RELATIONSHIP OF LANGUAGE AND MUSIC, TEN YEARS AFTER: NEURAL ORGANIZATION, CROSS-DOMAIN TRANSFER AND EVOLUTIONARY ORIGINS

EDITED BY: Caicai Zhang, Chao-Yang Lee, William Shiyuan WANG and Mary Miu Yee Waye

PUBLISHED IN: Frontiers in Psychology, Frontiers in Communication and Frontiers in Neuroscience

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-898-1
DOI 10.3389/978-2-88976-898-1

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: frontiersin.org/about/contact

RELATIONSHIP OF LANGUAGE AND MUSIC, TEN YEARS AFTER: NEURAL ORGANIZATION, CROSS-DOMAIN TRANSFER AND EVOLUTIONARY ORIGINS

Topic Editors:
Caicai Zhang, The Hong Kong Polytechnic University Kowloon, SAR China
Chao-Yang Lee, Ohio University Athens, United States
William Shiyuan WANG, The Hong Kong Polytechnic University Kowloon, SAR China
Mary Miu Yee Waye, The Chinese University of Hong Kong, China

Citation: Zhang, C., Lee, C.-Y., WANG, W. S., Waye, M. M. Y., eds. (2022). Relationship of Language and Music, Ten Years After: Neural Organization, Cross-domain Transfer and Evolutionary Origins. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-898-1

Table of Contents

frontiers | Frontiers in Psychology

TYPE Editorial
PUBLISHED 28 July 2022
DOI 10.3389/fpsyg.2022.990857

OPEN ACCESS

EDITED AND REVIEWED BY
Xiaolin Zhou,
Peking University, China

*CORRESPONDENCE
Chao-Yang Lee
leec1@ohio.edu

SPECIALTY SECTION
This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

RECEIVED 10 July 2022
ACCEPTED 12 July 2022
PUBLISHED 28 July 2022

CITATION
Lee C-Y, Zhang C, Wang WS-Y and Waye MMY (2022) Editorial: Relationship of language and music, ten years after: Neural organization, cross-domain transfer and evolutionary origins. *Front. Psychol.* 13:990857. doi: 10.3389/fpsyg.2022.990857

Editorial: Relationship of language and music, ten years after: Neural organization, cross-domain transfer and evolutionary origins

Chao-Yang Lee[1]*, Caicai Zhang[2], William Shi-Yuan Wang[2] and Mary Miu Yee Waye[3]

[1]Division of Communication Sciences and Disorders, Ohio University, Athens, OH, United States, [2]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China, [3]The Nethersole School of Nursing, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China

KEYWORDS

language, music, speech, speech perception, music perception

Editorial on the Research Topic
Relationship of language and music, ten years after: Neural organization, cross-domain transfer and evolutionary origins

Language and music are both evolutionarily old and ubiquitous in human cultures. They also share many common features such as the use of basic acoustic attributes, the presence of complex hierarchical structures, and the ability to elicit and communicate emotions. These parallels have sparked questions regarding the neural organization of language and music, the cross-domain transfer between them, and their evolutionary origins. Ten years after the publication of a similar Research Topic in Frontiers, many intriguing questions remain. The 11 articles in this collection address the relationship between language and music from a wide range of perspectives, including six empirical studies on cross-domain transfer, three articles on clinical applications, and two articles on evolutionary perspectives (Figure 1).

Cross-domain transfer

A plethora of studies have demonstrated that long-term experience with either language or music can transfer to processing or learning in the other domain. However, the extent and direction of the transfer remains controversial. Several hypotheses have been proposed to explain the transfer, including the domain-general sharpening of sensory encoding (Chandrasekaran and Kraus, 2010), common acoustic processing and influence on abstract representations in the other domain (Besson et al., 2011), and the OPERA/expanded OPERA hypothesis (Patel, 2011, 2014). The use of lexically distinctive pitch in tonal languages has also provided a unique opportunity to evaluate the language-music relationship.

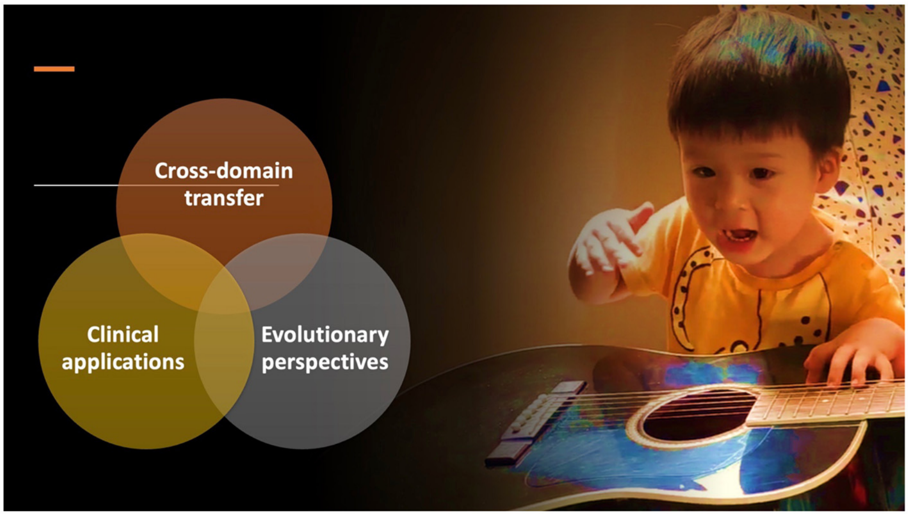


FIGURE 1
Three areas of study covered by the current Research Topic.

Zhu et al. showed that musicianship affects categorical perception of Mandarin tones by native listeners. Mismatch negativity (MMN) amplitude in response to tonal deviations was amplified in amateur musicians, indicating that musical training enhances pre-attentive processing of lexical tones. This finding also suggests that amateur musical training is sufficient to induce neural plasticity in speakers who already have long-term tonal experience. Chen et al. further showed that musicianship affects categorical perception of Mandarin tones in speakers of a second tone language. Perception of Mandarin pitch direction was more categorical in Cantonese-speaking musicians than non-musicians. The musicians were also more sensitive to stimulus duration and intrinsic F0 associated with vowel quality, suggesting that musicians are able to use their sensitivity to acoustics to form more robust representations of tones in a second language.

The impact of musical experience extends beyond phonetic categorization. Smit et al. found an association between pitch discrimination ability and learning of novel words. Surprisingly, better pitch discrimination was associated with worse word learning. The use of infant-directed speech in this study may have led to greater pitch variation, which turned out to be detrimental to learning for individuals with better pitch discrimination abilities. Nie et al. showed that musical training affects working memory. The authors compared a group of Mandarin-speaking children who received 1-year music training to another two groups of children that received either second-language training or no training. After controlling for initial group differences in the baseline performance, the authors found superior development of auditory working memory in the music group.

Choi examined the other direction of the cross-domain transfer, i.e., how language experience affects musical pitch perception. Cantonese language experience facilitated musical pitch perception, but the effect was limited to non-musicians. The lack of a language effect among musicians was attributed to perceptual saturation due to musical training or specialized pitch processing. With these findings, Choi proposed to revise the "Precision" criterion in the OPERA/expanded OPERA hypothesis to accommodate the language-to-music transfer.

The cross-domain transfer, however, does not appear to apply all the time. Tao et al. used the language-music connection to address a long-standing issue in speech perception: Does speech normalization require a general auditory mechanism or a speech-specific perceptual mechanism? The authors showed that familiarity with a music context did not give musicians an advantage in Cantonese tone normalization. Rather, tone normalization occurred only in the speech context, and there was no difference in tone normalization performance between musicians and non-musicians regardless of the context, suggesting that a speech-specific mechanism is responsible for perceptual normalization.

Clinical applications

Three articles addressed the clinical application of the language-music connection. Zhang et al. compared the effectiveness of melodic intonation therapy (MIT) and traditional speech therapy on Mandarin-speaking individuals with non-fluent aphasia. After 8 weeks of therapy, patients receiving MIT showed better listening comprehension, repetition, and spontaneous naming compared to those receiving traditional speech therapy. The authors concluded that MIT is an effective approach to rehabilitating language functions. Zhang, Li, et al. offered a systematic review of 39 randomized controlled trials examining the effect of MIT on the treatment of non-fluent aphasia. The review showed that behavioral measurements were used in most of the studies, and few studies provided brain imaging data. With this observation, the authors call for more clinical studies incorporating both behavioral and neurophysiological data to evaluate the effectiveness of MIT.

Zhang, Song, et al. compared the effects of vocal intonation therapy (VIT) to standard respiratory therapy on voice production in people with respiratory dysfunction resulting from cervical spinal cord injury (CSCI). After 12 weeks of treatment, patients in the VIT group outperformed those in the standard respiratory therapy group in measures of speech volume, singing volume, sustained note length, and fundamental frequency, suggesting that VIT is an effective treatment for respiratory dysfunctions in CSCI patients.

Evolutionary perspectives

Commonalities between speech and music in basic acoustic attributes and hierarchical structure have inspired many researchers to hypothesize a common evolutionary precursor. Previous studies have probed the co-evolution hypothesis from the perspective of pitch (Thompson et al., 2012; Fenk-Oczlon, 2017). Fenk-Oczlon extended the investigation to duration. By showing iconic association between vowel height and the duration of musical notes in 20 Alpine traditional yodels, Fenk-Oczlon provided new evidence for the "musical protolanguage" hypothesis (Darwin, 1871; Fitch, 2005, 2006). The close coupling of vowel height and music in singing with meaningless syllables is perhaps reminiscent of an ancient, prosodic protolanguage.

The evolutionary perspective is elaborated in Haiduk and Fitch. Following Darwin's speculation about the speech-music relationship (Darwin, 1871) and Hockett's "design features" of communication (Hockett, 1960), Haiduk and Fitch delineate the evolutionary circumstances which led to the distinctive trajectories that language and music took in their respective developments. They consider the differences and similarities in evolution when language and music are used in a variety of contexts. These ideas are new probes and need to be nourished with much more interaction and perhaps experimentation across diverse cultures. Cultural evolution works much faster than biological evolution, and humans have become a very different kind of animal since Darwin's time. We are living in a very different physical and social environment of our own making. The trajectories that language and music will take in the decades ahead are hard to predict and bound to be fascinating.

Ethics statement

Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

Author contributions

C-YL wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Conflict of interest

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Besson, M., Chobert, J., and Marie, C. (2011). Transfer of training between music and speech: common processing, attention, and memory. *Front. Psychol.* 2, 1–12. doi: 10.3389/fpsyg.2011.00094

Chandrasekaran, B., and Kraus, N. (2010). The scalp-recorded brainstem response to speech: neural origins and plasticity. *Psychophysiology* 47, 236–246. doi: 10.1111/j.1469-8986.2009.00928.x

Darwin, C. (1871). *The Descent of Man: and Selection in Relation to Sex*. London: John Murray.

Fenk-Oczlon, G. (2017). What vowels can tell us about the evolution of music. *Front. Psychol.* 8, 1581. doi: 10.3389/fpsyg.2017.01581

Fitch, W. T. (2005). The evolution of language: a comparative review. *Biol. Philos.* 20, 193–203. doi: 10.1007/s10539-005-5597-1

Fitch, W. T. (2006). The biology and evolution of music: a comparative perspective. *Cognition* 100, 173–215. doi: 10.1016/j.cognition.2005.11.009

Hockett, C. D. (1960). The origin of speech. *Sci. Am.* 203, 88–96. doi: 10.1038/scientificamerican0960-88

Patel, A. D. (2011). Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. *Front. Psychol.* 2, 142. doi: 10.3389/fpsyg.2011.00142

Patel, A. D. (2014). Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis. *Hearing Res.* 308, 98–108. doi: 10.1016/j.heares.2013..011

Thompson, W. F., Marin, M. M., and Stewart, L. (2012). Reduced sensitivity to emotional prosody in congenital amusia rekindles the musical protolanguage hypothesis. *Proc. Natl. Acad. Sci.* 109, 19027–19032. doi: 10.1073/pnas.1210344109

ORIGINAL RESEARCH
published: 15 March 2021
doi: 10.3389/fpsyg.2021.611189

Effects of Amateur Musical Experience on Categorical Perception of Lexical Tones by Native Chinese Adults: An ERP Study

Jiaqiang Zhu[1], *Xiaoxiang Chen*[1]* *and Yuxiao Yang*[2]

[1] *School of Foreign Languages, Hunan University, Changsha, China,* [2] *Foreign Studies College, Hunan Normal University, Changsha, China*

Music impacting on speech processing is vividly evidenced in most reports involving professional musicians, while the question of whether the facilitative effects of music are limited to experts or may extend to amateurs remains to be resolved. Previous research has suggested that analogous to language experience, musicianship also modulates lexical tone perception but the influence of amateur musical experience in adulthood is poorly understood. Furthermore, little is known about how acoustic information and phonological information of lexical tones are processed by amateur musicians. This study aimed to provide neural evidence of cortical plasticity by examining categorical perception of lexical tones in Chinese adults with amateur musical experience relative to the non-musician counterparts. Fifteen adult Chinese amateur musicians and an equal number of non-musicians participated in an event-related potential (ERP) experiment. Their mismatch negativities (MMNs) to lexical tones from Mandarin Tone 2–Tone 4 continuum and non-speech tone analogs were measured. It was hypothesized that amateur musicians would exhibit different MMNs to their non-musician counterparts in processing two aspects of information in lexical tones. Results showed that the MMN mean amplitude evoked by within-category deviants was significantly larger for amateur musicians than non-musicians regardless of speech or non-speech condition. This implies the strengthened processing of acoustic information by adult amateur musicians without the need of focused attention, as the detection of subtle acoustic nuances of pitch was measurably improved. In addition, the MMN peak latency elicited by across-category deviants was significantly shorter than that by within-category deviants for both groups, indicative of the earlier processing of phonological information than acoustic information of lexical tones at the pre-attentive stage. The results mentioned above suggest that cortical plasticity can still be induced in adulthood, hence non-musicians should be defined more strictly than before. Besides, the current study enlarges the population demonstrating the beneficial effects of musical experience on perceptual and cognitive functions, namely, the effects of enhanced speech processing from music are not confined to a small group of experts but extend to a large population of amateurs.

Keywords: amateur musical experience, categorical perception, Mandarin lexical tones, MMN, cortical plasticity

Edited by:
Caicai Zhang,
The Hong Kong Polytechnic University, Hong Kong

Reviewed by:
Linjun Zhang,
Beijing Language and Culture University, China
Zhen Qin,
Hong Kong University of Science and Technology, Hong Kong
Mireille Besson,
UMR 7291 Laboratoire de Neurosciences Cognitives (LNC), France

****Correspondence:***
Xiaoxiang Chen
xiaoxiangchensophy@hotmail.com

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

Received: *28 September 2020*
Accepted: *17 February 2021*
Published: *15 March 2021*

Citation:
Zhu J, Chen X and Yang Y (2021) Effects of Amateur Musical Experience on Categorical Perception of Lexical Tones by Native Chinese Adults: An ERP Study. Front. Psychol. 12:611189. doi: 10.3389/fpsyg.2021.611189

INTRODUCTION

Pertaining to the old relationship between music and language, it is believed that the spoken language evolves from music (Darwin, 1871), or music evolves from the spoken language (Spencer, 1857), or both of them descend from a common origin (Rousseau, 1781/1993). These viewpoints bolster the notion that music and language, both of which involve complex and meaningful sound sequences (Patel, 2008), are reciprocally connected. It has been considerably verified that musical experience impacts on language or speech processing (Besson et al., 2011a,b, Patel, 2014). One of the most common approaches to the assessment of speech processing is adopting categorical perception in experiments, in which individuals are required to perceptually categorize continuous auditory signals into discrete linguistic representations along a physical continuum (Fujisaki and Kawashima, 1971). The phenomenon of categorical perception has been investigated with preliminary foci on segments of consonants and vowels (e.g., Liberman et al., 1957; Fry et al., 1962; Miller and Eimas, 1996).

In recent years, a surge of interest has been observed concerning the suprasegmental aspect of speech processing (e.g., Yip, 2002; Hallé et al., 2004; Xu et al., 2006; Peng et al., 2010; Liu et al., 2018). One example is the research into lexical tones, which are phonemically contrastive and alter the semantic meanings of words in usage (Gandour and Harshman, 1978). Through shifts in pitch height and pitch contour, lexical tones can be distinguished and recognized in disparate categories (Francis et al., 2003). For instance, monosyllables "ma 55," "ma 35," "ma 214," and "ma 51" in Mandarin mean "mother," "hemp," "horse," and "to scold," when presented individually (Chao, 1968). The numerals following the syllables stand for the transcribed tones, depicting the relative pitch value within a five-point scale of the talker's normal frequency range (Chao, 1947). These four tones can also be annotated with respective pitch patterns as Tone 1 (T1), level; Tone 2 (T2), rising; Tone 3 (T3), falling-rising; and Tone 4 (T4), falling (Wang et al., 2017).

Categorical Perception of Lexical Tones

Lexical tones are privileged in Mandarin both phonetically and phonologically, and categorical perception of lexical tones has been broadly researched in present decades (Xu et al., 2006; Peng et al., 2010; Chen F. et al., 2017). In a classic experiment of categorical perception, participants complete an identification task to label a flow of tonal stimuli and a discrimination task to estimate some tonal contrasts as either "same" or "different" (Xu et al., 2006). The auditory stimuli are physically interpolated with variable pitch values along a continuum (Francis et al., 2003). Separated by categorical boundary between two tones as defined by identification curves (Peng et al., 2010), within-category stimuli stemmed from one category can be perceived analogously, but across-category stimuli extracted from two categories tend to be perceived differentially. Stimulus tokens from discrepant categories are more discriminable than those from the same category (Joanisse et al., 2006; Jiang et al., 2012; Chen and Peng, 2018).

Most of prior studies probed into lexical tones as a whole (e.g., Wang et al., 2001; Francis et al., 2003; Hallé et al., 2004), while more recent studies have switched to the dynamic interaction between acoustic and phonological information of lexical tones (e.g., Xi et al., 2010; Zhang et al., 2011; Yu et al., 2014). In general, the acoustic information consists of the physical features of lexical tones as estimated by F0 (e.g., pitch height and pitch contour), while the phonological information refers to the linguistic properties with tonal categories to distinguish lexical semantics (Yu et al., 2019). Although some secondary cues might influence the judgment of lexical tone contrasts, F0 remains most critical as amply confirmed by the seminal (Wang, 1976) and subsequent studies of categorical perception of lexical tones (Xu et al., 2006; Peng et al., 2010; Shen and Froud, 2016). It is reported that for Mandarin lexical tones, the perception of within-category pairs mainly depends on lower-level acoustic information of pitch, yet the perception of across-category comparisons is principally reliant on higher-level phonological information of lexical categories (Fujisaki and Kawashima, 1971; Xi et al., 2010; Yu et al., 2014). Importantly, distinguishing acoustic and phonological information of lexical tones re-paints a clear picture to interpret the mechanisms underlying lexical tone perception. For example, the study by Xi et al. (2010) revealed that when listening to Mandarin lexical tones, native speakers need to process the acoustic information and the phonological information simultaneously. A following study by Yu et al. (2014) manipulated phonological categories and acoustic intervals of lexical tones in experiments and replicated the results of Xi et al. (2010). Their additional findings uncovered the temporal pattern of lexical tone processing showing that phonological processing precedes acoustic processing at the pre-attentive cortical stage, which was then re-confirmed in a subsequent study by Yu et al. (2017). The pre-attentive stage refers to an earlier stage at which individuals involuntarily process the stimuli, in contrast to the later attentive stage at which individuals consciously process the stimuli (Neisser, 1967; Kubovy et al., 1999; Yu et al., 2019). According to Zhao and Kuhl (2015a,b), this pattern of temporal processing involves the dominant influences of the higher-level linguistic categories as compared to the lower-level acoustics relating to lexical tones.

Given that native speakers of tone languages outperform those of non-tone languages, a plethora of studies support the notion that lexical tone perception is plastic and experience-dependent (Hallé et al., 2004; Peng et al., 2010; Shen and Froud, 2016; Chen S. et al., 2017). To be exemplified, compared to Mandarin speakers who perceive native lexical tones categorically, participants from non-tone languages show impoverished performance in tasks requiring identification and discrimination of lexical tones. In the study by Hallé et al. (2004), although French listeners were observed to have substantial sensitivity to pitch contour differences, they failed to perceive lexical tones along the lines of a well-defined and finite set of linguistic representations as exhibited by across-category tonal contrasts. However, those speaking the tone language Taiwanese could perceive lexical tones in a quasi-categorical manner. The authors proposed that this disparity was attributable to the existence of phonological information relating to lexical tones in Taiwanese, which was, however, absent in French. Considering the gradation of identification and discrimination tasks, Peng et al. (2010) demonstrated that German listeners exhibited larger boundary widths and psychophysical boundaries rather than linguistic boundaries compared to their Mandarin and Cantonese counterparts. The results of Xu et al. (2006) also exhibited strong categorical perception for Chinese-speaking but not for English-speaking participants in their experiments. According to Chen et al. (2020), listeners from non-tone languages exhibited psychoacoustically based performance because of the lack of experience with lexical tones; however, Shen and Froud (2016) found that with increased exposure to lexical tones, English learners of Mandarin would show similar performance to native Mandarin speakers. Since categorical perception provides an ideal window to disentangle acoustic information from phonological information through pairing respective within- and across-category auditory stimuli, the current study would make use of this paradigm to research the processing of these two types of information, which is less studied among listeners with different levels of musical pitch expertise.

The Influence of Musicianship on Lexical Tone Processing

The brain perceptual plasticity of lexical tones induced by language experience has been further demonstrated via cross-domain research, suggesting that musical experience also impacts on the perception of lexical tones. Besson et al. (2011a) proposed that the musician's brain is a good model of brain plasticity. The links between music and language are grounded on findings from numerous empirical studies (e.g., Schön et al., 2004; Marques et al., 2007; Moreno, 2009; Posedel et al., 2012; Ong et al., 2020). It is believed that musicians show advantages in processing and encoding speech sounds due to increased plasticity and perceptual enhancements (Magne et al., 2006; Kraus and Chandrasekaran, 2010; Gordon et al., 2015). According to the expanded hypothesis of Overlap, Precision, Emotion, Repetition and Attention (OPERA-e), musical experience enhances speech processing because common sensory and cognitive processing mechanisms are shared by music and language (Patel, 2014). Based on the conceptual framework of OPERA-e, the perception of fine-grained musical pitches remains transferrable to that of coarse-grained lexical tones, as supported by the studies on music and lexical tone processing. For example, Alexander et al. (2005) found that a group of American musicians obtained higher scores than non-musicians in identifying and discriminating four lexical tones in Mandarin. Zhao and Kuhl (2015a) demonstrated that, given no prior experience of Mandarin or any other tone languages, English-speaking musicians were more sensitive to pitch variations of tonal stimuli from Mandarin T2–T3 continuum than non-musicians. A follow-up experiment of perceptual training of lexical tones was held, results demonstrating that compared to the non-musician counterparts, musician trainees showed improvement in identification in post-training test (Zhao and Kuhl, 2015a). It certified that short-term perceptual training altered perception, which spanned only about 2 weeks. The study by Moreno et al. (2008) investigated individuals trained via respective music and painting lessons. The results uncovered that participants with musical training were strengthened in both pitch discrimination and reading abilities, in contrast to those with painting training. It highlighted the influence of musical experience on speech processing. Zheng and Samuel (2018) also found that English-speaking musicians outperformed non-musicians in the processing of both speech (Mandarin short phrases) and non-speech (F0) sounds.

The research into music and lexical tone processing has benefitted from neurophysiological assessments applied in an array of studies, which allow monitoring the perception of auditory signals within the brain. One of the most prevalent approaches is the measurement of event-related potentials (ERPs) to quantify brain activities in response to specific events with a high-temporal resolution in millisecond. Because the auditory ERP component, known as the mismatch negativity (MMN), can evaluate automatic discrimination at the pre-attentive cortical stage, it has been pervasively employed in neural studies of lexical tone perception (e.g., Chandrasekaran et al., 2007a,b; Li and Chen, 2015; Nan et al., 2018). Specifically, MMN peak latency reflects the time course of cognitive processing, while MMN mean amplitude indexes the extent to which neural resources relate to our brain activities (Duncan et al., 2009). Through ERP measurements, Besson et al. (2011a) tested the perception of Mandarin tonal and segmental variations among French-speaking musicians and non-musicians. The results exhibited increased amplitude and shorter latency of ERP components of N2, N3, and P3 for musicians than non-musicians, suggesting that musical expertise impacts on the categorization of foreign linguistic contrasts. Chandrasekaran et al. (2009) examined the perception of non-speech tone homologs to Mandarin T1, T2, and a linear rising ramp (T2L). The results revealed that English-speaking musicians provoked larger MMN responses than their non-musician counterparts, regardless of within-category (T2/T2L) or across-category (T1/T2) tonal contrast. This finding indicates that experience-dependent effects of pitch processing are domain-general. In spite of the preexisting long-term experience of lexical tones, Tang et al. (2016) found that Mandarin-speaking musicians showed increased MMN amplitude to changes of lexical tones compared to non-musicians, which implicates that musical experience facilitates cortical plasticity of linguistic pitch processing.

Although the studies regarding the influence of musical experience on lexical tone processing are not rare, most concentrate on the performance of professional musicians, who usually complete long-term musical training for tens of years (e.g., Pantev et al., 2001; Marie et al., 2012; Dittinger et al., 2016), receive formal and theoretical musical education in music conservatories (e.g., Vuust et al., 2012; Lee et al., 2014; Tang et al., 2016), and start musical practice very early (before puberty) in life (e.g., Cooper and Wang, 2012; Zhao and Kuhl, 2015a,b). In stark contrast to expert musicians in a preponderance of studies mentioned above, amateur musicians involve those who are non-music majors with later onset age and shorter musical length for their limited musical experience. Since it has been widely accepted that music plays a strong modulatory role in boosting language or speech processing (Patel, 2008, 2014; Besson et al., 2011a,b; Strait and Kraus, 2014), a question then arises as to whether individuals with amateur musical experience can obtain similar advantageous effects to experts such as in categorical perception of lexical tones.

A very recent study identified that children who attended informal musical group activities demonstrated better neural sound discrimination than controls (Putkinen et al., 2019). However, it remains less known whether the facilitative effects of amateur musical experience can also be found in adults, who diverge from children in light of physiological maturation and language exposure. Gfeller (2016) argued that both music and language as communicative forms encompass many subskills, and these are impacted by maturation as well as auditory input and experience (Yang et al., 2020). In the study by Chen et al. (2010), prelingually deafened participants with cochlear implants completed the pitch ranking of tonal pairs. The results showed that the length of musical experience was beneficial only for young participants. Best (1994, 1995) and Best and Tyler (2007) concluded that children and adults performed differently in speech perception because their perceptual systems had been tuned to variable degrees as a function of native language exposure. For example, according to Best et al. (2016), the discriminability of two phones produced with the same articulatory organ improves with increased native language exposure, yet the same improvement was not observed among adults in Yang and Chen (2019). In addition, for Mandarin-speaking participants, the significant differences in voice onset time, defined as the time interval of the burst and the beginning of glottal pulse in stop consonants (Cho and Ladefoged, 1999), were observed in Ma et al. (2017), which were ascribed to the physiological differences between children and adults. Given the vast disparities in neuroplasticity and hearing history, adults and children might have different bioelectrical responses to auditory stimuli at the pre-attentive stage originating from the impact of amateur musical experience.

The Present Study

The study by Putkinen et al. (2019) serves as an encouraging indication of the benefits of musical exposure while highlighting the need to continue researching the effects of amateur musical experience throughout adulthood. We recruited adult participants with limited musical experience (mean ± standard deviation: 4.5 ± 0.3, range: 4–5 years) of playing orchestral instruments requiring intensive usage of musical pitch (Sluming et al., 2002; Vuust et al., 2012; Patel, 2014). Although these participants reported that they always enjoyed their musical practice, none of them had received an early musical education, taken private lessons, or obtained any professional certificates in musical practice; crucially, their involvement in music (all after 16 years old) was motivated by self-willingness rather than commercial performance (Marie et al., 2012). The musical expertise of these amateur musicians in this study was prominently lower than that of expert musicians investigated in previous studies (e.g., Pantev et al., 2001; Alexander et al., 2005; Vuust et al., 2012; Wu et al., 2015; Zhao and Kuhl, 2015a,b).

What has been unveiled thus far about adult amateur musicians is sparse; however, the population of adult amateur musicians worldwide is enormous, in contrast to the limited number of professional musicians. The investigation of the effects of amateur musical experience on adults' perception of lexical tones is of great importance, because it helps resolve the research issues mentioned above and address whether participants with amateur musical experience should be differentially grouped from non-musicians, when conducting experiments in relation to the processing of word-level lexical tones or sentence-level intonations (Qin et al., 2021). That is to say, among these tests, it might be inappropriate to ignore participants' musical expertise or simply regard those with informal musical training as non-musicians. In addition, both behavioral and neural studies have elucidated that musicianship brings positive impacts on language and cognition across the life span for children (Schellenberg, 2004), adults (Wang et al., 2015), and aging citizens (Román-Caballero et al., 2018). Therefore, findings from the current study might encourage more individuals to participate in musical activities no matter what levels of performance they maintain and what backgrounds they are from, in the hope of increasing their aesthetic appreciation as well as helping them balance their physical and mental health.

In the current study, categorical perception of lexical tones was adopted so as to tease apart acoustic and phonological information by pairing respective within- and across-category stimuli (Yu et al., 2014). Except for the investigation of lexical tones in the speech condition, pure tones as the non-speech stimuli with congruous F0 features were meanwhile exploited to examine whether music-driven and experience-dependent effects of pitch processing were domain-general (Chandrasekaran et al., 2009). Grounded on the framework of OPERA-e (Patel, 2014) that common sensory and cognitive processes relating to pitch permit the facilitative effects from music to lexical tone perception, discrepancies were anticipated between amateur musicians and non-musicians with respect to the coverage of MMN mean amplitudes and MMN peak latencies. Concretely, according to preceding studies (Chandrasekaran et al., 2009; Besson et al., 2011a; Wu et al., 2015; Tang et al., 2016), MMN mean amplitudes were expected to be larger in the perception of within-category stimuli by amateur musicians than non-musicians, yet both of them would be comparable when processing across-category stimuli; in other words, amateur musicians might be only enhanced in acoustic processing for native lexical tones. As to MMN peak latencies, there exist two competing views about the time course of acoustic and phonological processing of lexical tones. Luo et al. (2006) proposed a serial model, the two-stage model, arguing that only acoustic information of lexical tones is processed at earlier pre-attentive stage and phonological information is processed at later attentive stage. However, many recent ERP studies showed that phonological information of lexical tones is processed in parallel with acoustic information at the pre-attentive stage (Xi et al., 2010; Yu et al., 2014, 2017). In this regard, we hypothesized that acoustic and phonological information would be processed concurrently, which contradicted the two-stage model of lexical tone processing (Luo et al., 2006).

To the best of our knowledge, the present study is the first attempt to clarify the aforementioned issues in Chinese adult population from a neural perspective. By exploring whether the facilitative effects from music to speech processing could be grasped by a large group of amateur musicians similar to experts in previous studies, this study aimed to provide an in-depth understanding of neuroplasticity in addition to the processing of lexical tones after re-visiting the relationship between music and language.

MATERIALS AND METHODS

Participants

Thirty adult Mandarin-speaking college students (18 males and 12 females, aged 21–30 years, mean age 24) were recruited from universities in Shenzhen, China through online advertising. All participants were confirmed as having no history of speech or hearing disorders, learning disabilities, brain injuries, or neurological problems (experienced themselves or by relatives). Based on the well-documented criteria for classifying musical expertise (e.g., Pantev et al., 2001; Alexander et al., 2005; Marie et al., 2012; Hutka et al., 2015), the participants were divided into two groups: non-musicians (NM) and amateur musicians (AM). The AM group consisted of 15 amateur musicians (6 males and 9 females, aged 21–25 years, mean age 23), none of whom majored in music. Their limited musical experience ranged from 4 to 5 years with the mean and standard deviation at 4.5 and 0.3, individually. The NM group served as the control group, which consisted of 15 participants (12 males and 3 females, aged 22–30 years, mean age 24) with no musical experience (e.g., playing an instrument or vocal training). Although no power analysis was performed for the calculation of sample size, the sample size of the current study was comparable with one seminal ERP study by Xi et al. (2010) that also focused on the processing of acoustic versus phonological information via categorical perception of Mandarin lexical tones. All participants were paid monetarily for their participation.

Both AM and NM members were right-handed according to a handedness questionnaire adapted from a modified Chinese version of the Edinburgh Handedness Inventory (Oldfield, 1971). Consent forms were signed by participants prior to the experiment, which was approved by the Ethics Review Board at the School of Foreign Languages of Hunan University.

Stimuli

Sampled at 44.1 kHz and digitized at 16 bits, the Chinese monosyllable /pa/ was recorded with respective T2 and T4 in a sound-attenuated room by a native female speaker from northern Mainland China. The primary cue to distinguish tonal contrasts in Mandarin refers to pitch, known as the psychological percept of F0 (Abramson, 1978; Gandour, 1983). Nevertheless, the comparison of some pairs of Mandarin lexical tones may be affected by cues in addition to pitch. For example, distinguishing T2 and T3 always confuses both native and non-native Mandarin speakers (Li and Chen, 2015), and the timing of the turning point in pitch contour is also critical for the discrimination (Shen and Lin, 1991). The signal properties of T2 and T3 are not very distinctive and their acoustic similarities are further compounded by the Mandarin tone sandhi (Hao, 2012). Unlike similar tones of T2 and T3, T2 and T4 have disparate pitch contours and remain phonologically distinctive (Chao, 1948). Hence, Mandarin T2 and T4 were purposefully selected, not only because they had been appreciably employed in categorical perception of lexical tones in previous studies, but also because the discrimination of T2 and T4 was reliant on the detection of pitch variations rather than other confounding features (e.g., phonation) in the acoustic signals (Xi et al., 2010; Zhang et al., 2011; Li and Chen, 2015; Zhao and Kuhl, 2015a). The examination of T2 and T4 was likely to maximize the potential differences between amateur musicians and non-musicians in the processing of across-category and within-category stimuli along the tonal continuum.

The two lexical tones were normalized to a sound pressure level of 70 dB and a duration of 200 ms using the Praat software (Boersma and Weenink, 2019). In addition, the Mandarin T2–T4 continuum was manipulated by applying the pitch-synchronous overlap and added function (Moulines and Laroche, 1995) via Praat. As shown in **Figure 1**, 11 stimulus sets were created spanning the continuum with an equalized acoustic interval between each step. Prototypically, the first stimulus (S1) referred to T2 and the last stimulus (S11) signaled T4. The non-speech stimuli were pure tones, with exactly the same pitch, intensity, and duration as the speech stimuli, which were resynthesized following the procedures of Peng et al. (2010).

According to prior studies on categorical perception, the third (S3) and the last (S11) stimuli were chosen as deviants, with the seventh (S7) being the standard from the current continuum, which would be played in the neural tests (Xi et al., 2010;

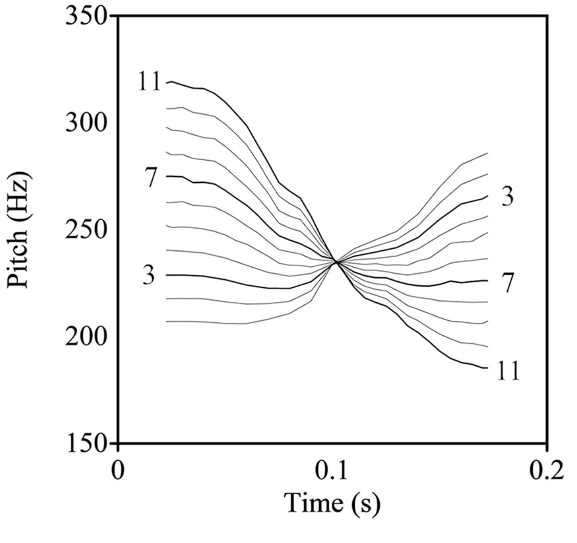


FIGURE 1 | The schematic illustration of the tonal continuum (the thick lines with numerals represent the stimuli used in the neural tests).

Zhang et al., 2011; Yu et al., 2014). Although both deviants were equidistant in frequency size to the standard, the stimuli of S3 and S11 were defined as across-category and within-category deviants, respectively. Crucially, to further assure the feasibility of stimulus deployment, an identification task was conducted in order to locate the categorical boundary (Peng et al., 2010). The categorical boundary was computed using Probit analysis, which involved the commensurate 50% crossover point in the continuum (Finney, 1971). Based on the boundary position, across-category and within-category stimuli for each participant could be paired in agreement with Jiang et al. (2012). For instance, if one participant retained the boundary position as 4.9 in the identification task, pairs S3–S5 and S4–S6 that straddled the position would be coded as across-category comparisons, whereas the remaining pairs that did not cross the boundary would be taken as within-category comparisons (Chen F. et al., 2017).

The identification task was completed by 10 participants who did not attend the following electroencephalogram (EEG) recording phase. All 11 stimuli were presented randomly through a laptop using the E-Prime 2.0 program (Psychology Software Tools Inc., United States). Each stimulus was played nine times. The design of two-alternative forced choices was applied, thereby participants had to make a choice when they heard the sounds. Both T2 and T4 were labeled on the keyboard and participants pressed the target buttons to respond. **Figure 2** demonstrates the identification curves.

In the current study, categorical boundary positions in speech and non-speech conditions were 6.47 and 6.46, respectively, which indicated that the pairing of S3–S7 was across-categorical, while S7–S11 remained within-categorical. Therefore, as mentioned above, the present stimulus deployment was operationalized and could be applied to the next EEG data collection. In addition, for both AM and NM participants, one more active behavioral identification task was carried out after their recordings of ERPs. All participants correctly identified the three lexical tones as either T2 or T4 with 0% error rate out of the four choices from T1, T2, T3, and T4 in Mandarin. The results revealed that amateur musicians and non-musicians perceived S3–S7 as an across-category comparison and S7–S11 as a within-category comparison.

ERP Procedure

In line with Näätänen et al. (2004) and Pakarinen et al. (2007), the current study adopted a multifeature passive oddball paradigm, which consists of more than one type of deviant in one block (Partanen et al., 2013; Yu et al., 2019). The 15 standard stimuli were played first to prompt participants to establish a standard perceptual template. Then, 1,000 stimuli (800 standards and 200 deviants) were played binaurally. The number of each type of deviant was 100. The stimulus-onset asynchrony (SOA) was 800 ms, and each sound was presented for 200 ms. The deviants were repeated pseudo-randomly with any two adjacent deviants separated by at least three standards, as displayed in **Figure 3**. The speech stimuli were set into one block and the non-speech stimuli were contained in another block. Two

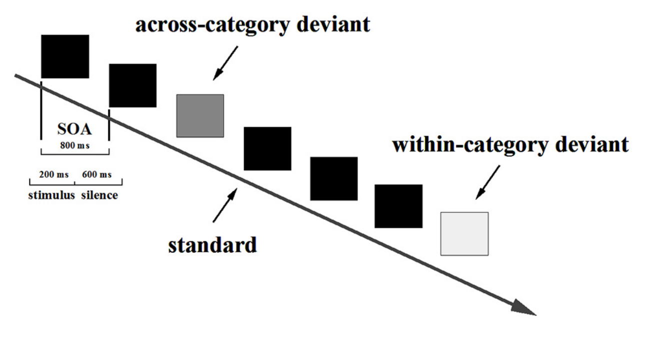


FIGURE 3 | The schematic diagram for typical trials in the improved passive oddball paradigm.

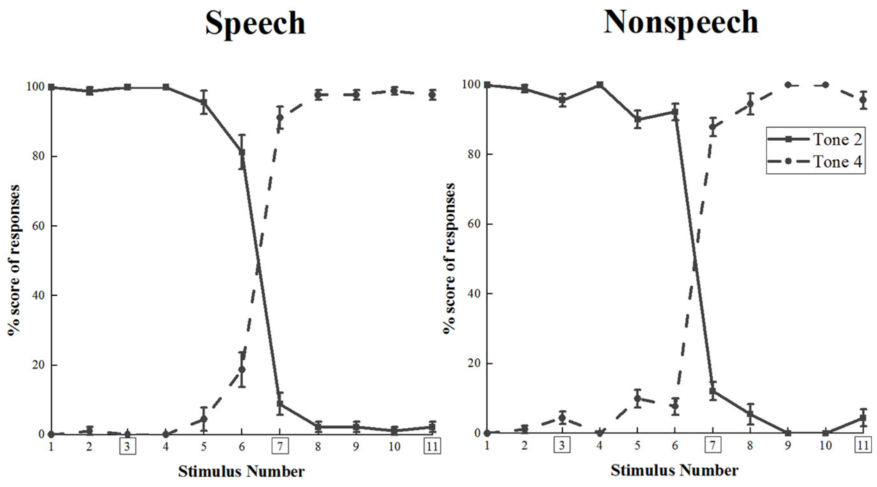


FIGURE 2 | The identification curves for the speech and non-speech stimuli among native Chinese adults (vertical bars represent one stand error). T2 and T4 were coded as S1 and S11; besides, S3, S7 and S11 with rectangles represented across-category deviant, standard and within-category deviant stimuli, respectively.

blocks were presented for all participants in a counterbalanced sequence. The whole experiment lasted around 1 h, including a 5-min break between blocks and a 10-min show of a movie before the tests.

The experiment was conducted in an acoustically and electrically shielded chamber. The participants were seated in front of two active loudspeakers placed to their right and left side with a 45° angle, both of which were kept 0.5 m distance from their ears. An electronic tablet was provided for the participants to play a movie that none had already watched to distract their attention from the sounds. Although the movie was kept silent, the subtitles appeared as normal. The participants were told that they should watch the movie carefully in the whole process because questions would be asked about the movie before and after the EEG recording. For example, prior to the EEG recording, the participants needed to answer one question after viewing the movie. When the right answer was provided, the formal EEG recording would proceed during which participants were instructed to minimize head motion and eye blinking while sitting quietly in the reclined chair.

EEG Recording

An EGI GES 410 system with 64 channel HydroCel GSN electrode nets was employed for the EEG data collection. The vertex (Cz) was settled as the reference electrode when the continuous EEG data were recorded. The vertical and horizontal electrooculograms were monitored by the electrodes placed on the supra- and infra-orbital ridges of each eye and the electrodes near the outer canthi of each eye, respectively. The data were digitized at 1 kHz and amplified with a band-pass filter of 0.5–30 Hz. The impedance of each contact channel was maintained below 50 kΩ (Electrical Geodesics, 2006).

Data Analysis

The EEG data were analyzed off-line with custom scripts and EEGLAB running in the MATLAB environment (Mathworks Inc., United States). With re-reference to the average of all electrodes, the data were adjusted by eliminating the interference of horizontal and vertical eye-movements. The recordings were off-line band-pass filtered with 1–30 Hz and segmented into a 700-ms time window with a 100-ms pre-stimulus baseline. The baseline was then corrected and the recorded trials with ocular or movement artifacts were rejected if they exceeded the range of −50 to 50 μV. Only those data with at least 80 accepted deviant trials for each deviant type were used. The ERPs elicited by standard and deviant stimuli were computed on average of trails of each participant, whereby the MMNs were obtained through the deviant-minus-standard formula.

Consistent with the extant literature, three recording sites of F3, F4, and Fz were selected for statistical analysis (Xi et al., 2010). The time window for MMN typically peaks around 200–350 ms based on the studies by Näätänen et al. (1978, 2007). As shown by Yu et al. (2014), there exist multiple time windows for MMN in different experiments, such as 100–350 ms, 150–300 ms, and 230–360 ms. In line with a recent study by Luck and Gaspelin (2017), an approach termed "Collapsed Localizers" (which is becoming increasingly common) was applied to identify the current time window for MMN[1]. The MMNs were firstly obtained by subtraction of ERP waveforms of the standard from those of the deviants for both conditions. After obtaining MMNs, these difference waveforms were averaged across all participants (AM and NM) and conditions (speech and non-speech), whereby the collapsed waveform was unbiasedly inspected (without showing group and condition differences). In the present study, the time window for MMN was fixed at 100–300 ms based on this collapsed waveform. The MMN mean amplitude was computed as the mean voltage from the range of 20 ms before and after the MMN peak at Fz. The statistical analyses of MMN mean amplitude and MMN peak latency were implemented on the three chosen recording electrodes (F3, F4, and Fz).

RESULTS

The grand average waveforms of the ERPs elicited by the standard and deviant stimuli in speech and non-speech conditions at three locations of F3, F4, and Fz are presented in **Figure 4**. The MMNs obtained via deviant-minus-standard formula of the ERPs for both conditions at F3, F4, and Fz are portrayed in **Figure 5**. Two three-way repeated measures analyses of variance (ANOVAs) were conducted for MMN peak latency and MMN mean amplitude, respectively, with Condition (speech and non-speech) and Deviant type (within-category and across-category stimuli) as two within-subject factors, and Group (AM and NM) as the between-subject factor. For all analyses, the degrees of freedom were adjusted according to the Greenhouse–Geisser method.

MMN Mean Amplitude

The MMN mean amplitudes are shown in **Figure 6**, which presents the clear differences between AM and NM groups in the processing of within-category deviants. It can be seen that the differences become less pronounced when both groups perceived across-category deviants. ANOVA indicated that the main effect of Group was not significant, $F(1,28) = 1.875$, $p = 0.182$, the main effect of Condition was not significant, $F(1,28) = 0.243$, $p = 0.626$, and the main effect of Deviant type was not significant, $F(1,28) = 1.425$, $p = 0.243$. However, a marginally significant interaction between Deviant type and Group was yielded, $F(1,28) = 3.962$, $p = 0.056$. Further simple effects analysis for this interaction revealed that, regardless of speech or non-speech condition, the AM group showed significantly larger MMN mean amplitude than the NM group in the processing of within-category deviants, $F(1,28) = 5.211$, $p < 0.05$. For across-category deviants, there was no significant difference between the two groups in terms of MMN mean amplitude, $F(1,28) = 0.004$, $p = 0.951$. Moreover, for the NM group, MMN mean amplitude evoked by across-category stimuli was significantly larger than that by within-category stimuli, $F(1,28) = 5.07$, $p < 0.05$, but there was no significant difference

[1] The way of Collapsed Localizers has been introduced in Luck and Gaspelin (2017), depicting that the researchers average the waveforms across conditions and then use the timing and scalp distribution from the collapsed waveforms to define the analysis parameters for the non-collapsed data.

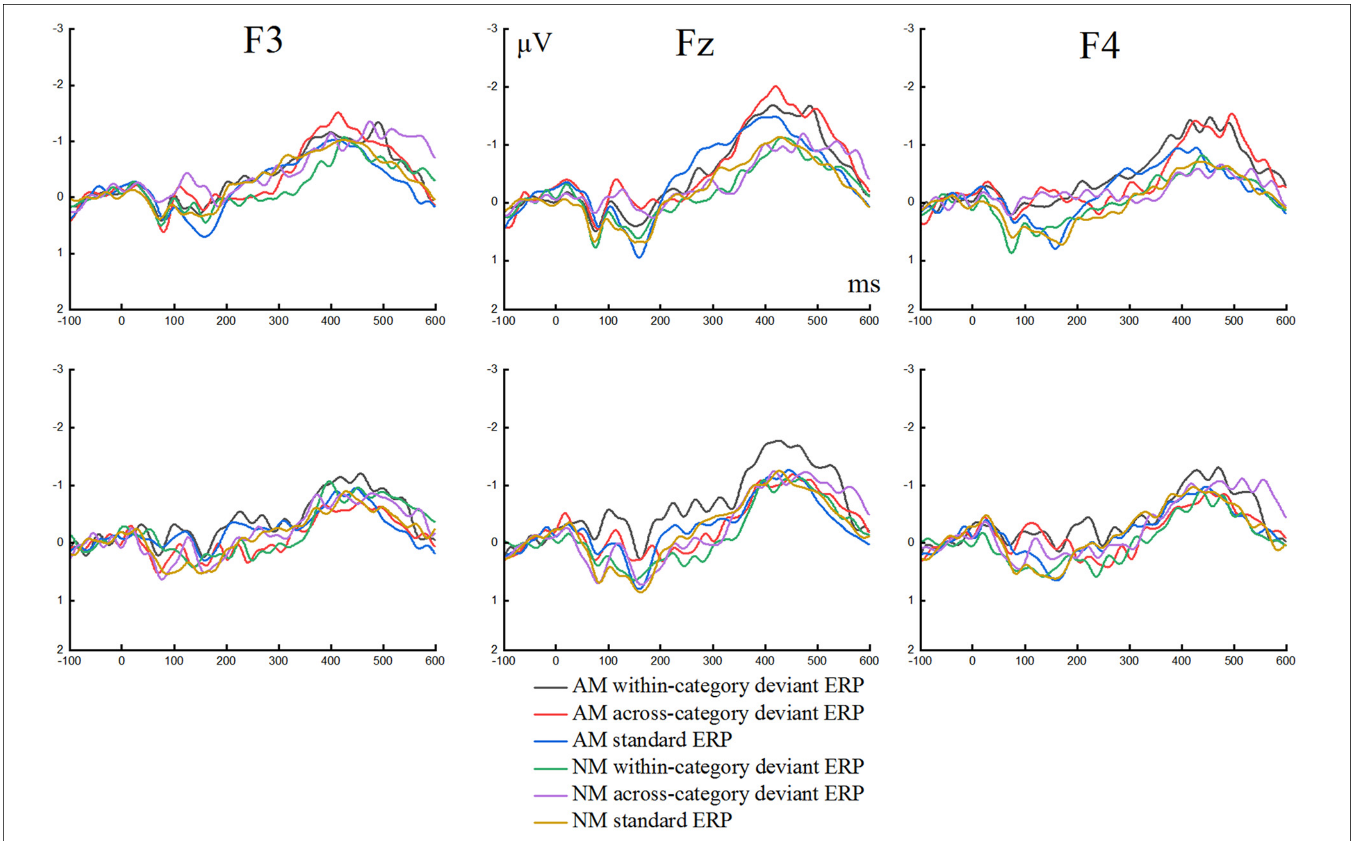


FIGURE 4 | Grand average waveforms elicited by standard and deviant stimuli in speech condition **(the upper row)** and non-speech condition **(the lower row)** at three electrodes for amateur musicians (AM) and non-musicians (NM).

between deviant types for the AM group, *F*(1,28) = 0.317, *p* = 0.578. The interaction between Deviant type, Condition and Group was not significant, *F*(1,28) = 0.176, *p* = 0.678. Taken together, the results of MMN mean amplitude confirmed that amateur musicians were enhanced in processing within-category deviants and more sensitive in detecting pitch shifts, evidenced through their larger MMN mean amplitude, as compared to non-musicians.

MMN Peak Latency

The MMN peak latencies are displayed in **Figure 7**, which shows the clear differences between across-category and within-category deviants in the speech condition, whereas the distinctions are less prominent in the non-speech condition. ANOVA revealed that the main effect of Group was not significant, *F*(1,28) = 0.002, *p* = 0.961, the main effect of Condition was not significant, *F*(1,28) = 0.551, *p* = 0.464, but there was a significant main effect of Deviant type, *F*(1,28) = 6.428, *p* < 0.05, across-category deviant < within-category deviant. No significant interaction was found between Deviant type and Group, *F*(1,28) = 1.618, *p* = 0.214, and no significant three-way interaction was found between Deviant type, Condition and Group, *F*(1,28) = 0.381, *p* = 0.542. Meanwhile, the other effects did not reach statistical significance (*ps* > 0.1). The results of MMN peak latency indicated that both amateur musicians and non-musicians perceived across-category deviants earlier than within-category deviants at the pre-attentive cortical stage.

DISCUSSION

Results of the ERP measurements indicated that both AM and NM groups provoked the significantly shorter MMN peak latency for across-category stimuli than within-category stimuli, which partially certifies our hypotheses of latency showing that not only these two types of information were processed concurrently but phonological information was processed prior to acoustic information of lexical tones at early pre-attentive stage, irrespective of speech or non-speech condition. Meanwhile, the AM group exhibited the significantly larger MMN mean amplitude than the NM group in the processing of within-category deviants in both speech and non-speech conditions, which certifies our hypotheses of amplitude indicative of the AM group's better automatic discrimination of pitch at the pre-attentive cortical stage. These findings manifest that amateur musicians and non-musicians differ in their MMN profiles, suggesting that perceptual processing of lexical tones by amateur musicians is divergent from that by their non-musician counterparts via distinctive neurocognitive mechanisms. This

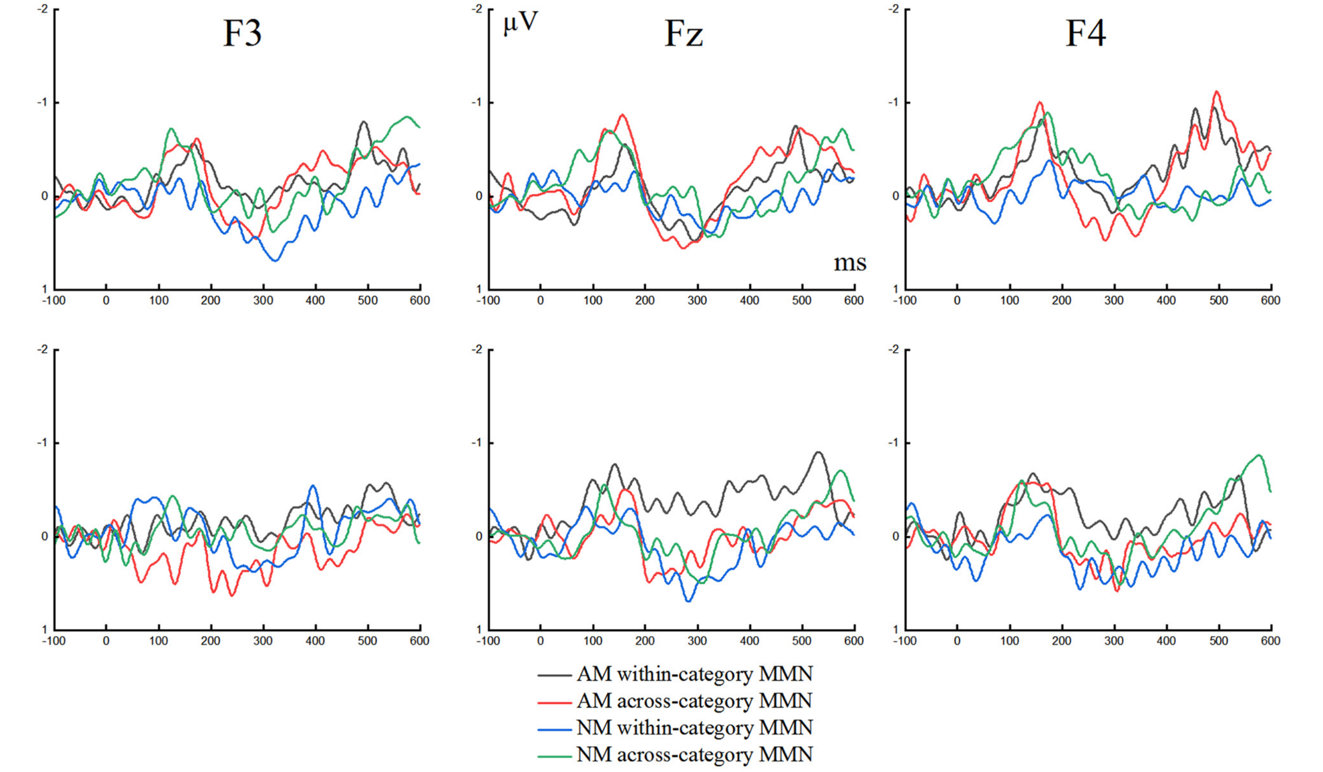


FIGURE 5 | The difference waveforms evoked by across-category and within-category changes in speech condition **(the upper row)** and non-speech condition **(the lower row)** at three electrodes for amateur musicians (AM) and non-musicians (NM).

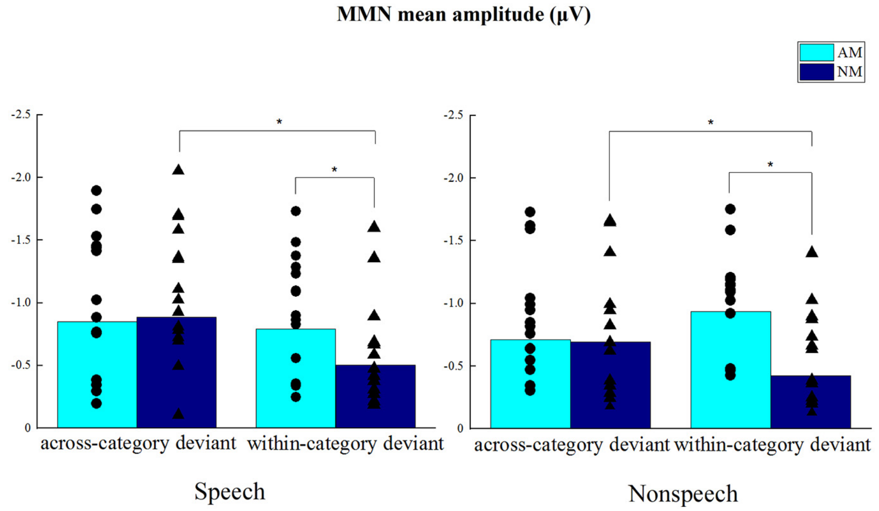


FIGURE 6 | MMN mean amplitudes from electrodes F3, F4, and Fz in respective speech and non-speech conditions. *$p < 0.05$.

lends support to the notion that musicianship modulates categorical perception of lexical tones. Besides, an association between language and music, even at an amateur level of musical expertise, is evidenced by the empirical data. Furthermore, alterations in plasticity can also be induced in adulthood, since the facilitative effects from music to linguistic pitch processing appeared for adult native speakers who had preexisting long-term tone language experience. Coming out of the traditional conception, this proves a novel point that the advantageous effects from music to speech processing can be obtained by a large population of amateurs rather than only by a small group of experts.

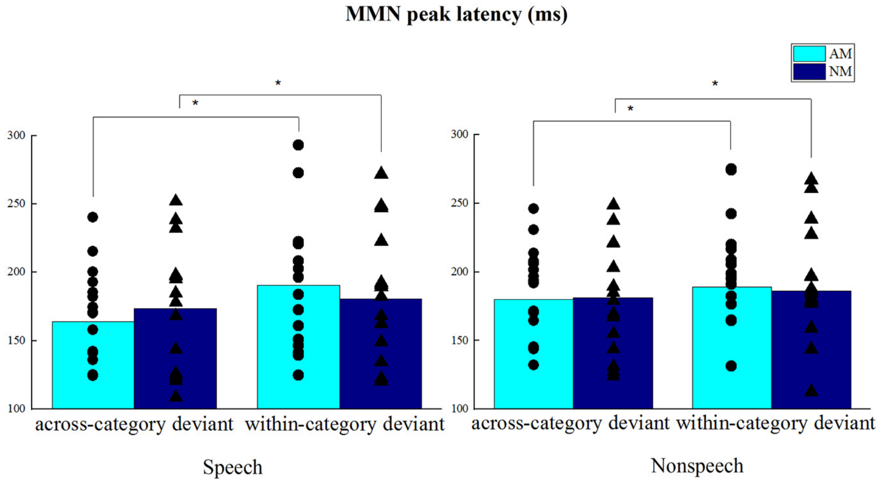


FIGURE 7 | MMN peak latencies from electrodes F3, F4, and Fz in respective speech and non-speech conditions. *$p < 0.05$.

Enhanced Acoustic Processing but Comparable Phonological Processing of Lexical Tones Between AM and NM

The results exhibited that amateur musicians provoked the significantly larger MMN mean amplitude when processing within-category deviants than non-musicians independent of speech or non-speech condition, suggesting that amateur musicians were more sensitive to acoustic information of lexical tones. According to the majority of previous studies (Xu et al., 2006; Peng et al., 2010; Yu et al., 2014, 2017; Shen and Froud, 2016), native tone language users mainly decode phonological information when across-category comparisons are heard; nonetheless, the discrimination of within-category comparisons demands fine-grained pitch resolution. The experimental stimuli (S3, S7, and S11) were adopted from the Mandarin T2–T4 continuum. Viewed from the perspective of signal properties, although S3 and S11 both maintained four steps apart from S7, the combinations of S3–S7 and S7–S11 were essentially different in perception for native speakers. For the pairs of S3–S7 and S7–S11, the former tones had straddled the categorical boundary between T2 and T4, yet the latter only belonged to the same category of T4. Therefore, the discrepancies between across-category (S3–S7) and within-category (S7–S11) comparisons influenced perceptual processing differentially. Moreover, considering that other cues, such as duration and intensity, had already been normalized, and the perception of Mandarin T2 and T4 is determined by pitch variations instead of some confounding features (Li and Chen, 2015; Zhao and Kuhl, 2015a), the results that the AM group showed larger MMN mean amplitude in perceiving within-category deviants implied that the AM participants had stronger pitch-processing abilities as compared to their non-musician counterparts.

Like in the speech condition, the higher MMN mean amplitude in perceiving within-category deviants was evoked in the AM group than the NM group in the non-speech condition. This could be the result of increased demands for the acoustic analysis of pitch in the non-speech condition (Xu et al., 2006), and amateur musicians had more experience with fine-grained musical pitch by playing orchestral instruments (Sluming et al., 2002; Vuust et al., 2012; Patel, 2014). It is believed that in the speech condition, lexical tones distinguish the semantic meanings of words at a higher lexical level, and hence, phonological representations are exploited to a larger extent than acoustic processing by native Mandarin listeners; however, in the non-speech condition, the internal acoustic analysis for tone analogs tends to be implemented (Xu et al., 2006; Peng et al., 2010; Xi et al., 2010). As proposed by Wang (1973), lexical tones and segments, including nuclear vowels and optional consonants, are compulsory elements of Mandarin syllables. The lack of segments led pure tones to be non-speech signals, even though pure tones were interpolated with congruent cues of pitch, intensity, and duration to lexical tones. From this perspective, the perception of pure tones was largely contingent on the processing of pitch information, thereby in the non-speech condition, amateur musicians also showed improved performance in the processing of within-category deviants than non-musicians.

As regards categoricality of lexical tone perception, only non-musicians were observed to provoke the significantly larger MMN mean amplitude in the processing of across-category deviants than within-category deviants, suggesting that amateur musicians were less categorical than non-musicians in the current study. This was in partial agreement with a recent study which showed that Mandarin-speaking musicians do not consistently perceive native lexical tones more categorically than non-musicians (Chen et al., 2020). Based on a latest study by Maggu et al. (2021) that focused on absolute pitch (AP, the ability to name or produce a pitch without a reference) and found that listeners with AP are more sensitive to both across-category and within-category distinctions of lexical tones compared to their non-AP counterparts, our AM members might be non-AP listeners whereby they did not outperform NM in the processing of across-category deviants. This also complies with the study by Levitin and Menon (2003) demonstrating that it is very rare for individuals to have AP when they started musical experience after age 6. Crucially, the insignificant group difference of amplitudes in perceiving across-category deviants uncovered the dominant higher-level influences of linguistic categories relating to lexical tones (Zhang et al., 2011; Zhao and Kuhl, 2015b; Si et al., 2017). Zhao and Kuhl (2015b) found that Mandarin musicians' overall sensitivity to lexical tones links with musical pitch scores, suggesting lower-level contributions; however, Mandarin musicians' sensitivity to lexical tones along a continuum remains analogous to non-musicians. In the study by Wu et al. (2015), no group difference was observed in terms of across-category discrimination accuracy and peakedness in the discrimination function between Mandarin musicians and non-musicians when processing lexical tones. Similar to the studies mentioned above, both groups in the present study were already tonal-language experts and the phonetic inventories for the native language had been acquired and refined early in their lives (Zhang et al., 2005). In other words, as revealed by prior studies (Kuhl, 2004; Best et al., 2016; Chen F. et al., 2017), our listeners had developed robust linguistic representations before their inception of musicianship, which was consequently resistant to plastic changes driven by music (Besson et al., 2011a,b; Tang et al., 2016). Therefore, we had not tracked any clues to mirror that amateur musicians were augmented in tonal representations. This also echoes a study researching segmental vowel perception, which showed that musicians were not advantageous in identifying native vowels and thus they had no strengthened internal representations of native phonological categories in comparison to the non-musician counterparts (Sadakata and Sekiyama, 2011).

In the current study, the AM group outperformed the NM group in the perception of within-category deviants regardless of being speech or non-speech, indicating their superior abilities in pitch processing across domains, in line with Chandrasekaran et al. (2009). This speculation provides novel but persuasive support for OPERA-e (Patel, 2014) in that musical-pitch extends to lexical-pitch detection, even when derived from amateur musical experience. The facilitative pitch processing could be explained by taking into account the differences in pitch precision between music and lexical tones. According to OPERA-e, pitch variations in music can be smaller in frequency size than that in language, thereby pitch precision from music plays a profitable role in perceiving within-category lexical tone stimuli (Patel, 2014). Previous studies identified that a pitch interval as small as one semitone remains perceptually salient in music. For example, a C versus a C# in the key of C can be explicitly discerned (Patel, 2014; Tang et al., 2016). Nonetheless, for categorical perception of lexical tones, the smallest frequency range for discrimination is about 4–8 Hz for normal Chinese speakers in light of just-noticeable differences (JNDs, Liu, 2013). Pitch threshold is important in lexical tone perception. Amusia is a musical-pitch disorder influencing both music and speech processing (Peretz et al., 2002, 2008; Tillmann et al., 2011; Vuvan et al., 2015). Tone agnosics, a subgroup of individuals with amusia (Nan et al., 2010), struggle to perceive fine-grained lexical tones because the elevated pitch threshold ranges from 20 to 30 Hz (Huang et al., 2015a,b), which results in their impoverished performance as compared to typical listeners in categorical perception of lexical tones (Zhang et al., 2017). In the current study, as measured via Praat, the tonal stimuli were about 9 Hz distant for every step along the continuum of Mandarin T2–T4, and both deviants (S3 and S11) were four steps apart from the standard (S7). The frequency size of stimuli in the present study was far larger than that in music and JNDs. Thus, within-category pitch differences were detected more easily by amateur musicians than non-musicians.

However, other possibilities for enhanced within-category pitch perception should be acknowledged. Although the present sample demographic characteristics were well-controlled (Ayotte et al., 2002; Patel, 2014; Tang et al., 2016), the capacities of lexical tone perception in the AM participants before they started their musical experience were unknown. In other words, some participants might be sensitive to pitch information prior to their musical experience. Longitudinal studies with pre- and post-tests are thus highly recommended to further estimate the effects of amateur musical experience on speech processing. Some heritable differences in auditory functions should also be cautiously controlled (Drayna et al., 2001), since there exist naturally occurring variations in pitch perception capacities (Moreno et al., 2008; Qin et al., 2021). Meanwhile, although the overall gender ratio was nearly equal, different males and females were found between AM and NM groups. Previous studies claim that there are no gender effects in amplitude and latency of MMN among male and female participants (Kasai et al., 2002; Ikezawa et al., 2008; Tsolaki et al., 2015; Yang et al., 2016), while others hold an opposite position (Aaltonen et al., 1994; Barrett and Fulfs, 1998). Future studies should try to exclude the inconclusive effects by gender.

Earlier Processing of Phonological Information Than Acoustic Information of Lexical Tones by Mandarin Listeners

The results showed that in both groups, across-category deviants elicited the significantly shorter MMN peak latency than within-category deviants, suggesting that phonological processing precedes acoustic processing for lexical tones. Note that unlike MMN mean amplitude, the two groups showed comparable MMN peak latency as revealed by the significant main effect of Deviant type. No group difference in terms of MMN peak latency might be ascribed to musical experience; in other words, the AM participants were not expert musicians, which mediated their abilities of pitch processing. For this reason, together with MMN mean amplitude and peak latency, the findings possibly suggest that the effects of amateur musical experience on lexical tone perception are somehow constrained. Therefore, only if musical expertise reaches the professional level, then a significant difference between the two groups in terms of latency can be anticipated. This tentative speculation needs to be further elaborated in future studies.

The current results relating to MMN peak latency provide counter-evidence to the findings by Luo et al. (2006) concerning the two-stage model. According to Luo et al. (2006), acoustic and phonological information about lexical tones are processed at pre-attentive and attentive stages, respectively. Nevertheless, many recent studies have shown that both acoustic and phonological information might be processed at pre-attentive and attentive stages in parallel (Xi et al., 2010; Yu et al., 2014, 2017). In accordance with these studies, the present results revealed that across-category stimuli were processed earlier than within-category stimuli, indicating that phonological information was processed ahead of acoustic information at pre-attentive stage, which differed from that proposed in the serial model. It is worth noting that the mentioned studies (Luo et al., 2006; Xi et al., 2010; Yu et al., 2014) recruited non-musician participants with similar neurocognitive mechanisms for processing lexical tones. Although the findings from Yu et al. (2014, 2017) comply with the notion that phonological information can be processed earlier than acoustic information, the current study further shows that this pattern of temporal processing occurs regardless of listeners' musical background. In contrast to the studies using only non-musicians as participants, amateur musicians in the current study performed similarly to professional musicians showing advantages in processing within-category deviants, suggesting their enhanced processing of acoustic information (Wu et al., 2015; Chen et al., 2020). However, analogous to non-musicians, these amateur musicians still elicited the significantly longer MMN peak latency for within-category deviants than across-category deviants. This indicated that irrespective of the strengthened processing of acoustic information, phonological information was processed earlier than acoustic information at the pre-attentive cortical stage, as opposed to the two-stage model (Luo et al., 2006); besides, it confirmed the dominant role of higher-level linguistic categories relating to lexical tones from a neural perspective (Zhao and Kuhl, 2015b; Si et al., 2017). This provides the meaningful insight to the neural mechanisms which underlie the perceptual processing of lexical tones.

Presumably, some factors contributing to the current results of latency are worthy of consideration. First, the earlier processing of across-category than within-category deviants might be attributable to the properties of MMN. Näätänen (2001) illustrated that in auditory presentation, the differences between several infrequent deviant stimuli embedded in a flow of frequent and repeated standard stimuli can be automatically detected as signaled by MMN, with stronger incongruity leading to shorter latency onset. In the current study, across-category stimuli (S3) diverged from the standard (S7) in both phonological information of categories and acoustic information of pitch, while within-category stimuli (S11) only differed in acoustic information. Therefore, a shorter MMN peak latency was evoked by S3 than S11. Second, it is argued that tonal representations influence lexical tone processing (Xu et al., 2006; Chandrasekaran et al., 2007a; Chen S. et al., 2017). Although pure tones were non-speech sounds, the perception of them might be facilitated as a function of long-term phonological memory traces for lexical tones (Chandrasekaran et al., 2007a; Kraus and White-Schwoch, 2017). As explicated in Yu et al. (2014), the activation of long-term memory traces means that phonological information relating to lexical categories has some effects on the pitch detection of non-speech analogs, which copy identical acoustic cues from lexical tones. In the current study, the perception of pure tones as non-speech analogs to lexical tones might be impacted by memory traces for lexical tones, through which across-category deviants also elicited shorter latencies than within-category deviants in the non-speech condition regardless of group.

Some researchers have addressed that pitch type influences the temporal processing of lexical tones. As demonstrated in previous studies, T2 and T3 are acoustically similar such that their perception even burdens native speakers (Shen and Lin, 1991; Hao, 2012). Accordingly, Chandrasekaran et al. (2007b) found that MMN peak latency of T1–T3 is shorter than that of T2–T3 in Mandarin. Chandrasekaran et al. (2007b) concluded that MMN peak latency can be impacted by pitch type and likewise, the study by Yu et al. (2017) systematically investigated pitch type and latency, showing that pitch height is always processed ahead of pitch contour. The current study revealed that across-category deviants (S3) were processed earlier than within-category deviants (S11) with S7 being the standard. As shown in **Figure 1**, the contