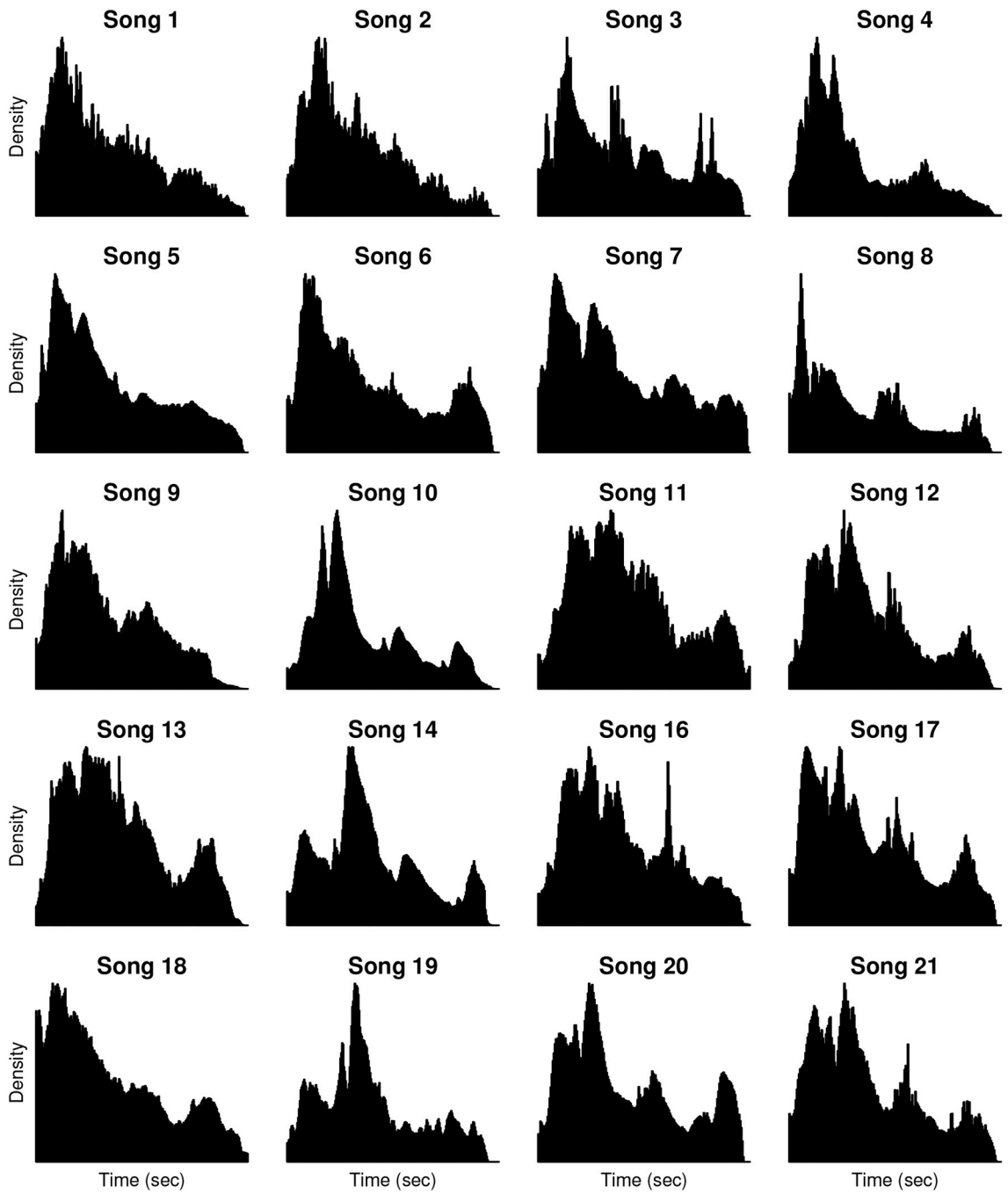


FIGURE 2 | Shazam query offset histograms. Histograms are computed from the query offsets of the 20 hit songs analyzed in the study (summarized in **Table 1**). Each histogram presents the density of Shazam queries (y-axis) over time (x-axis) for a given song. Histograms are scaled to maximum density and song duration on a per-song basis. The number of queries per song ranges from 3,020,785 (Song 21) to 19,974,795 (Song 16), with a median of 8,148,686 queries per song.

The median total variation distance between randomly sampled disjoint subsets as a function of subsample size across the song set is shown in **Figure 7**. As shown in the left panel (**Figure 7**), the trajectory of these results is consistent across songs. The distance between distributions of two disjoint subsamples for a given song decreases rapidly as a function of sample size, leveling off well below 500,000 queries. While there exists no standard metric of "good" total variation distance, we identify the median subsample size necessary to achieve total variation distance of 0.1 and 0.05 (**Figure 7**, right panel). A median subsample size of 26,000 queries is required to achieve total variation distance of 0.1—somewhat in line

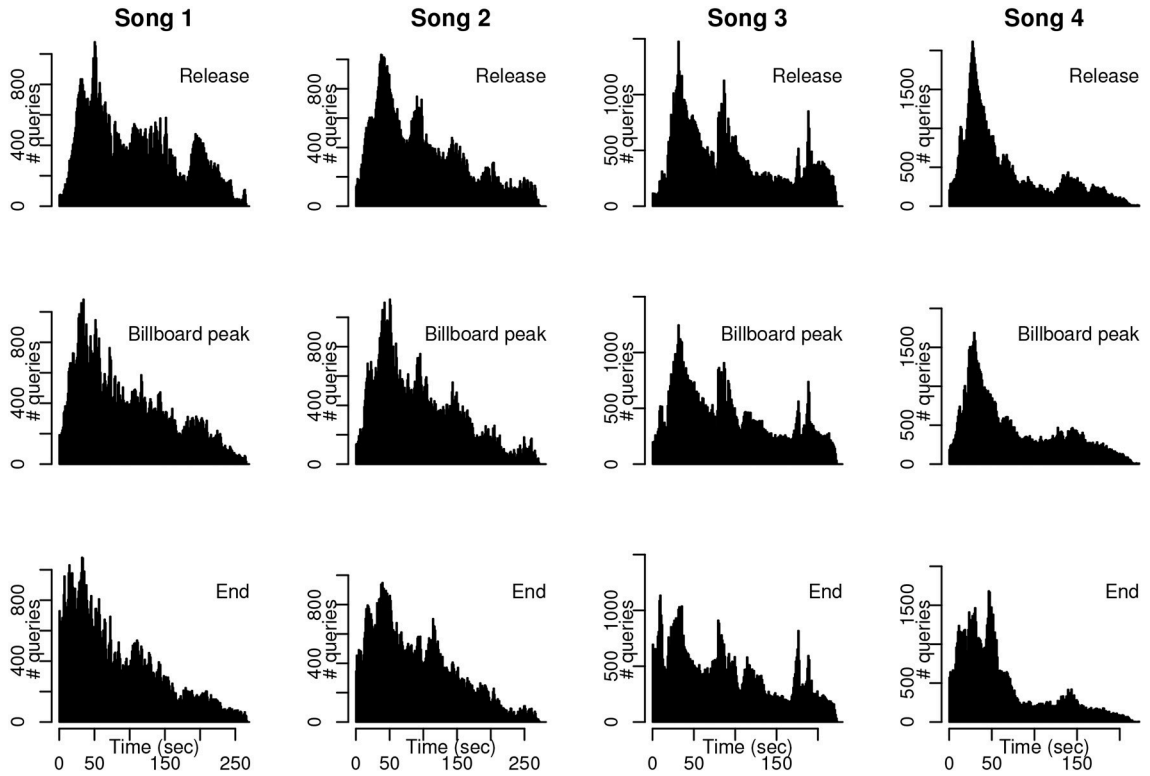


FIGURE 3 | Example histograms throughout hit song life cycle. Subsampled query distributions at the time of song release, date of peak position on Billboard, and end of the dataset for four songs. The distribution of query offsets for the end dates in particular exhibit a pronounced shift toward the beginning of each song.

with our observations of the histograms in **Figure 6**—while 104,000 queries correspond to a median total variation distance of 0.05.

4. DISCUSSION

In this study, we investigated music discovery behavior on a large scale by analyzing the timing of Shazam queries during popular songs. Using a dataset of over 188 million queries of 20 hit songs, our findings suggest a relationship between musical events and the timing of Shazam queries. We show that query offsets are not uniformly distributed throughout a song, but rather vary over the course of a song, and may thus be driven in particular by salient musical and structural elements of the song. Furthermore, the shapes of the offset histograms themselves change over the course of the hit song life cycle, showing that the musical content that compels listeners to query a song changes as a function of song popularity or listener exposure to a song. A closer analysis of salient song parts reveals that the onset of vocals and the first occurrence of the chorus in particular drive an increase in queries. Finally, having ample data, we assessed the consistency of the data as a function of data size, and propose that Shazam query offsets for the present song set reach consistent distributions with around 26,000 queries.

Shazam's user data offer several advantages for the study of music discovery. First and foremost is the scale and scope of the data, representing a massive global user base that performs millions of queries each day. Also, while the current study focused on only a small set of songs, Shazam's music catalog contains over 30 million deduplicated tracks. Thus, in terms of both size and demographic diversity of the experimental sample (users), as well as number of stimuli (song catalog), Shazam data capture music discovery at a scale not attainable in controlled studies. The dataset analyzed here is comparable in size to other recently released industrial datasets for music research. For example, the #nowplaying dataset currently exceeds 56 million tweets (Zangerle et al., 2014), while Gracenote's GNMID14 dataset exceeds 100 million music identification matches (Summers et al., 2016). Shazam data are also ubiquitous, meaning that they reflect music discovery in a variety of contexts throughout daily life. As a result, the user data reflect a wide range of music discovery scenarios. Third, Shazam data possess an ecological validity lacking in controlled laboratory studies, as users engage the application in real-world information-seeking scenarios, and were not asked to adopt this behavior as part of a study. Finally, what uniquely differentiates Shazam's data from most other data—including other large-scale social media data—is its objectivity. By this, we mean that under the assumed primary use case of learning the identity of a musical excerpt, Shazam queries are motivated by interest in some aspect of the musical content, even while the queried excerpt may be unknown to the user. Therefore, interest in musical content may be reflected more

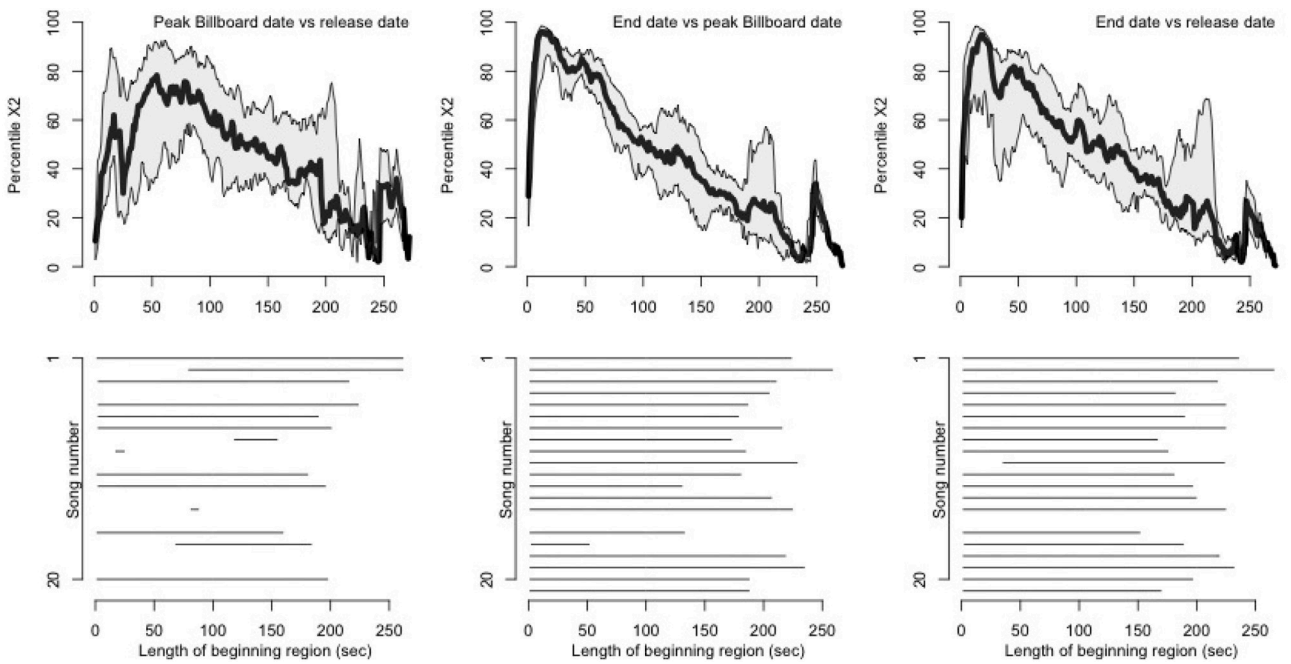


FIGURE 4 | Changes in histogram shape during the hit song life cycle. We performed Chi-squared tests of proportions to assess whether distributions of query offsets migrate toward the beginning of the song as a song achieves popularity. One-sided tests compared each later-vs.-earlier pair of subsamples using a beginning window of increasing duration. **(Top)** Median percentile Chi-squared statistics, across songs, with first and third quartiles, for each pairwise test over beginning windows of increasing length. We converted statistics to percentiles on a per-song basis to impose a common scale across songs. For peak Billboard vs. end date and release date vs. end date, window lengths of around 50 s or less produce notably high Chi-squared statistics, demonstrating that query offsets for the latest dates are more concentrated at the beginnings of songs. **(Bottom)** Raster plot of beginning window lengths producing $p < 10^{-10}$ in the tests of proportions for individual songs (no correction for multiple comparisons).

directly in Shazam queries than in other formats such as tweets, where the content of a posted tweet (and decision whether to post it) has been mediated by the user, reflecting a confluence of musical taste and the user's conscious awareness of how the posted content aligns with his or her expressed identity (Lonsdale and North, 2011; Rentfrow, 2012).

4.1. Musical Correlates of Shazam Queries

4.1.1. Query Volume Varies Throughout a Song

In our first analysis, we tested the uniformity of the offset histograms. Visual inspection of the offset histograms of our song set (**Figure 2**) and results of statistical tests indicate that the query offset distributions are not uniform, and that queries are more likely to occur at some points during the songs than others. In this way, Shazam query offset histograms may facilitate the "locate" research proposed by Honing (2010), in that they reveal points in a song that a number of listeners found engaging.

The timing and heights of histogram peaks vary from song to song. We surmised that this was a reflection of the variation in song structure (e.g., arrangement of choruses, verses, and other elements) across the song set, but that the peaks might reflect structurally salient events that occur across the songs. By analyzing regions of the histograms time-locked to such events, we were able to show that the initial onset of vocals and occurrence of the first chorus drive increases in query volume—represented by high percentiles of histogram slopes—in a consistent fashion across songs.

In relating offset histogram peaks to musical events, it is important to keep in mind that users are assumed to successfully query a given broadcast of a song only once. This is reflected to some extent in the overall downward trend in query volume over the duration of a song. Musical content driving Shazam queries may be better characterized, then, as the *first* content in a song that compelled a user to take action and perform the query. Therefore, this content was presumably more engaging than content that came before, but not necessarily more engaging than content that comes after—the user just would not need to query the song a second time, as he had already received the benefit of the query result. Under this reasoning, songs for which the highest histogram peak is not the first peak (for example, Song 14, Song 19, and Song 20) may be of particular interest, as these represent a break from the conventional histogram shape, and may highlight especially engaging musical material occurring later in the song. Furthermore, as shown in **Figure 8**, histogram peak heights can vary even across occurrences of the same song part (here, most notably for the second verse compared to the first), which may reflect changes in texture, instrumentation, or other musical content. Finally, our present analysis used histogram slopes as indicators

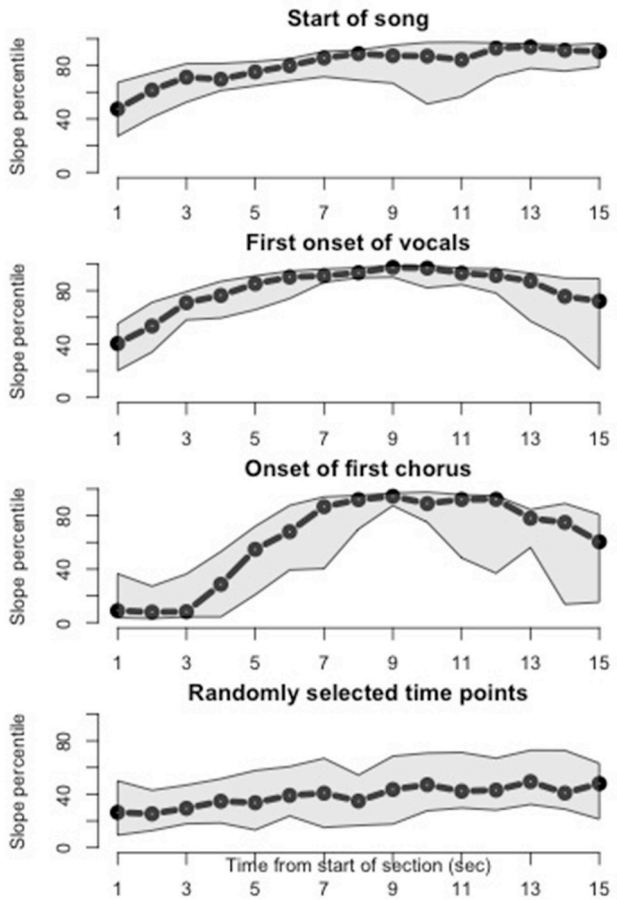


FIGURE 5 | Slopes of histograms after salient musical events. Median (black) with first and third quartiles (gray) of histogram slopes across songs are plotted, time-locked to specific musical events. Histogram slopes for each song were converted to percentiles prior to plotting. Histogram slopes increase following the start of the song **(top plot)**, the first onset of vocals in the song **(second plot)**, and the onset of the first chorus **(third plot)**. In particular, histogram slopes are consistently high across songs around 9 s after the first onset of vocals and first onset of the chorus. **(Bottom plot)** When randomly selected time points, rather than salient musical events, are considered, the median histogram slope across songs over time peaks around the 50th percentile.

of upcoming histogram peaks; future analyses could utilize other histogram features, such as the density or timing of the peaks themselves, or the local minima interspersed between the peaks.

4.1.2. Inferring Intent-to-Query Time

A Shazam query typically does not occur at the exact moment the user was compelled to perform the query. In many cases, the user must retrieve his or her mobile device, unlock it, and load the Shazam application before the query can be performed. Therefore, there exists in the offset data an unknown latency between intent-to-query and query time, which can range from 0 to 10 s or more. We did not attempt to estimate or correct for this latency in our present analyses. However, the histogram slopes following salient musical events may provide some insight into the duration of this delay. If our musical events of interest in fact drive increased queries, we might interpret the time point after such events, at which histogram slopes are consistently high across songs, as an estimate of the mean latency between onset of the song part and initiation of the query. Based on the present results (shown in **Figure 5**), histogram slopes become consistently high around 9 s after the onset of vocals or the first chorus.

We find that peaks and troughs of an offset histogram are better aligned with structural segmentation boundaries of the song when the histogram is shifted to account for an estimated latency. For example, **Figure 8** shows the offset histogram for Song 10, with structural segmentation boundaries visualized in the background. When all offsets are shifted back by 6 s as shown in the figure, the resulting histogram aligns well with the structural segmentation boundaries. Visualizing the other songs in a similar fashion reveals some variation in adjustments required to optimally align histograms with song part boundaries.

Even so, the assumption that histogram slope percentiles or minima convey the intent-to-action delay remains speculative at this stage. Furthermore, the histogram slopes over our time window of interest vary from song to song, as does the optimal time shifting of histograms to align local minima with song-part boundaries. Therefore, additional research—perhaps in a controlled experimental setting—will be required to better characterize this delay, and to determine whether our current proposed approaches for inferring it are appropriate.

4.1.3. Impact of Hit Song Life Cycle

As shown in our second analysis, the shapes of offset histograms change over the life cycle of the hit songs in our song set. As a song attained and receded from its peak position on the Billboard chart, queries tended to occur closer to the start of the song. Therefore, even though the underlying musical content was unchanged, users tended to query the audio earlier once a song became successful. As we will later discuss, the intent of the query may have changed, e.g., users querying later in the life cycle may have been doing so for reasons other than to learn the identity of the song. However, it may also be that repeated exposures to such popular songs, which—even while the identity of the song may remain unknown—enhance familiarity, processing fluency, and even preference (Nunes et al., 2015), could compel the user to query the song earlier than he would have done prior to so many exposures. Therefore, it would be interesting to repeat this analysis with songs that never achieved ubiquitous broadcast and widespread popularity, in order to assess in finer detail the impact of popularity and exposure on changes in music discovery behavior.

In interpreting the changes in histogram shape over a song's life cycle, we note that the earliest and latest subsets of data (release date and end date) are always disjoint, but that repeated observations may exist with either of these subsets and the

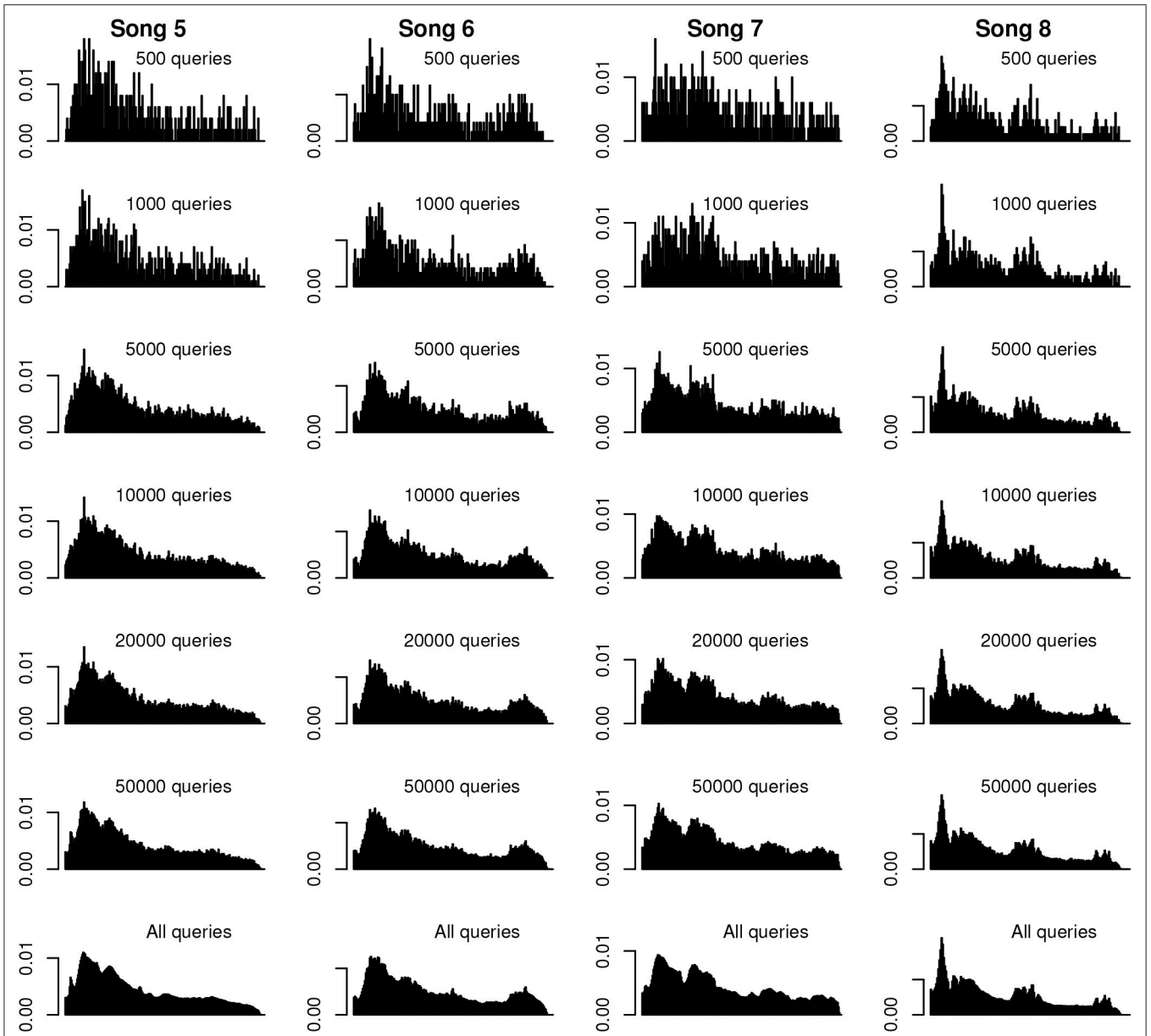


FIGURE 6 | Example subsampled histograms. Histograms (density plots) of various quantities of random subsamples for four of the songs. Histograms are scaled to common density and time axes on a per-song basis. The most prominent peaks of the full-sample histogram emerge with as few as 1,000 queries, and are visually similar by 20,000 queries. The finer details of the full-distribution histograms are discernible with subsamples of 50,000 queries.

Billboard peak date subset—for example, if a song peaked on Billboard soon after its release.

4.1.4. Disentangling Discovery and Preference

Under the premise that Shazam queries are primarily searches for identities of unknown songs, it would be erroneous to equate a user's Shazam history with his or her most-loved music. However, if we may assume that users query songs because they are in some way attracted to, or at least aroused by, the songs' musical content, we may conclude that musical attributes of a user's queried songs reflect, to some extent, the musical preferences of that user. In other words, a queried song's musical content, especially around the query offset, may contain features that compel the user to take action and want to know more. In this sense, one's discovered music, more so than freely chosen songs, may be more widely representative of musical preferences, as it encompasses music (and musical features) beyond the scope of what a user could have articulated in advance that he wanted to hear—and possibly across a broader range of musical genres. And, given that known recommended tracks have been shown to be received more positively by listeners than unknown recommendations (Mesnage et al., 2011), music discovery data

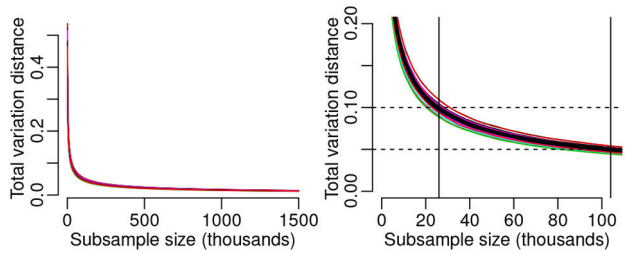


FIGURE 7 | Histogram consistency as a function of data size. (Left) Median of per-song total variation distance computed across the songs, as a function of subsample size in each of the two distributions being compared. Results of individual songs (colored curves) lie close to the median. Total variation distance shows a sharp drop for subsample sizes up to around 200,000 observations followed by a gradual decrease to a subsample size of 1.5 million. **(Right)** The median subsample size corresponding to a total variation distance of 0.1 is 26,000 observations. Median total variation distance of 0.05 is attained with a subsample size of 104,000 queries.

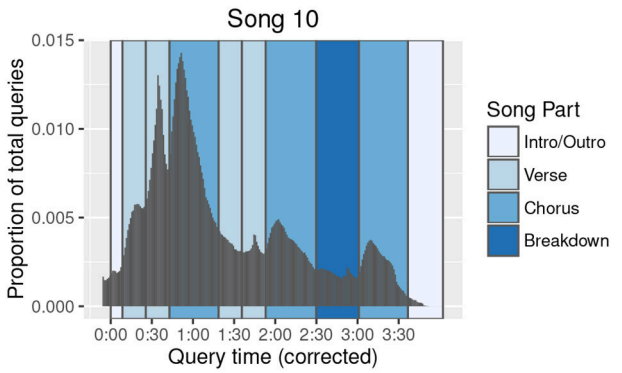


FIGURE 8 | Song 10 query offset histogram annotated with song parts. The query offset histogram (density plot) of Song 10 is plotted with structural segmentation annotations. The entire distribution has been shifted back in time by 6 s to demonstrate better alignment of the histogram shape with structural segmentation boundaries once an estimated intent-to-action query latency is considered. Prominent peaks in the histogram are now seen to correspond to specific song parts.

may be especially valuable in deepening our understanding of positive reception of new music, since it largely reflects music that was both unknown to, and positively received by, the user.

4.1.5. Inferring User Intent

While the typical Shazam use case is assumed to be the identification of an unknown audio excerpt, this is by no means the only use case of the service. Other use cases include querying a song in order to access other features of the query result, including the music video, lyrics, artist information; to purchase the song or add it to a third-party playlist; to establish a static access point for the song; to share the song via messaging or social media services; or to demonstrate or test the performance of the application. The shift in query offsets toward the beginning of songs that have peaked in popularity could thus reflect a change in user intent, whereby fewer users are using Shazam to learn the identity of the song at that point, and are instead reflecting an alternative use case.

In fact, in the realm of web searches, informational need is known to account for <50% of queries, with navigational (attempting to reach a specific site) and transactional (reaching a site where further interactions will take place) thought to account for the remainder of use cases (Broder, 2002). This framework of query intent has more recently been extended to the case of multimedia search, for example text queries for videos (Hanjalic et al., 2012). The Shazam use cases mentioned thus far could arguably be categorized as informational (e.g., learn song identity, information about a song) or transactional (e.g., add song to Spotify playlist). However, user intent is not always communicated clearly in a query, and in fact may not even be clear to the user as the query is being performed (Kofler et al., 2016). In the case of Shazam, audio queries are invariant—all initiated by a button press—and therefore provide no insight into user intent. However, it could be possible to infer intent through other factors, such as day or time of query, geography, song popularity, or previous users' interactions with the query result, and to adjust the content of the query result accordingly.

4.2. Considerations

While the dataset used in the present study provides several advantages for studying music discovery on a large scale, there exist several unknown contextual factors underlying the queries. First, as our analysis takes into account only query offset and date, we gain no insights from the time or location of the queries. Furthermore, from the present data we do not know how the user reacted to the query result, or whether the query reflects positive reception of the musical content.

In addition, Shazam's utility varies according to the music listening setting. Streaming services and personal playlists provide ubiquitous metadata, which can be accessed with often greater ease than performing a Shazam query. Therefore, Shazam is likely used primarily to identify unknown songs in settings where the user does not otherwise have easy access to song metadata. This could include radio listening as well as public settings in which the user does not control music play (e.g., club, retail, or restaurant). While streaming and playlist listening scenarios typically involve "zero-play" music consumption—that is, the song is likely heard from its start (Frank, 2009)—in radio and other Shazam-worthy settings, we cannot assume the user was exposed to the song from its onset, which could affect the interpretation of some of the present results.

Issues related to the performance of the application should be noted as well. Spurious observations were addressed to some extent during data cleaning, but likely persist throughout the data. Due to a pre-recording functionality of Shazam that begins at application launch, time stamps of query offsets may precede the time of the actual query by up to 3 s for an unknown percentage of users. Certain listening environments, such as those with heavy reverberation, can impede the performance

of the application and could therefore require multiple query attempts in order to obtain a result. The presence of vocals during a song may also complicate interpretation of results. While we might interpret a connection between vocals and increased queries as a reflection of musical engagement, it could also be the case that portions of the song with highly prominent vocals may be easier for the Shazam algorithm to match successfully. Prominent vocals may also be easier for a human listener to pick out in a noisy environment. Therefore, disentangling "vocalness" from "catchiness" (by which we mean engaging in the moment, not necessarily memorable in the long term; Burgoyne et al., 2013) could be a useful topic for future research.

In sum, conclusions from the current study must be taken in the context of various unknowns pertaining to users, listening settings, application performance, and other uncontrolled factors. The research questions addressed here could therefore benefit from further investigation in a human-subjects laboratory study setting, where potential confounds and unknowns can be controlled.

4.3. Future Work

4.3.1. Hooks and Catchiness

Through an analysis of offset histogram slopes, this study provides first insights into Shazam queries following song starts, initial onsets of vocals, and first occurrences of choruses. This approach could be broadened to consider more generally the role of "hooks" in music discovery. Musical hooks are defined in many ways, largely describing the part(s) of a song that grab the listener's attention and stand out from other content (Burns, 1987). Hooks need not be restricted only to popular music (Mercer-Taylor, 1999), but are often discussed in the context of popular songs and are thought to occur primarily at structural segmentation boundaries (i.e., starts of song parts; Burns, 1987; Mercer-Taylor, 1999; Burgoyne et al., 2013). The construction of a hook can involve musical features such as rhythm, melody, and harmony, as well as production decisions such as editing and mix (Burns, 1987). The study of musical hooks historically involved human analysis of hand-picked excerpts (Mercer-Taylor, 1999; Kronengold, 2005); in recent years, computational approaches have also evolved (Burgoyne et al., 2013; Van Balen et al., 2013, 2015), which may facilitate hook research over large audio corpuses.

Singability is considered to be a characteristic of hooks (Kronengold, 2005), and is thought to increase listener engagement, both by increasing familiarity and by inspiring the listener to sing along (Frank, 2009). In addition to such intrinsic factors as singability or catchiness, the arrangement of structural elements within a song is also critical to engaging the listener (Mercer-Taylor, 1999). Shazam query offset histograms could prove useful in exploring all of these topics further. While we used annotated lyrics to guide our identification of salient song parts, future research could consider computational models of catchiness—perhaps constructed from computationally extracted audio features (McFee et al., 2015),[7] higher-level musical features (Van Balen et al., 2015),[8] and structural segmentation boundaries (Nieto and Bello, 2016)[9]—and use Shazam query distributions to validate the models. Alternatively, a model could be learned directly from features of the audio corresponding to the histogram peaks themselves. In addition to increasing our understanding of what types of musical features attract listeners, these analyses have the potential to explain the appearance of higher histogram peaks later in a song, as in Song 10 (**Figure 8**).

4.3.2. Modeling and Prediction of Hit Songs

Large-scale music discovery data may also provide new insights into modeling and predicting hit songs. Hit prediction remains an open area of research (Pachet and Roy, 2008; Pachet, 2012), and has been attempted with audio and lyrics features (Dhanaraj and Logan, 2005; Herremans et al., 2014) and Twitter data (Kim et al., 2014; Zangerle et al., 2016) with varying success. Other recent studies have found instrumentation (Nunes and Ordanini, 2014) and lexical repetition (Nunes et al., 2015) to be predictive of peak chart position for past Billboard hits. The potential of Shazam's data for hit prediction has been discussed in news articles.[10] Audio, lyrics, instrumentation, and other features found to be predictive of success in the past studies mentioned above could be explored using query offset histograms. While the present analysis considered only hit songs, query offsets—or other Shazam data attributes—of a song set with more variation in popularity could lead to the formulation of unique predictors of eventual song success.

4.3.3. Other Time-Based Analyses

When thinking about Shazam queries, time can signify many things. Our present analyses considered two types of time: The timing of queries over the course of a song, and the longer-term time scale of the hit song life cycle, spanning several months. Other approaches to time could include day of week—known to impact listening behavior (Schedl, 2013) as well as Shazam query volume—and time of day.

4.3.4. Other Behaviors and Data Attributes

The present study provides novel insights into music discovery, using only two of Shazam's many data attributes. A variety of additional musical questions could be addressed using Shazam user data. User interactions with the application after receiving a query result could provide insight into user preference and user intent. Other analyses could model music discovery or preference by considering specific geographies, musical genres, or even individual users. Large-scale data have been used to address specific musical questions including the long tail in music-related microblogs (Schedl et al., 2014), social media behavior of Classical music fans (Schedl and Tkalčič, 2014), the relationship between musical taste and personality factors (Bansal and Woolhouse, 2015), and Twitter activity around a specific musical event (Iren et al., 2016).

[7] https://github.com/librosa/librosa.

[8] https://github.com/jvbalen/catchy.

[9] https://github.com/urinieto/msaf.

[10] http://www.theatlantic.com/magazine/archive/2014/12/the-shazam-effect/382237/.

Using Shazam data in this way—to address specific musical questions—promises interesting approaches for future research endeavors.

AUTHOR CONTRIBUTIONS

Conceived and designed the research: BK, FR, CB, JB. Aggregated the data: BK, CB. Analyzed the data: BK, FR. Wrote the paper: BK, FR, CB, JB.

FUNDING

This research was supported by the Wallenberg Network Initiative: Culture, Brain, Learning (BK, JB), the Roberta Bowman Denning Fund for Humanities and Technology (BK, JB), Shazam Entertainment, Ltd. (BK, CB), and the E. K. Potter Stanford Graduate Fellowship (FR).

ACKNOWLEDGMENTS

The authors thank Martha Larson, Fabio Santini, and Julius Smith for helpful discussions relating to this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2017.00416/full#supplementary-material

REFERENCES

Bansal, J., and Woolhouse, M. (2015). "Predictive power of personality on music-genre exclusivity," in *Proceedings of the 16th International Society for Music Information Retrieval Conference* (Malaga), 652–658.

Bonferroni, C. E. (1936). *Teoria Statistica delle Classi e Calcolo delle Probabilita*. Libreria Internazionale Seeber.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum* 36, 3–10. doi: 10.1145/792550.792552

Burgoyne, J. A., Bountouridis, D., Van Balen, J., and Honing, H. (2013). "Hooked: a game for discovering what makes music catchy," in *Proceedings of the 16th International Society for Music Information Retrieval Conference* (Curitiba), 245–250.

Burns, G. (1987). A typology of 'hooks' in popular records. *Popular Music* 6, 1–20. doi: 10.1017/S0261143000006577

Conover, W. J. (1971). *Practical Nonparametric Statistics*. New York, NY: John Wiley and Sons.

Cunningham, S. J., Bainbridge, D., and McKay, D. (2007). "Finding new music: a diary study of everyday encounters with novel songs," in *Proceedings of the 8th International Conference on Music Information Retrieval* (Vienna), 83–88.

Dhanaraj, R., and Logan, B. (2005). "Automatic prediction of hit songs," in *Proceedings of the 6th International Conference on Music Information Retrieval* (London), 488–491.

Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability*, Vol. 66. New York, NY: CRC Press.

Frank, J. (2009). *Futurehit.DNA: How the Digital Revolution is Changing Top 10 Songs*. Nashville, TN: Futurehit, Inc.

Hanjalic, A., Kofler, C., and Larson, M. (2012). "Intent and its discontents: the user at the wheel of the online video search engine," in *Proceedings of the 20th ACM International Conference on Multimedia, MM '12*, (New York, NY: ACM), 1239–1248. doi: 10.1145/2393347.2396424

Hauger, D., Schedl, M., Košir, A., and Tkalčič, M. (2013). "The million musical tweets dataset: what can we learn from microblogs," in *Proceedings of the 14th International Society for Music Information Retrieval Conference* (Curitiba), 189–194.

Herremans, D., Martens, D., and Sörensen, K. (2014). Dance hit song prediction. *J. New Music Res.* 43, 291–302. doi: 10.1080/09298215.2014.881888

Honing, H. (2010). Lure(d) into listening: the potential of cognition-based music information retrieval. *Empirical Musicol. Rev.* 5, 121–126.

Iren, D., Liem, C. C. S., Yang, J., and Bozzon, A. (2016). "Using social media to reveal social and collective perspectives on music," in *Proceedings of the 8th ACM Conference on Web Science, WebSci '16* (New York, NY: ACM), 296–300. doi: 10.1145/2908131.2908178

Kim, Y., Suh, B., and Lee, K. (2014). "#nowplaying the future Billboard: mining music listening behaviors of Twitter users for hit song prediction," in *Proceedings of the First International Workshop on Social Media Retrieval and Analysis, SoMeRA '14* (New York, NY: ACM), 51–56. doi: 10.1145/2632188.2632206

Kofler, C., Larson, M., and Hanjalic, A. (2016). User intent in multimedia search: a survey of the state of the art and future challenges. *ACM Comput. Surv.* 49, 36:1–36:37. doi: 10.1145/2954930

Kronengold, C. (2005). Accidents, hooks and theory. *Popular Music* 24, 381–397. doi: 10.1017/S0261143005000589

Laplante, A., and Downie, J. S. (2011). The utilitarian and hedonic outcomes of music information-seeking in everyday life. *Libr. Inform. Sci. Res.* 33, 202–210. doi: 10.1016/j.lisr.2010.11.002

Levin, D. A., Peres, Y., and Wilmer, E. L. (2009). *Markov Chains and Mixing Times*. Providence, RI: American Mathematical Society.

Lonsdale, A. J., and North, A. C. (2011). Why do we listen to music? A uses and gratifications analysis. *Br. J. Psychol.* 102, 108–134. doi: 10.1348/000712610X506831

McDonald, J. H. (2014). *Handbook of Biological Statistics, 3rd Edn*. Baltimore, MD: Sparky House Publishing.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., et al. (2015). "librosa: audio and music signal analysis in Python," in *Proceedings of the 14th Python in Science Conference* (Austin, TX), 18–25.

Mercer-Taylor, P. (1999). Two-and-a-half centuries in the life of a Hook. *Popul. Music Soc.* 23, 1–15. doi: 10.1080/03007769908591729

Mesnage, C. S., Rafiq, A., Dixon, S., and Brixtel, R. P. (2011). "Music discovery with social networks," in *2nd Workshop on Music Recommendation and Discovery* (Chicago, IL).

Moore, J. L., Joachims, T., and Turnbull, D. (2014). "Taste space versus the world: an embedding analysis of listening habits and geography," in *Proceedings of the 15th International Society for Music Information Retrieval Conference* (Taipei), 439–444.

Nieto, O., and Bello, J. P. (2016). "Systematic exploration of computational music structure research," in *Proceedings of the 17th International Society for Music Information Retrieval Conference*, 547–553.

Nunes, J. C., and Ordanini, A. (2014). I like the way it sounds: the influence of instrumentation on a pop song's place in the charts. *Music. Sci.* 18, 392–409. doi: 10.1177/1029864914548528

Nunes, J. C., Ordanini, A., and Valsesia, F. (2015). The power of repetition: repetitive lyrics in a song increase processing fluency and drive market success. *J. Consum. Psychol.* 25, 187–199. doi: 10.1016/j.jcps.2014.12.004

Pachet, F. (2012). "Hit Song Science," in *Music Data Mining*, eds T. Li, M. Ogihara, and G. Tzanetakis (Boca Raton, FL: CRC Press), 305–326.

Pachet, F., and Roy, P. (2008). "Hit song science is not yet a science," in *Proceedings of the 9th International Conference on Music Information Retrieval* (Philadelphia, PA), 355–360.

Pichl, M., Zangerle, E., and Specht, G. (2014). "Combining Spotify and Twitter data for generating a recent and public dataset for music recommendation," in

Proceedings of the 26th GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken) (Bozen).

Pichl, M., Zangerle, E., and Specht, G. (2015). "#nowplaying on #spotify: leveraging Spotify information on Twitter for artist recommendations," in *Current Trends in Web Engineering: 15th International Conference, ICWE 2015* (Rotterdam), 163–174. doi: 10.1007/978-3-319-24800-4_14

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rentfrow, P. J. (2012). The role of music in everyday life: current directions in the social psychology of music. *Soc. Personal. Psychol. Compass* 6, 402–416. doi: 10.1111/j.1751-9004.2012.00434.x

Schedl, M. (2010). "On the use of microblogging posts for similarity estimation and artist labeling," in *Proceedings of the 11th International Society for Music Information Retrieval Conference* (Utrecht), 447–452.

Schedl, M. (2013). "Leveraging microblogs for spatiotemporal music information retrieval," in *Advances in Information Retrieval: 35th European Conference on Information Retrieval, ECIR 2013* (Berlin; Heidelberg: Springer), 796–799. doi: 10.1007/978-3-642-36973-5_87

Schedl, M., Hauger, D., and Urbano, J. (2014). Harvesting microblogs for contextual music similarity estimation: a co-occurrence-based framework. *Multimedia Syst.* 20, 693–705. doi: 10.1007/s00530-013-0321-5

Schedl, M., and Tkalčič, M. (2014). "Genre-based analysis of social media data on music listening behavior: are fans of classical music really averse to social media?," in *Proceedings of the First International Workshop on Internet-Scale Multimedia Management, WISMM '14* (New York, NY: ACM), 9–13. doi: 10.1145/2661714.2661717

Shazam Entertainment, Ltd. (2016). "Shazam research dataset—offsets (SRD-O)," in *Stanford Digital Repository*. Available online at: http://purl.stanford.edu/fj396zz8014

Summers, C., Tronel, G., Cramer, J., Vartakavi, A., and Popp, P. (2016). "GNMID14: a collection of 110 million global music identification matches," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16* (New York, NY: ACM), 693–696. doi: 10.1145/2911451.2914679

Tsai, C.-G., Chen, R.-S., and Tsai, T.-S. (2014). The arousing and cathartic effects of popular heartbreak songs as revealed in the physiological responses of listeners. *Music. Sci.* 18, 410–422. doi: 10.1177/1029864914542671

Van Balen, J., Burgoyne, J. A., Wiering, F., and Veltkamp, R. C. (2013). "An analysis of chorus features in popular song," in *Proceedings of the 14th International Society for Music Information Retrieval Conference* (Curitiba), 107–112.

Van Balen, J. M. H., Burgoyne, J. A., Bountouridis, D., Müllensiefen, D., and Veltkamp, R. C. (2015). "Corpus analysis tools for computational hook discovery," in *Proceedings of the 16th International Society for Music Information Retrieval Conference*, 227–233.

Wang, A. (2003). "An industrial strength audio search algorithm," in *Proceedings of the 4th International Conference on Music Information Retrieval* (Baltimore, MD), 7–13.

Zangerle, E., Gassler, W., and Specht, G. (2012). "Exploiting Twitter's collective knowledge for music recommendations," in *2nd Workshop on Making Sense of Microposts (#MSM)* (Lyon), 14–17.

Zangerle, E., Pichl, M., Gassler, W., and Specht, G. (2014). "#nowplaying music dataset: extracting listening behavior from Twitter," in *Proceedings of the First International Workshop on Internet-Scale Multimedia Management, WISMM '14* (New York, NY: ACM), 21–26. doi: 10.1145/2661714.2661719

Zangerle, E., Pichl, M., Hupfauf, B., and Specht, G. (2016). "Can microblogs predict music charts? An analysis of the relationship between #nowplaying tweets and music charts," in *Proceedings of the 17th International Society for Music Information Retrieval Conference* (New York, NY), 365–371.

Conflict of Interest Statement: Authors BK and CB are present or former paid employees of Shazam Entertainment, Ltd. Authors FR and JB declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 05 April 2017
doi: 10.3389/fpsyg.2017.00431

Listening Niches across a Century of Popular Music

*Carol Lynne Krumhansl**

Department of Psychology, Cornell University, Ithaca, NY, USA

This article investigates the contexts, or "listening niches", in which people hear popular music. The study spanned a century of popular music, divided into 10 decades, with participants born between 1940 and 1999. It asks about whether they know and like the music in each decade, and their emotional reactions. It also asks whether the music is associated with personal memories and, if so, with whom they were listening, or whether they were listening alone. Finally, it asks what styles of music they were listening to, and the music media they were listening with, in different periods of their lives. The results show a regular progression through the life span of listening with different individuals (from parents to children) and with different media (from records to streaming services). A number of effects found in previous studies were replicated, but the study also showed differences across the birth cohorts. Overall, there was a song specific age effect with preferences for music of late adolescence and early adulthood; however, this effect was stronger for the older participants. In general, music of the 1940s, 1960s, and 1980s was preferred, particularly among younger participants. Music of these decades also produced the strongest emotional responses, and the most frequent and specific personal memories. When growing up, the participants tended to listen to the older music on the older media, but rapidly shifted to the new music technologies in their late teens and early 20s. Younger listeners are currently listening less to music alone than older listeners, suggesting an important role of socially sharing music, but they also report feeling sadder when listening to music. Finally, the oldest listeners had the broadest taste, liking music that they had been exposed to during their lifetimes in different listening niches.

Keywords: dehumanization, reminiscence bump, music technology, popular music, music and emotion, age cohort, music decade

OPEN ACCESS

Edited by:
Frank A. Russo,
Ryerson University, Canada

Reviewed by:
Matthew Woolhouse,
McMaster University, Canada
Annabel Joan Cohen,
University of Prince Edward Island, Canada

***Correspondence:**
Carol Lynne Krumhansl
clk4@cornell.edu

Specialty section:
This article was submitted to Cognition, a section of the journal Frontiers in Psychology

Received: *30 October 2016*
Accepted: *08 March 2017*
Published: *05 April 2017*

Citation:
Krumhansl CL (2017) Listening Niches across a Century of Popular Music. Front. Psychol. 8:431. doi: 10.3389/fpsyg.2017.00431

INTRODUCTION

The survey reported in this article seeks to characterize the contexts, or "listening niches", in which people hear popular music throughout their lifetimes. It is an extension of a study that investigated autobiographical memories and life-long preferences for music in young adults (Krumhansl and Zupnick, 2013). That study used top *Billboard* hits from five-and-a-half decades, 1955–2009. For each half decade, a clip was made with a compilation of short, recognizable segments of the top two hits from each year. Participants reported the percentage of songs from each half-decade that they recognized, how much they liked the songs, and how highly they rated the quality of the songs. They also reported their emotional response to the songs from each half decade. Finally, they reported whether they had personal memories associated with the songs and, if so, whether these memories were from listening with parents, alone, or with other people while growing up, or listening alone or with other people recently.

All these measures showed the typical increase for music released over the two decades of their lives, with the highest ratings for the music of the most recent half decade. This is consistent with previous studies showing preferences for music from late adolescence and early adulthood (Holbrook and Schindler, 1989; Schulkind et al., 1999; Janssen et al., 2007). More generally, the term "reminiscence bump" has been used to describe the peak in autobiographical memories and knowledge of events occurring during this period of people's lives (Rubin et al., 1986). However, we found an unexpected effect in as much as the same measures peaked for the music of their parents' late adolescence and early adulthood, music of the 1980s. In other words, they were familiar with, and liked, the music that was popular when their parents were the same age as they are now. We knew from their reports that they were listening to the music of the 1980s with their parents, but were not listening to it currently. We called the effect the "cascading reminiscence bump".

These results suggested it would be interesting to investigate in more detail the contexts in which people of different birth cohorts have listened to and developed preferences for music throughout their lives. The sample includes nearly 1900 participants born between 1940 and 1999, divided into six birth cohorts, those born in the 1940s, 1950s, 1960s, 1970s, 1980s, or 1990s. A short segment was extracted from the most popular song from each year from 1910 to 2009 (based on Whitburn, 1999, for years before 1955, and the *Billboard's* year-end Hot 100 chart for years since). Ten excerpts were joined together to form a clip for each of 10 decades: 1910s, 1920s, . . ., 1990s, 2000s.

For the clip of music from each decade, the participants reported whether they knew the songs, whether they liked the songs, what their emotional reactions to the songs were, and whether they had they had personal memories associated with the songs. If so, they were asked how specific the memory is and with whom they were listening. Because the sample of participants varied widely in age, the choices included parents, siblings and other family members, friends and peers, spouses or partners, children, and listening alone. To understand more about the contexts in which they were listening to music, they were asked what styles of music they were listening to during three periods of their lives: growing up, ages 18–25, and now. For the same three periods, they were also asked with what music media they were listening. Because the music spanned a century, the choices included radio, record, tape cassette, dances and parties, concerts, performed by others or by themselves, CDs, and various digital media other than CD, such as digital download and streaming.

Music information systems currently being developed promise new insights into how music is consumed, chosen and distributed, who listens to what styles of music, and how people share information about music with one another. Potentially, this kind of information may provide new information about fundamental issues that have been studied in music psychology. These issues include which aspects of musical structure contribute to memory and preference (e.g., Krumhansl, 1990; Narmour, 1990; Pearce and Wiggins, 2012), how personality traits and context affect musical choices (e.g., Hargreaves and North, 1997; Rentfrow and Gosling, 2003; Gabrielsson, 2011), and the nature of and mechanisms generating musical emotions (e.g., Blood and Zatorre, 2001; Sloboda and O'Neill, 2001; Krumhansl, 2002; Juslin and Västfjäll, 2008; Eerola and Vuoskoski, 2010). Practical insights about the therapeutic use of music and the value of music in public and private spaces may also derive from the analysis of large-scale data on music and its uses.

In particular, streaming services, such as Pandora and Spotify, would seem to greatly expand the amount of data on musical behaviors potentially available. Spotify, in particular, stresses a data-based culture for understanding music behavior, consumption, and choice. These services offer access to huge libraries of music and provide tools to aid listeners' discovery of new music. Luck (2016) identified psychological factors that make such services attractive, including freedom from ownership responsibility, enhanced discovery and emotional engagement, and nostalgia-fulfilment. However rich the potential of such information, there are limitations. A poll conducted by CivicScience in 2015 showed that 45% of Pandora and 62% of Spotify active users are less than 30 years old[1]. In addition, given the emphasis on discovering new music, the services tend to feature recent, innovative styles. It is hoped that the results of this broad, retrospective survey reported here can be seen as complementing what we can learn from contemporary music information systems.

MATERIALS AND METHODS

Stimulus Materials

Appendix A lists the 100 songs that were used to make up the 10 clips that the listeners heard. For the years 1910–1954, before *Billboard* magazine began publishing the year-end Hot 100 chart, the song that was used in the clip was the top single listed in Joel Whitburn's (1999) *A Century of Pop Music*. His criteria for choosing the top single varied depending on the year. The number of sources and the size of the charts varied, but for each year Whitburn listed the total number of weeks the song appeared on any one of the charts. We chose for each year the song that charted for the greatest number of weeks. For the years 1955–2009, the song was the top single from every year-end Hot 100 chart[2]. These more recent Billboard charts are compiled from national samples of radio air-play, top 40 radio playlists, retail sales and, more recently, internet sales reports.

There were 10 clips, each spanning a 10-year period, with an excerpt from the top song for each year. The excerpts were taken from the songs' choruses to maximize recognition. Thus, there were a total of 10 songs per clip for each of 10 music decades (1910–1919, 1920–1929, . . ., 2000–2009). Musical clips averaged 56.6 s ($SD = 18.89$). A practice clip consisted of the second most popular songs from 1955 to 1964. All excerpts were recorded from Spotify's streaming music service with the exception of a

[1]https://www.emarketer.com/Article/Pandora-Maintains-Strong-Audience-Lead-Over-Spotify/1012476

[2]http://en.wikipedia.org/wiki/List_of_Billboard_YearEnd_number_one_singles_and_albums#cite_note-221

couple from the 1910-1919 era, which were taken from Internet Archive[3].

Procedure

The experiment was designed with the *Qualtrics* research suite of tools and participants linked to the questionnaire by way of the Cornell Music Cognition[4]. **Appendix B** lists the questions asked in the survey. After each clip, participants reported the percentage of songs they recognized and how much they liked the songs. All responses were given on a Likert-type scale (0–10), except for the percent recognized (0–100). Participants also rated their emotional responses: sad, happy, nostalgic, romantic, and energized (with 0 = *Does not describe my feelings*, 10 = *Describes my feelings*). Next, they were asked if they would choose to hear similar songs, if given the opportunity. This was included to be a measure of the appeal of the songs from that decade independently of whether or not they were previously familiar with them. Finally, participants reported whether they had personal memories associated with the music. If so, then they were asked how specific are the memories on a scale from 0 to 10, from what period in their life (childhood up to 13 years old, teens ages 13–19, 20s, 30s, 40s, ages 50–65, over 65) and in what social context (listening alone, with parents, spouse/partner, children, siblings or other family members, and friends or peers). For these, they could select all that apply. They first made these responses with the practice clip, and then the 10 clips for each of the 10 decades which were presented in random order.

Following the ratings of the music clips, the participants answered a number of demographic questions: gender, year born, year mother born, year father born, years when children (if any) were born, their nationality, and the country in which they are currently living and, if they were living in the USA, for how many years.

Finally, a number of questions inquired about their music listening histories for each of three periods of their lives: growing up, ages 18–25, and now. For each of these periods, they indicated how many hours they listened to these styles: pop and rock, rhythm and blues, country and folk, classical, jazz, ethnic and world, and other. Then, for the same period they indicated where they heard popular music with these options: radio, record, tape cassette, dances and parties, concerts, heard performed by family and friends, played myself, CDs, subscription services (e.g., Spotify, Rhapsody, etc.), YouTube, Internet radio (e.g., Pandora), digital download (e.g., mp3), and other. They could select all that apply. They answered all of these questions for growing up, before proceeding to ages 18–25, and then they finally answered these questions for now. The protocol was approved by the Cornell University Institutional Review Board. Participants volunteered, granted their informed consent to record their responses, and were not compensated.

Participants

1910 (729 Males, 1181 Females) participants voluntarily completed the questionnaire. After the publication of Krumhansl and Zupnick (2013), the results were covered in various press media worldwide. The link to Cornell Music Cognition[4] was included in the NPR coverage[5], which is most likely the major source of participants, especially the older participants living in the USA. The majority (1085) were living in the USA, but more than 100 participants came from the Netherlands (268), Mexico (183), and Croatia (139), and it was not possible to determine how they found the link to the questionnaire. The questionnaire was discontinued and the data were compiled in October 2013.

The birth years of the participants ranged from 1928–2001. For the statistical analyses, there were enough participants born in each of six decades: 1940–1949 ($N = 64$), 1950–1959 ($N = 214$), 1960–1969 ($N = 243$), 1970–1979 ($N = 392$), 1980–1989 ($N = 601$), and 1990–1999 ($N = 384$). This gives a total number of 1899 participants included in the data analysis. They will be identified in the figures by the midpoint of the decade of their birth, for example 1945 for those born in the decade 1940–1949, and they will be referred to as the 1940s cohort. For the participants currently residing in the USA, their average birth year was 1973. The average birth year of those currently living outside the USA was 1981. When analyzed separately, it was difficult to separate effects of current residency from effects of age differences, so the two groups will not be separated in the statistical analyses that are reported. The average age of their father when they were born was 30.7 years (range 29.0–32.0), with the youngest fathers for the 60s and 70s cohorts. The average age of their mother when they were born was 28.1 years (range 26.5–29.4), with the youngest mothers for the 60s and 70 cohorts.

Figure 1 shows the number of hours per week the participants listened to different styles of music. As can be seen, for participants in all cohorts and all three spans of their lives, the most hours were spent listening to rock and pop music. Thus, the focus on *Billboard* top hits in the study was appropriate given their listening histories. The distribution of hours listening across the three time periods of their lives was quite consistent; the correlation between the distributions growing up and ages 18–25 was $r(5) = 0.97$, between growing up and now was $r(5) = 0.95$, and between 18 and 25 and now was $r(5) = 0.95$. Despite these general patterns, some differences were found between the cohorts. The older cohorts listened more to classical, country and folk, and rhythm and blues, whereas the younger cohorts listened more to ethnic and world music, and music that did not fall in any of the categories listed in the questionnaire.

RESULTS

Age and Who Was in the Listening Niche

The first analysis was undertaken to get an overview of who was in the participants' listening niches at different periods of their lives. The data used in the analysis were, for each of six cohorts,

[3]https://www.archive.org

[4]http://music.psych.cornell.edu

[5]http://www.npr.org/sections/health-shots/2013/09/05/219278386/turns-out-your-kids-really-did-love-that-music-you-played

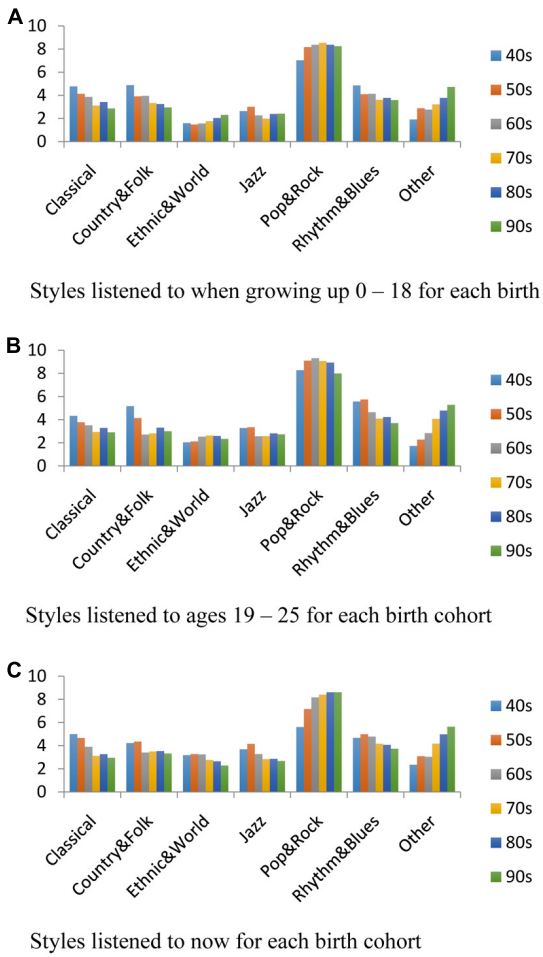


FIGURE 1 | (A) Styles of music listened to when growing up for the participants born in each decade, the birth cohorts. **(B)** Styles listened to when ages 19–25. **(C)** Styles listening to now.

how much they were listening to the music of each of 10 music decades (6 cohorts × 10 music decades). This was found for the different periods of their lives (0–12 years, 13–19 years, 20–29 years, 30–39 years, 40–49 years, and 50–65 years); the data for listening when over 65 was too sparse to include. The same data (6 cohorts × 10 music decades) were compiled for whom they were listening to the music with (parents, siblings and other family members, with friends or peers, with spouse or partner, with children, or alone).

Figure 2 shows the results of a principal components analysis done on these data. The arrows point in similar directions if they were listening to similar music at these times of their lives with these individuals. It shows that when the participants were ages 0–12, they were most often listening to music with their parents, by ages 13–19, they were listening more with siblings and other family members. Then later, through their 20s, they were more often listening alone or with friends and peers. By ages 30–39, music was listened to with spouse or partner, and then with children for participants in their 30s and 40s. The first (horizontal) dimension accounted for 48.1% of the variance in the data; the second (vertical) dimension accounted for 32.3% of the variance, for a total of 80.4% of the variance. Overall, the results suggest a regular progression of listening with different groups of people throughout the life span ranging from parents in early life to children in later life.

Song Specific Age

The next analysis looked at the liking ratings as a function of the participants' age at the time the music was popular, the "song specific age" (Holbrook and Schindler, 1989). It was calculated as the approximate age they were when the song was popular. For example, the song specific age for the cohort born in the 1960s and the music of the 1980s was 20. The analysis was also done on 5-year cohorts, with similar results and will not be reported.

The results showed an increase in how much they liked the music up to the age of about 20 and then a decrease for music that was popular later in their lives. This was confirmed by a polynomial regression which accounted for 62% of the variance [$F(2,57) = 46.9$, $p < 0.0002$] and both the linear and quadratic effects were significant [$F(1,57) = 45.3$ and 48.4, respectively, both $p < 0.0001$). Overall, liking ratings were lowest for the songs that were popular long before the participants were born, and for the most recent songs for those in the oldest age cohort.

However, a closer look showed notable differences between the three oldest cohorts (40s, 50s, 60s) and the three youngest cohorts (70s, 80s, 90s). The liking ratings for the two groups as a function of the song specific age are shown in **Figure 3**. It is apparent that the song specific age effect is stronger and more regular for the older cohorts than for the younger cohorts; the peak is more distinct and occurs somewhat later for the older cohorts than the younger cohorts.

Music Decade

The next analysis considered whether there were overall preferences for different decades of music. To look at this, the decade of music was added to the analysis of variance with linear and quadratic effects of song specific age (as above). In other words, the analysis looked to see whether once the effect of song specific age was factored out there was a residual effect of the decade of the music. The analysis with both the song specific age and decade accounted for 86% of the variance in the liking ratings [$F(11,48) = 26.6$, $p < 0.0001$] and the effect of decade was highly significant [$F(9,48) = 9.0$, $p < 0.0001$].

There were peaks for music popular in the 1940s and in the 1960s. A contrast comparing music from the 1940s to the music from the 1930s and 1950s produced a marginally significant effect [$F(1,48) = 3.5$, $p = 0.066$, which would be significant by a one-tailed test]. A contrast comparing music from the 1960s to the music from the 1950s and 1970s produced a significant effect [$F(1,48) = 10.4$, $p = 0.0023$]. Thus, the peaks for music of the 1940s and 1960s were confirmed statistically. A contrast was also computed testing whether the average liking ratings for music of the 1980s exceeded that for the 1970s or 1990s because the earlier paper (Krumhansl and Zupnick, 2013) found a peak for

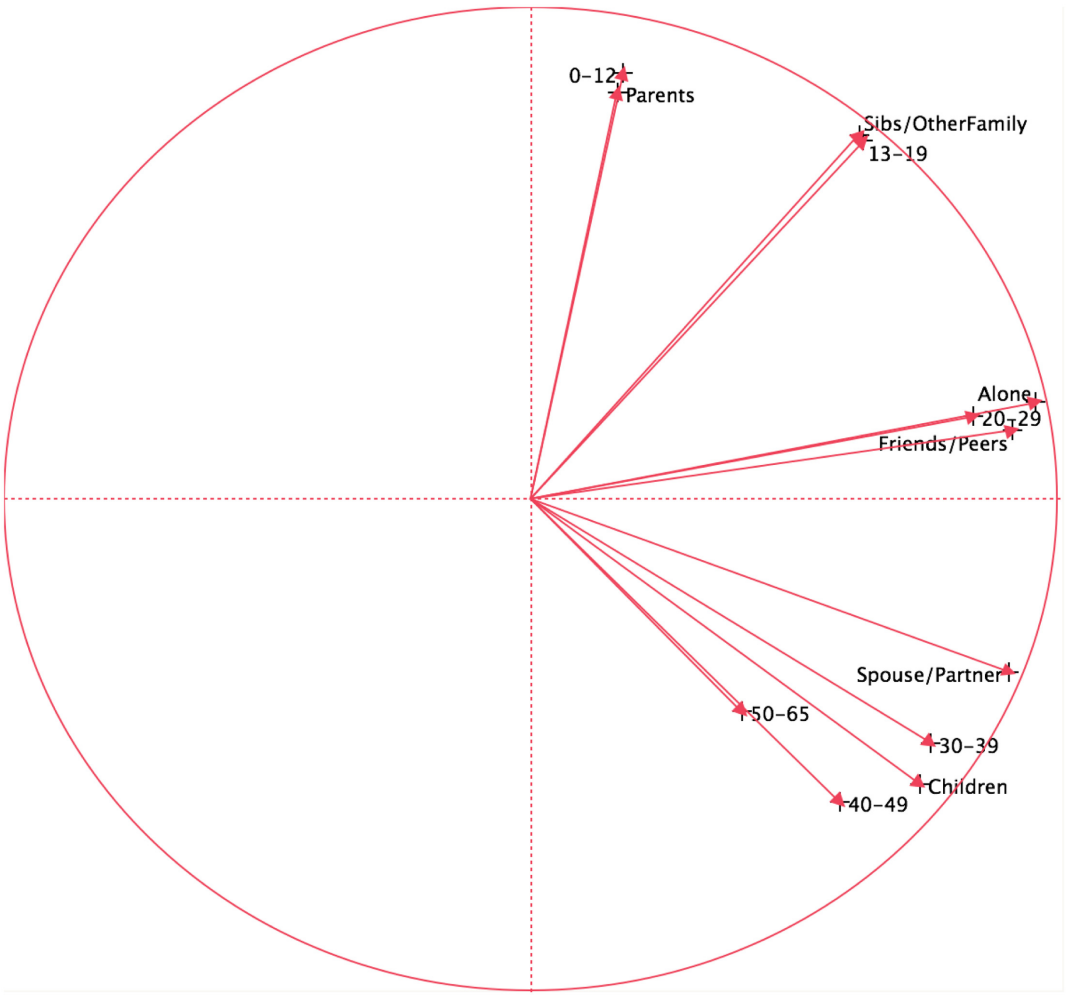


FIGURE 2 | Principal components analysis of music listened to at different ages and with whom.

the music of the 1980s in college age participants. The result was non-significant [$F(1,48) = 1.74$, $p = 0.19$].

However, as can be seen in **Figure 4** the decade of music effect was stronger for the younger cohort than the older cohort. Their liking ratings showed clear peaks for the music in the decades of the 1940s, 1960s, and 1980s. In contrast, the liking ratings for the older cohort were more evenly distributed with a broad peak around the music of the 1960s and 1970s, which is consistent with the song specific age effect described earlier.

Emotional Reactions

Figure 5 shows the emotional reactions to music of the different decades. There was a significant effect of decade for all the emotion scales, with the weakest effect for sad [energized $F(9,50) = 21.1$, $p < 0.001$, happy $F(9,50) = 18.3$, $p < 0.001$, nostalgic $F(9,50) = 7.7$, $p < 0.001$, romantic $F(9,50) = 12.0$, $p < 0.001$, sad $F(9,50) = 2.8$, $p = 0.01$] For all the scales (except sad) there was an increasing trend from the earliest decade to the music of the 1980s, and then a decrease. For sad, a test comparing means showed that the only significant difference is between the 1910s (the saddest) and the 2000s (the least sad). Distinctive peaks relative to neighboring decades can be seen in the curves for happy, nostalgic, and energized for music of the 1940s, 1960s and 1980s (except for nostalgia, possibly because the music is relatively recent).

The next analysis considered how much the emotional reactions accounted for how well they liked the music. A multiple regression predicting liking from these five emotional responses accounted for 99.2% of the variance [$F(5,54) = 1346.9$, $p < 0.0001$], which indicates that the emotional reaction to the music is a very strong predictor of how well the music is liked. Each of the five emotions was significant in the multiple regression [energized $F(1,54) = 5.20$, $p = 0.03$; happy $F(1,54) = 42.7$, $p < 0.0001$, nostalgic $F(1,54) = 40.2$, $p < 0.0001$, romantic $F(1,54) = 16.7$, $p < 0.0001$, sad $F(1,54) = 38.0$, $p < 0.0001$], suggesting they are each making independent contributions to how well the music is liked. The regression coefficient for all of the emotions except sad was positive, suggesting that sadder popular music is less preferred. It should

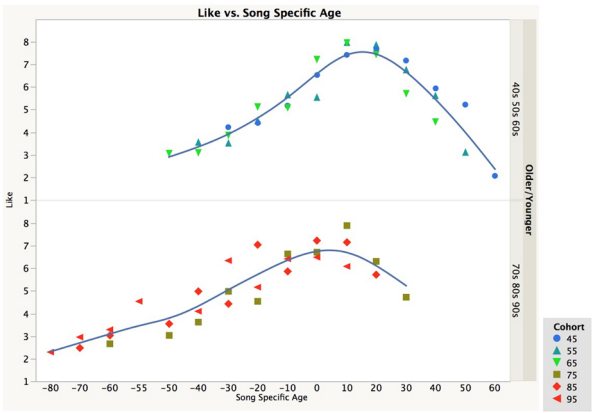


FIGURE 3 | Plot of liking ratings as a function of song specific age (the age of the participant when the song was popular), divided up between the three oldest and three youngest birth cohorts.

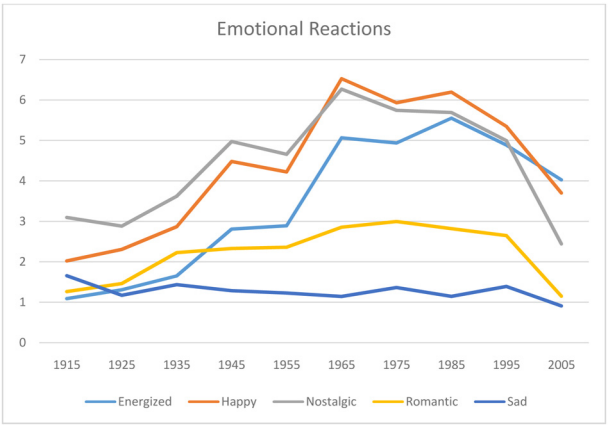


FIGURE 5 | The emotional reactions to the music of the different decades.

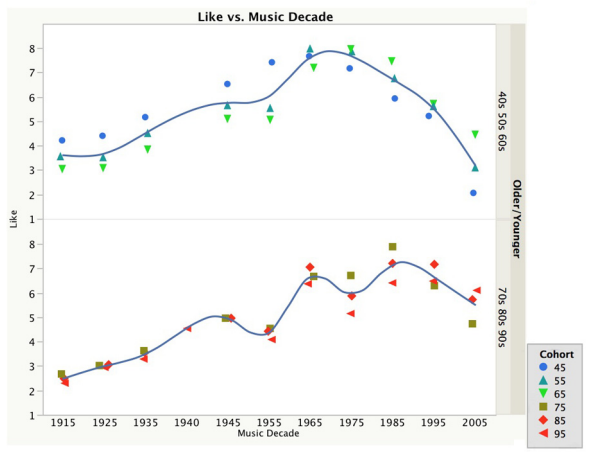


FIGURE 4 | The liking ratings as a function of the decade of the music, divided up between the three oldest and three youngest birth cohorts.

be noted, however, that the music from none of the decades is rated highly on sad.

Because the liking ratings might be influenced by whether the participants recognized the songs (and the correlation between the two was, in fact, $r(58) = 0.92$, $p < 0.0001$), the survey included another question about whether they would choose to hear music like that in each decade again. The correlation with whether they recognized the music and whether they would like to hear music like that again was still fairly strong [$r(58) = 0.87$, $p < 0.0001$]. However, there was a possibly interesting difference in the emotions that predicted whether they said they would like to hear music like that again. The five emotion ratings accounted for 98.2% of the variance [$F(5,54) = 607.4$, $p < 0.0001$], but only happy and romantic contributed positively [happy $F(1,54) = 97.4$, $p < 0.0001$, romantic $F(1,54) = 6.9$, $p = 0.011$] and energized contributed *negatively* [$F(1,54) = 12.9$, $p = 0.0007$]; the other two scales were marginally significant and in the same direction as before. Thus, hearing music that makes the participants feel energized made them less likely to want to hear music like that again.

The final analysis considered whether the different birth cohorts had different emotional reactions to the songs of different decades. Even though the younger participants didn't know the older songs and the older participants didn't know the most recent songs, they agreed on their emotional reactions to the music. To look at this statistically, for each birth cohort, an emotion profile was made of the five emotion scales for the 10 decades of music. For example, the emotion profile for the 40s cohort was the rating on the five emotion scales for all 10 decades of music, for a total of 50 values. The correlations between the emotion profiles for all pairs of cohorts were highly significant (at $p < 0.001$, when Bonferonni corrected for multiple comparisons). This might be an artifact of the low ratings on sad, so the same analysis was done after that scale was excluded and the correlations between all pairs of cohorts were still highly significant (except for the correlation between the oldest and the youngest cohorts when corrected for multiple comparisons).

Personal Memories

Overall, 53.6% of the participants reported having personal memories associated with the songs in the 10 decades, and those memories were rated an average of 5.73 on specificity (0–10). There was no effect of birth cohort on either the percent of associated memories or their specificity. Both measures correlated strongly with whether they liked and recognized the songs, and wanted to hear songs like that again. Listeners' reported memories correlated most strongly with music they heard when they were 13–19 years old [$r(58) = 0.81$, $p < 0.0001$) and 20–29 years old ($r(58) = 0.83$, $p < 0.0001$], although how much they listened to music from all periods of their lives (except ages 50–65) correlated significantly with the proportion of people reporting associated memories; the same was true for the specificity of the memory.

The incidence of personal memories was also associated positively with the song decades that were rated high on making them feel energized, happy, nostalgic, and romantic [$r(58) = 0.94$, $r(58) = 0.95$, $r(58) = 0.86$, $r(58) = 0.66$, respectively, all $p < 0.0001$], and negatively on sad [$r(58) = -0.29$, $p = 0.023$]. The proportion of participants reporting personal memories correlated most strongly with music they heard listening alone [$r(58) = 0.96$, $p < 0.0001$] and with friends and peers [$r(58) = 0.91$, $p < 0.0001$], the music they heard most often in their teens and early adulthood, but the correlations were significant for all periods of their lives. As for how specific the memories were, the ratings correlated most strongly with music they heard listening alone [$r(58) = 0.90$, $p < 0.0001$] and with friends and peers [$r(58) = 0.91$, $p < 0.0001$], but the correlations were significant for all of the music they listened to with others except for music they listened to with parents.

Music Media

Participants also indicated which media they were using when listening to music during three periods of their lives: growing up, 19–25 years, and now. **Figure 6** shows the percentage of people in each birth cohort who were listening to music on the most common media: concerts, parties, radio, records, tape, CDs, and Digital. Digital was the composite of digital download, YouTube, internet radio, and subscription services. The responses for "played myself" were not included because of the ambiguity of the question: whether they were performing it themselves, or playing a recording of someone else performing the music.

For all periods of their lives, they were listening to music on radio at a fairly high level although note the decreasing use of radio presently. The youngest birth cohort is listening to music almost as much in digital formats. Beyond that, we see effects of the period of their lives that relate to music media. Growing up, the older participants were listening to music on records, whereas younger birth cohorts were listening to music on tape, and the youngest on CDs and in other digital formats. For music in late teenage and early adult years, the oldest listeners were hearing music on records, but also tapes; the middle birth cohorts had clearly switched to tape, and the youngest participants were listening to music on CDs and on digital media. Finally, nearly no one is listening to music on records or tapes now, but more on CDs and other digital formats, even including the oldest birth cohorts. Finally, participants seem to have heard music at concerts and parties most often when they were ages 19–25 years.

Differences between Birth Cohorts

The results described above showed that the decade effect (preferences for music of the 1940s, 1960s, and 1980s) was stronger for the younger generations and the song specific age effect (with a peak in preference for music popular in late teens and early 20s) was stronger for the older generations. When looking for other differences between the birth cohorts, some obvious effects emerged. For example, the younger cohorts were less familiar with the older music and liked it less than the more recent music; the opposite was true for the older cohorts. Three less obvious findings emerged, however.

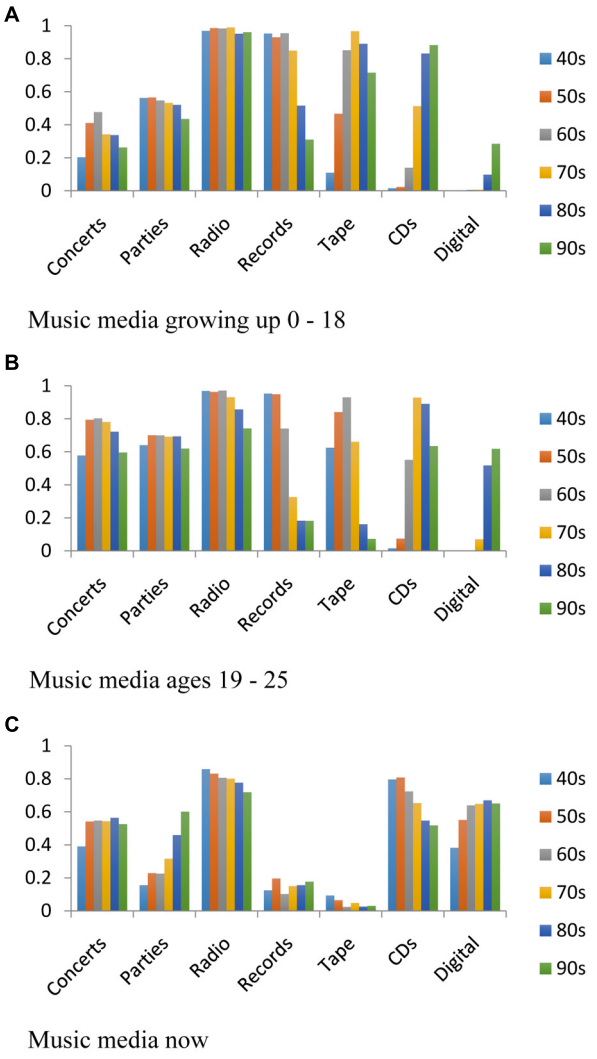


FIGURE 6 | (A) The music media used when growing up. **(B)** Media used ages 19–25. **(C)** Media used now.

One finding concerned the overlap between the music they listened to with their parents and their friends and peers. **Figure 7** shows for each cohort the decades of the music they listened to with their parents and their friends and peers. The oldest three birth cohorts listened to the older music with their parents and the newer music with their friends, with very little overlap. When it comes to the cohort born in the 1970s, we start to see them listening to the older music with their parents, particularly the music of the 1940s and 1960s, and only the newer music, the music of their early adulthood, with their friends and peers. This pattern became stronger for the birth cohorts from the 1980s and 1990s.

As described earlier, there was a predictable pattern of who was listening to music with the participants as they moved through different stages of their lives, from parents to children. However, there was a somewhat surprising effect of birth cohort on how much they listened to music alone. **Figure 8A** shows

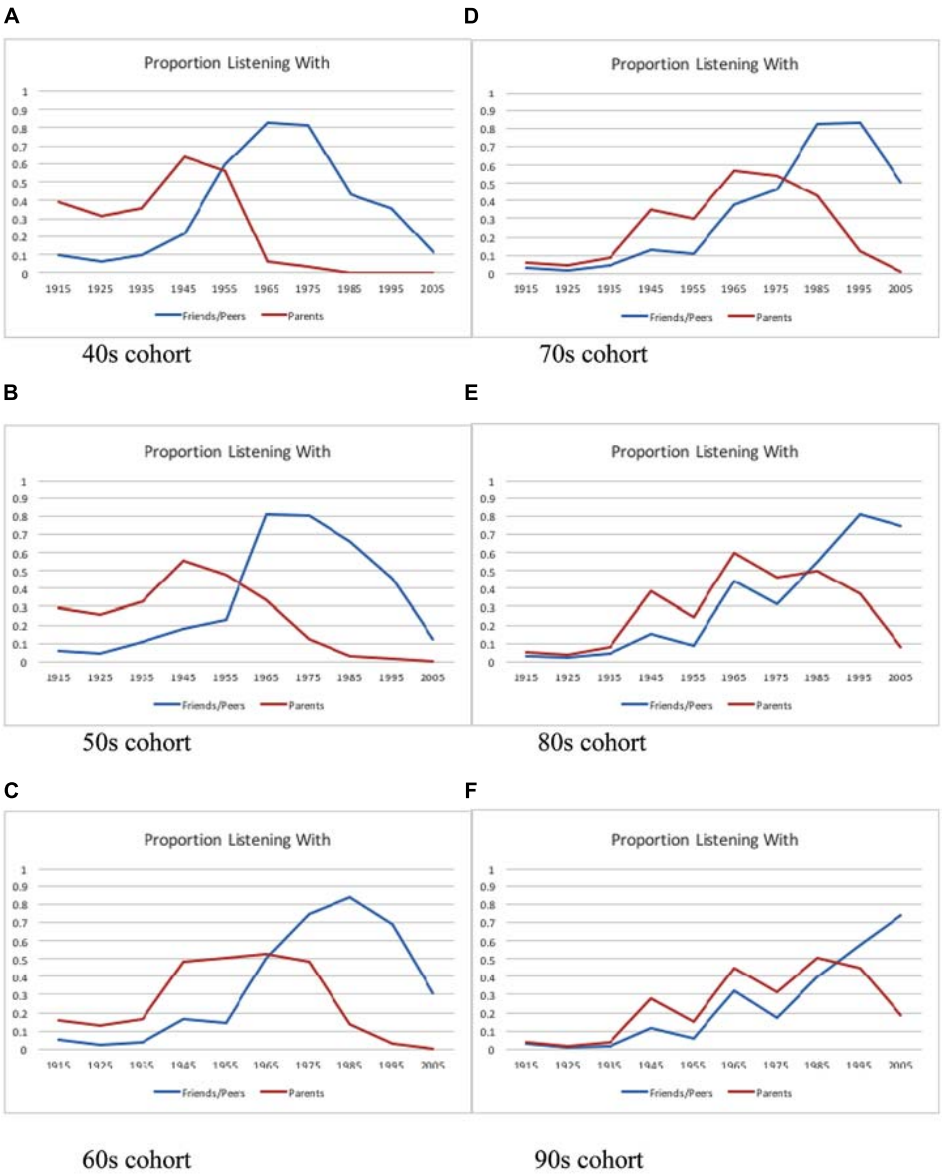


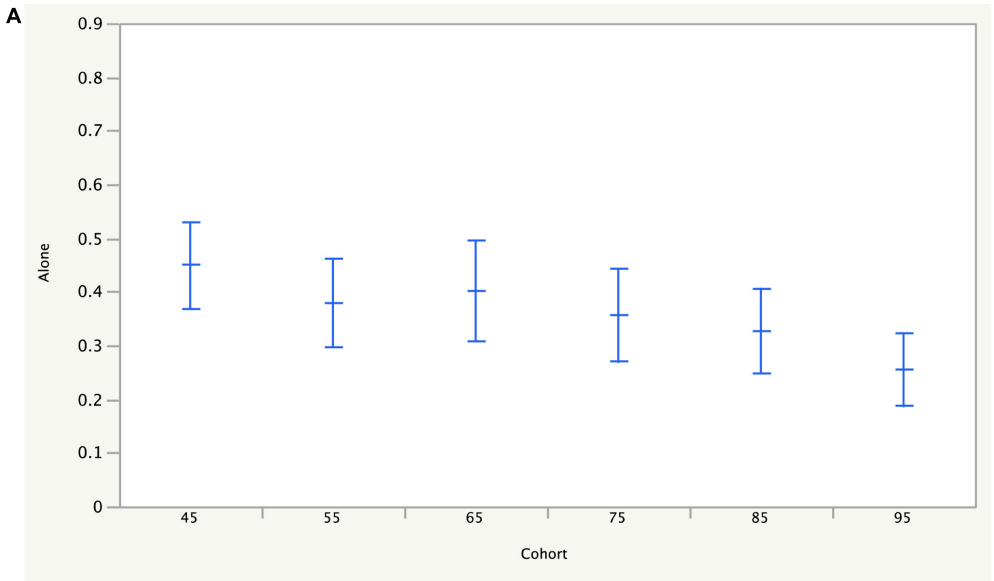
FIGURE 7 | The decade of the music listened to with friends and parents for participants born in each decade. (A) 40s cohort; **(B)** 50s cohort; **(C)** 60s cohort; **(D)** 70s cohort; **(E)** 80s cohort; **(F)** 90s cohort.

the percentage of people in the different birth cohorts who reported listening to music alone, showing a decline for younger participants. When decade of music was included in the analysis, a linear contrast found a quite significant decreasing effect of birth cohort [$F(1,54) = 10.9, p = 0.0017$] on how much they were listening to music alone. Given the prevalence in more recent years of personal listening devices, one might have expected the opposite effect.

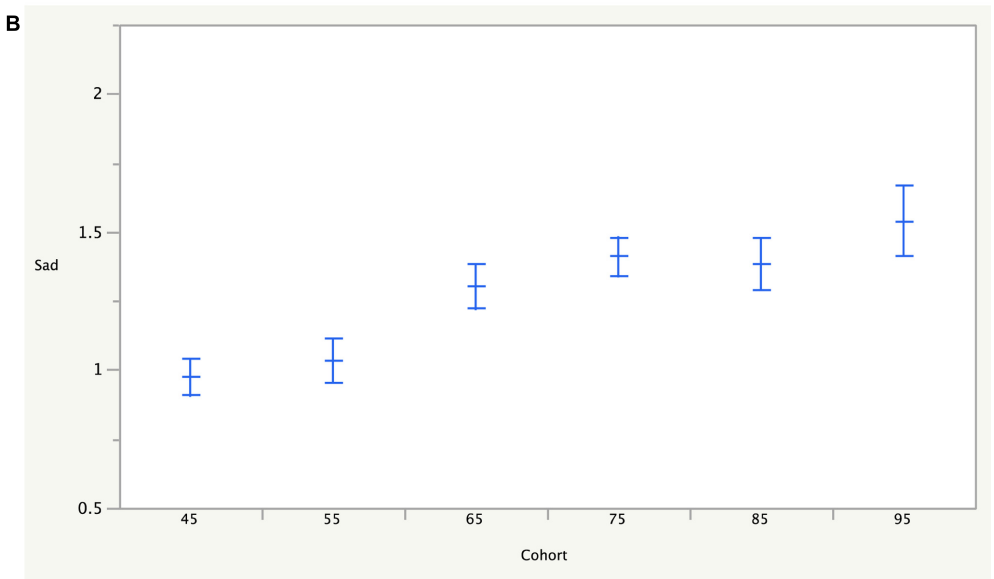
The final effect concerned the different cohorts' overall emotional responses to the music. No significant effect of birth cohert was found on any of the emotion scales, with the exception that the younger birth cohorts generally gave notably higher ratings on sad. As can be seen in **Figure 8B**, the younger birth cohorts judged the music of all decades to make them feel sadder than the older birth cohorts [$F(1,58) = 29.8, p < 0.0001$].

Cumulative Effects of Listening Niches on Musical Preferences

Figure 9A graphs how much each birth cohort liked the music of each decade. As can be seen, those born in the 1940s had a broader liking curve than any of the other birth cohorts. This may be because they have, over their lifetimes, listened to music with more different types of people. **Figure 9B** shows, for each decade of music with whom they were listening. With their parents, they were listening to the music of the 1940s, during the decade in which they were born. They were also listening to music of the



Percentage of each birth cohort reporting listening alone



Ratings of feeling sad for each birth cohort

FIGURE 8 | (A) Shows for each birth cohort how much music they listened to music alone (average and mean error bars). **(B)** Ratings of each cohort of feeling sad while listening to music (average and mean error bars).

1910s, music that their parents would have been listening to with *their* parents in the decade in which they were born, that is, with our participants' grandparents. With siblings and other family members, they were listening to this same music and also to the music of the 1950s, music that was contemporary when they were young. The listened alone most to music of the 1950s, 1960s, and 1970s, the music of their teen years, 20s, and 30s. With friends and peers, they listened most to music of the 1960s and 1970s, in their teen and early adult years. Overall this birth cohort listened to music most often during this period of their lives. They listened with spouse or partner most to music of the 1960s and 1970s, when in their 20s and 30s. And, finally, they listened with their children most to music of the 1980s, when their children would have been in their teens. Note that after this they were listening relatively little to newer music, especially music of the 1990s and 2000s. Note that in **Figure 9A**, they liked the music of these decades least, the music that they also were not hearing in any of their listening niches.

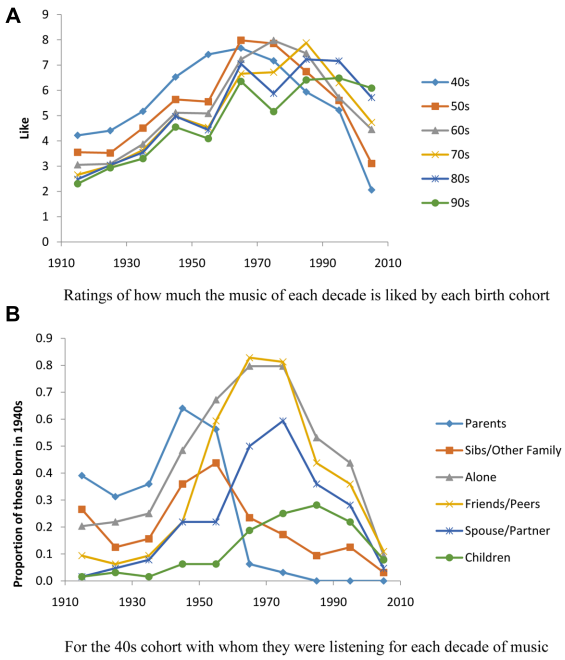


FIGURE 9 | (A) The liking ratings of the birth cohorts for each decade of music. **(B)** For the generation born in the 1940s, with whom they were listening for each decade of music.

DISCUSSION

The main objective of the present study was to gain a more detailed understanding of the contexts in which people listen to and develop preferences for music. One important component of the "listening niches" was with whom they were listening. Given the wide range of ages of the participants, it was possible to trace a regular progression throughout the life span: they were listening with parents as children, with siblings and other family members in their teen years, with friends and peers and alone in their twenties, with spouse or partner in their thirties, and finally with children when they were in their forties.

The present study replicated the song specific age effect found in many studies (e.g., Holbrook and Schindler, 1989; Schulkind et al., 1999; Schubert, 2016). The effect is an overall preference for songs that were popular in late adolescence and early adulthood. A recent study by Rathbone et al. (2016) found, however, that the reminiscence bump was pronounced only if the music was personally significant to the listener. Other factors that might contribute to the reminiscence bump found for music (and also for other domains, such as public events, sports, and films) include the occurrence of personally significant events during these years, physiological changes, formation of personal values, and music as a badge of social identity (see Rubin et al., 1998, for a review).

Another effect found in this study was a decade effect. Music of the 1940s was preferred to music of its neighboring decades (i.e., the 1930s and 1950s), and the same was true for music of the 1960s. The music of the 1980s also showed a peak, but it was different from its neighbors only for the younger participants. To try to understand the decade effect, the emotional responses to music of the different decades were considered. Consistent with the decade effect, the music of the 1940s and 1960s was judged to make the participants feel happier and more energized and nostalgic than the music of their neighboring decades. The same was true of the music of the 1980s, although the effect of nostalgia was somewhat muted possibly owing to its relative recency. These results are in line with the finding that popular music is generally judged to be positive in both valence and arousal (e.g., Platz et al., 2015).

In general, the popular music used here was not judged to be sad, except perhaps for the oldest decade of music, the 1910s. Schellenberg and von Scheve's (2012) analysis of 1000 Top 40 recordings found an increase over the period from 1965 to 2009 of minor mode and slower tempo. Consistent with this, the minor mode songs in the present study were predominantly from the most recent decade. However, it might be noted, the study by Platz et al. (2015) did not support the shift to sadder songs over this period in German popular songs; their study included music from the period 1930 to 2010. Schellenberg and von Scheve (2012) hypothesized that this shift to minor mode and slower tempo would make more recent songs sound sadder, although they did not test this empirically. Our participants did not rate the more recent music as making them feel sadder than earlier songs (in fact, none of these top hits were rated as making them feel sad), but they did rate the more recent songs as making them feel less energized, happy, nostalgic, and romantic.

The same influences of emotion were found for personal memories associated with the songs: their incidence was positively related to songs that made them feel energized, happy, nostalgic, and romantic, and negatively to those that made them feel sad. Despite the century long span of the music, more than half the participants reported personal memories associated with the music in the study. This complements the finding that 30% of the time listeners in Janata et al. (2007) study had somewhat or strongly autobiographical memories associated with 1500 randomly selected popular songs. The prevalence and specificity of personal memories were greatest for music heard in the teens and 20s, but also came from all periods of their lives. They were most prevalent and specific for music heard with friends and peers, and alone, but were associated with all contexts, except for music listened to with parents possibly because autobiographical memory emerges gradually in development (Nelson and Fivush, 2004). The older participants judged their personal memories to be as specific as the younger participants, but it should be noted that there are general shifts from episodic to semantic details in autobiographical memories with aging (Levine et al., 2002). Overall, these results are consistent with the frequency, durability, strength and rich content of autobiographical memories associated with music (e.g., Gabrielsson, 2001, 2011; Janata et al., 2007; Belfi et al., 2015).

The emotion rating scales almost perfectly predicted how well the music from the different decades was liked. However, other factors might be involved. In the 1940s, WW II made popular both songs that brought the war home and sentimental ballads for those remaining at home (Sanjek, 1988), which have

been absorbed into film and other popular media (Basinger, 2003). After the war, high-quality, low cost tape recorders helped establish independent labels broadening the musical styles available on recordings (Burgess, 2014). The 1960s was a time of political unrest and tremendous artistic innovation, including that of the Beatles and the Rolling Stones, but also Motown, country, folk and, late in the decade, disco and hip hop. The 1980s ushered in a conservative political era and saw the introduction of music videos on MTV, and influential albums by Michael Jackson, Madonna, Springsteen, Prince, and others. Burgess (2014) also details technical advancements in music production during these decades. It is impossible to assess from the current survey how influential these, and cultural and artistic factors, have been in establishing the participants' preferences and emotional responses.

The survey does, however, provide some information about the media the participants were using to hear popular music. They reported how they were hearing music during three periods of their lives: when they were growing up, when they were 19–25 years of age, and now. Radio has been a major source of music for all birth cohorts during all periods of their lives, although a decline was apparent for the youngest birth cohorts. In the 1940s, the transistor radio was invented, and car radios came in by the late 1940s. In the 1960s, radio developed the long-playing FM format, and AM radio innovated the Billboard Hot 100 in 1959. Internet radio was pioneered in the 1990s. Thus, radio in its various forms has been a constant source of music delivery for all the birth cohorts. Other music media have undergone shifts, however, and this might be a partial cue to the decade effect found.

Important changes in how music could be heard occurred in the 1940s, 1960s, and 1980s (Burgess, 2014). Columbia Records introduced the 33 1/3 RPM long playing record in 1948 with greatly improved signal to noise ratio and longer playing times. The survey found that records were the predominant music media (together with radio) while growing up for the cohorts born in the 1940s, 1950s, 1960s, and even the 1970s, suggesting that young participants were listening to their parents' music on their parents' media, records. However, by the time they were listening to music during ages 19–25 they shifted to the new media of the 1960s, tape. Phillips compact cassette was introduced in 1963, making it possible to listen to music almost anywhere and inexpensively sharing it with others. For the participants born in the 60s and 70s, tape was the predominant music media while growing up, again suggesting that they were listening to their parents' music on their parents' media, tape. But by the time they were listening to music during ages 19–25 they shifted to the new media of the 1980s, CDs. Sony and Philips introduced the CD format in 1983. For the cohort born in the 1960s, 1970s, and 1980s, tape was still the primary music media while growing up, again suggesting they were listening to their parents' music on their parents' media, tape. However, by the time they were 19–25 years of age, they were primarily listening with the new technology, music on CDs.

Stepping away from these particular results, one factor contributing to the preferences for music of the 1940s, 1960s, and 1980s may be the introduction of music media that were significant improvements over previous media. The most likely candidates, based on the survey results, are: long-playing records, cassette tapes, and CDs. While growing up, listeners appear to have heard the music of the previous birth cohorts on the older technologies, but actively sought new music on the new technologies in their teens and twenties. Perhaps it is during that period of their lives that they began building their own music collections in the new media, developing their musical preferences, and establishing associated personal memories and emotional responses. Radio has been a major source of music for all birth cohorts, although the digital formats (other than CD) seem to be overtaking radio for the youngest cohorts. An interesting question, given the adoption of streaming services with no physical musical artifacts (Luck, 2016), is whether intergenerational transfer of music will be less prevalent in the future, or whether the easy access to very large music libraries will actually facilitate sharing music across generations.

Finally, the study turned up some generational differences. Listeners born in the 1940s, 1950s, and 1960s listened to very different music with their parents and their friends. They listened to the older music with their parents, but more contemporary music with their friends. This is consistent with the idea that the older birth cohorts used music, particularly the music of the 1960s and 1970s, to distance themselves from their parents. In contrast, those born in the 1970s, 1980s, and 1990s listened to some of the older music with both their parents and their friends, especially music of the 1940s and 1960s and, for the youngest two birth cohorts, the music of the 1980s, replicating Krumhansl and Zupnick (2013).

Other generational differences were found. The oldest three birth cohorts showed a stronger effect of song-specific age, whereas the youngest three birth cohorts showed a stronger effect of the decade of the music. One possible explanation for this is that the older participants may generally have had less access to a wide variety of music. Other than music heard on radio, they would have had to purchase records, tapes, and CDs. In contrast, because the younger participants have had relatively easy access to a greater variety of music, they could freely sample music of widely different styles and eras, especially that from the preferred decades.

Another generational effect was that the younger participants tended to listen alone less than the older participants. One might have thought, with the availability of personal listening devices, they would be listening alone more. A survey done by Edison Research[6] found that listeners report friends and family were among the most important sources to keep up-to-date with music, together with AM/FM radio, suggesting they discover music by listening with others. The present finding that younger listeners listen alone less also fits with the idea that music sharing is used to as a way to convey information about ourselves to others (Rentfrow and Gosling, 2003, 2006; Lonsdale and North, 2009).

A somewhat surprising result was that the older participants generally found the music less sad than the younger participants.

[6]http://www.edisonresearch.com/the-infinite-dial-2016/

This may be because the older individuals tend to focus on more positive things in general (Mather and Carstensen, 2005), so that they might have focused on the more upbeat songs in each decade. Alternately, the effect might be specific to music, with older participants having had more experience with the older sadder music and thus responded less to the sad content in the songs while, conversely, the younger participants were more experienced with less happy music and were thus responding to the less happy content. Music is multivalent, in as much as it can express multiple emotions simultaneously (Krumhansl, 1997; Vines et al., 2005), so that the same piece of music might be, for example, happy, sad and nostalgic at the same time.

The final generational effect came from looking at the oldest birth cohort, those born in the 1940s, to see the cumulative effect of listening to music over approximately 70 years. This birth cohort had the most eclectic taste of all the cohorts, that is, they liked music from all periods of their lives except from the last two decades, as will be discussed below. The finding argues against the stereotype of that generation (mostly "baby boomers") has musical tastes confined to music of the 1960s. Although that music played a strong role in defining their identities, their musical tastes are considerably broader than just the music of their youth.

Schubert's (2016) younger participants reported their tastes broadening over time. The result for this older generation suggests that this process might continue well into the lifetime. This kind of "open-earedness" (Hargreaves, 1982) may be facilitated by the variety of listening niches the oldest participants have occupied. Listening with parents, siblings and other family members, friends and peers, spouse, or partner, and finally with children have given them broad exposure to, and developed their liking for, music of many decades. Cohen (2000) has suggested reduced plasticity with age makes it difficult to acquire the grammar of new styles of popular music, and this might be reflected in the steep drop off in preferences for the most recent music. It may also be that people in their 60s and 70s no longer typically occupy multigenerational listening niches.

The music industry is currently undergoing rapid changes in how music is produced, delivered, and shared between individuals. What will come of these changes is a question of great interest. If the present findings offer any guidance, various forces are likely to play a stabilizing role in future developments. One is that people move through a generally regular sequence of listening niches that are populated by different individuals and media over time. They adapt to new technologies in a gradual way. Musical tastes tend to broaden with age, and listening to music is a social activity with people sharing music recommendations with one another, increasingly across generations. All these forces, at least as they have operated over the last century, have produced systematic patterns of change over time despite the marked evolution of musical styles and technologies. Rather than creating ruptures in music listening patterns, periods of particularly rapid evolution have in fact resulted in enhanced preferences for, and emotional responses to, music from those decades.

AUTHOR CONTRIBUTIONS

CK designed the study, analyzed the results, prepared the figures, and wrote the manuscript.

ACKNOWLEDGMENTS

The author is grateful to Justin Zupnick for his assistance in designing the study. He created the stimulus materials and the survey. The helpful suggestions of the two reviewers are gratefully acknowledged.

REFERENCES

Basinger, J. (2003). *The World War II Combat Film: Anatomy of a Genre*. Middletown, CT: Wesleyan University Press.

Belfi, A. M., Karlan, B., and Tranel, D. (2015). Music evokes vivid autobiographical memories. *Memory* 24, 1–11. doi: 10.1080/09658211.2015.1061012

Blood, A. J., and Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proc. Natl. Acad. Sci. U.S.A.* 98, 818–823. doi: 10.1073/pnas.191355898

Burgess, R. J. (2014). *The History of Music Production*. New York, NY: Oxford University Press.

Cohen, A. J. (2000). Development of tonality induction: plasticity, exposure, and training. *Music Percept.* 17, 437–459. doi: 10.2307/40285828

Eerola, T., and Vuoskoski, J. K. (2010). A comparison of the discrete and dimensional models of emotion in music. *Psychol. Music* 39, 18–49.

Gabrielsson, A. (2001). "Emotions in strong experiences with music," in *Music and Emotion: Theory and Research*, eds P. N. Juslin and J. A. Sloboda (New York, NY: Oxford University Press), 431–449.

Gabrielsson, A. (2011). *Strong Experiences with Music*. New York, NY: Oxford University Press.

Hargreaves, D. J. (1982). The development of aesthetic reactions to music. *Psychol. Music* 10, 51–54.

Hargreaves, D. J., and North, A. C. (1997). *The Social Psychology of Music*. New York, NY: Oxford University Press.

Holbrook, M. B., and Schindler, R. M. (1989). Some exploratory findings on the development of musical tastes. *J. Consum. Res.* 16, 119–124. doi: 10.1086/209200

Janata, P., Tomic, S. T., and Rakowski, S. K. (2007). Characterisation of music-evoked autobiographical memories. *Memory* 15, 845–860. doi: 10.1080/09658210701734593

Janssen, S. M., Chessa, A. G., and Murre, J. M. (2007). Temporal distribution of favorite books, movies, and records: differential encoding and re-sampling. *Memory* 15, 755–767. doi: 10.1080/09658210701539646

Juslin, P. N., and Västfjäll, D. (2008). Emotional responses to music: the need to consider underlying mechanisms. *Behav. Brain Sci.* 31, 559–575. doi: 10.1017/S0140525X08005293

Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch*. New York, NY: Oxford University Press.

Krumhansl, C. L. (1997). An exploratory study of musical emotions and psychophysiology. *Can. J. Exp. Psychol.* 51, 336–353. doi: 10.1037/1196-1961.51.4.336

Krumhansl, C. L. (2002). Music: a link between cognition and emotion. *Curr. Dir. Cogn. Sci.* 11, 45–50. doi: 10.1111/1467-8721.00165

Krumhansl, C. L., and Zupnick, J. A. (2013). Cascading reminiscence bumps in popular music. *Psychol. Sci.* 24, 2057–2068. doi: 10.1177/0956797613486486

Levine, B., Svoboda, E., Hay, J. F., Winocur, G., and Moscovitch, M. (2002). Aging and autobiographical memory: dissociating episodic from semantic retrieval. *Psychol. Aging* 17, 677–689. doi: 10.1037/0882-7974.17.4.677

Lonsdale, A. J., and North, A. C. (2009). Musical taste and in-group favouritism. *Group Process. Intergroup Relat.* 12, 319–327. doi: 10.1177/1368430209102842

Luck, G. (2016). The psychology of streaming: exploring music listeners' motivations to favour access over ownership. *Int. J. Music Bus. Res.* 5, 46–61.

Mather, M., and Carstensen, L. L. (2005). Aging and motivated cognition: the positivity effect in attention and memory. *Trends Cogn. Sci.* 9, 496–502. doi: 10.1016/j.tics.2005.08.005

Narmour, E. (1990). *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model*. Chicago, IL: University of Chicago Press.

Nelson, K., and Fivush, T. (2004). The emergence of autobiographical memory: a social cultural developmental theory. *Psychol. Rev.* 111, 486–511. doi: 10.1037/0033-295X.111.2.486

Pearce, M. T., and Wiggins, G. A. (2012). Auditory expectation: the information dynamics of music perception and cognition. *Top. Cogn. Sci.* 4, 625–652. doi: 10.1111/j.1756-8765.2012.01214.x

Platz, F., Kopiez, R., Hasselhorn, J., and Wolf, A. (2015). The impact of song-specific age and affective qualities of popular songs on music-evoked autobiographical memories (MEAMs). *Musicae Sci.* 19, 327–349. doi: 10.1177/1029864915597567

Rathbone, C. J., O'Connor, A. R., and Moulin, C. J. A. (2016). The tracks of my years: personal significance contributes to the reminiscence bump. *Mem. Cogn.* 45, 136–150. doi: 10.3758/s13421-016-0647-2

Rentfrow, P. J., and Gosling, S. D. (2003). The do re mi's of everyday life: the structure and personality correlates of music preferences. *J. Pers. Soc. Psychol.* 84, 1236–1256. doi: 10.1037/0022-3514.84.6.1236

Rentfrow, P. J., and Gosling, S. D. (2006). Message in a ballad: the role of music preferences in interpersonal perception. *Psychol. Sci.* 17, 236–242.

Rubin, D. C., Rahhal, T. A., and Poon, L. W. (1998). Things learned in early adulthood are remembered best. *Mem. Cogn.* 26, 3–19. doi: 10.3758/BF03211366

Rubin, D. C., Wetzler, S. E., and Nebes, R. D. (1986). "Autobiographical memory across the adult lifespan," in *Autobiographical Memory*, ed. D. C. Rubin (Cambridge: Cambridge University Press), 202–221.

Sanjek, R. (1988). *American Popular Music and its Business: The First Four Hundred Years*. New York, NY: Oxford University Press.

Schellenberg, E. G., and von Scheve, C. (2012). Emotional cues in American popular music: five decades of the top 40. *Psychol. Aesthet. Creat. Arts* 6, 196–203. doi: 10.1037/a0028024

Schubert, E. (2016). Does recall of a past music event invoke a reminiscence bump in young adults? *Memory* 24, 1007–1014. doi: 10.1080/09658211.2015.1061014

Schulkind, M. D., Hennis, L. K., and Rubin, D. C. (1999). Music, emotion, and autobiographical memory: they're playing your song. *Mem. Cogn.* 27, 948–955. doi: 10.3758/BF03201225

Sloboda, J. A., and O'Neill, S. A. (2001). "Emotions in everyday listening to music," in *Music and Emotion: Theory and Research*, eds P. N. Juslin and J. A. Sloboda (Oxford: Oxford University Press), 415–429.

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., Dalca, I., and Levitin, D. J. (2005). Dimensions of emotion in expressive musical performance. *Ann. N. Y. Acad. Sci.* 1060, 462–466. doi: 10.1196/annals.1360.052

Whitburn, J. (1999). *A Century of Pop Music*. Menomonee Falls, WI: Record Research Inc.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX A

Songs Used in the Survey

1910–1919

Casey Jones	Billy Murray
Alexander's Ragtime Band	Arthur Collins and Byron Harlan
Moonlight Bay	American Quartet
When Irish Eyes Are Smiling	Chauncey Olcott
The Song That Stole My Heart Away	Henry Burr
It's A Long Way To Tipperary	John McCormack
M-o-t-h-e-r (A Word That Means The World To Me)	Henry Burr
Over There	American Quartet
Just A Baby's Prayer At Twilight	Henry Burr
Till We Meet Again	Henry Burr and Albert Campbell

1920–1929

Dardanella	Ben Selvin and His Orchestra
Wang-Wang Blues	Paul Whiteman
April Showers	Al Jolson
Parade Of The Wooden Soldiers	Paul Whiteman Orchestra
It Ain't Gonna Rain No Mo	Wendell Hall
The Prisoner's Song	Vernon Dalhart
Valencia (A Song Of Spain)	Paul Whiteman and His Orchestra
My Blue Heaven	Gene Austin
Sonny Boy	Al Jolson
Tiptoe Through The Tulips	Nick Lucas

1930–1939

Stein Song	Rudy Vallee
El Manicero (The Peanut Vendor)	Don Azipiazu and The Havana Casino Orchestra
Night and Day	Leo Reisman
The Last Round Up	George Olson and His Music
June In January	Bing Crosby
Cheek To Cheek	Fred Astaire
Pennies From Heaven	Bing Crosby
Sweet Leilani	Bing Crosby
A-Tisket A-Tasket	Ella Fitzgerald, Chick Webb
Deep Purple	Larry Clinton

1940–1949

In The Mood	Glenn Miller
Amapola (Pretty Little Poppy)	Jimmy Dorsey and His Orchestra
White Christmas	Bing Crosby
I've Heard That Song Before	Harry James and His Orchestra
Swinging On A Star	Bing Crosby
Rum and Coca Cola	The Andrews Sisters
The Gypsy	The Ink Spots
Near You	Francis Craig
Buttons And Bows	Dinah Shore
Riders In The Sky (A Cowboy Legend)	Vaughn Monoroe and His Orchestra

1950–1959

The Tennessee Waltz	Patti Page
Cry	Johnnie Ray and The Four Lads
You Belong To Me	Jo Stafford
Vaya Con Dios (May God Be With You)	Les Paul, Mary Ford
Little Things Mean A Lot	Kitty Kallen
Cherry Pink And Apple Blossom White	Perez Prado
Heartbreak Hotel	Elvis Presley
All Shook Up	Elvis Presley
Volare (Nel Blue Dipinto Di Blu)	Demenico Modugno
The Battle of New Orleans	Johnny Horton

1960–1969

Theme From "A Summer Place"	Percy Faith
Tossin' And Turnin'	Bobby Lewis
Stranger On The Shore	Mr. Acker Bilk
Sugar Shack	Jimmy Gilmer and The Fireballs
I Want To Hold Your Hand	The Beatles
Wooly Bully	Sam The Sham and The Pharoahs
The Ballad Of The Green Berets	Sgt. Barry Sadler
To Sir With Love	Lulu
Hey Jude	The Beatles
Sugar, Sugar	Archies

1970–1979

Bridge Over Troubled Water	Simon and Garfunkel
Joy To The World	Three Dog Night
The First Time Ever I Saw Your Face	Roberta Flack
Tie A Yellow Ribbon 'Round The Ole Oak Tree	Tony Orlando
The Way We Were	Barbara Streisand
Love Will Keep Us Together	Captain and Tennille
Silly Love Songs	Wings
Tonight's The Night (Gonna Be Alright)	Rod Stewart
Shadow Dancing	Andy Gibb
My Sharona	Knack

1980–1989

Call Me	Blondie
Bette Davis Eyes	Kim Carnes
Physical	Olivia Newton-John
Every Breath You Take	The Police
When Doves Cry	Prince
Careless Whisper	Wham!
That's What Friends Are For	Dionne and Friends
Walk Like An Egyptian	Bangles
Faith	George Michael
Look Away	Chicago

1990–1999

Hold On	Wilson Phillips
(Everything I Do) I Do It For You	Bryan Adams
End Of The Road	Boyz II Men
I Will Always Love You	Whitney Houston
The Sign	Ace of Base
Gangsta's Paradise	Coolio
Candle In The Wind	Elton John
Too Close	Next
Believe	Cher

2000–2009

Breathe	Faith Hill
Hanging By A Moment	Lifehouse
How You Remind Me	Nickelback
In Da Club	50 Cent
Yeah!	Usher featuring Lil' Jon and Ludacris
We Belong Together	Mariah Carey
Bad Day	Daniel Powter
Irreplaceable	Beyonce
Low	Flo Rida featuring T-Pain
Boom Boom Pow	The Black Eyed Peas

APPENDIX B

Questions on Survey

For each decade (1910s, 1920s, 1930s, 1940s, 1950s, 1960s, 1970s, 1980s, 1990s, and 2000s):

Percent recognized (0–100)
How much do you like these songs? (0–10)
How much do these songs make you feel sad? (0–10)
How much do these songs make you feel happy? (0–10)
How much do these songs make you feel nostalgic? (0–10)
How much do these songs make you feel energized? (0–10)
How much do these songs make you feel romantic? (0–10)
If given the opportunity, would you choose to hear more songs like this?
Are any of these songs associated with personal memories? (Y/N)

if so:
How specific are the memories (for example, who you were with, where, when)? (0–10)
During what period(s) in your life?-Childhood (up to 13 years old) (Y/N)
During what period(s) in your life?-Teens (13–19 years old) (Y/N)
During what period(s) in your life?-20s (including college) (Y/N)
During what period(s) in your life?-30s (Y/N)
During what period(s) in your life?-40s (Y/N)
During what period(s) in your life?-50s-65 (Y/N)
During what period(s) in your life?-Over 65 (Y/N)
What context(s)?-Listening alone (Y/N)
What context(s)?-Listening with parents (Y/N)
What context(s)?-Listening with spouse/partner (Y/N)
What context(s)?-Listening with children (Y/N)
What context(s)?-Listening with siblings or other family members (Y/N)
What context(s)?-Listening with friends or peers (Y/N)

Demographics

What is your gender? (M/F)
What year were you born?
What year was your mother born?
What year was your father born?
Do you have children?

if so:
What year was your first child born?
What year was your second child born?
What year was your third child born?
What year was your forth child born?

What is your nationality?
Are you living in the USA now? (Y/N)
If so, how many years have you lived in the USA?

For each of three periods of life (growing up at home, about 18 - 25, within the last year or so):
How much did you listen to?

Pop and Rock (hours per week)
Rhythm and Blues (hours per week)
Country and Folk (hours per week)
Classical (hours per week)
Jazz (hours per week)
Ethnic and World (hours per week)
Other (hours per week)

Where did you hear popular music?

Radio (Y/N)
Record (Y/N)
Tape cassette (Y/N)
Dances and parties (Y/N)
Concerts (Y/N)
Heard performed by family and friends (Y/N)
Played myself (Y/N)
CDs (Y/N)
Subscription services (e.g., Spotify, Rhapsody, etc.) (Y/N)
YouTube (Y/N)
Internet radio (e.g., Pandora) (Y/N)
Digital download (e.g., mp3) (Y/N)
Other (Y/N)

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership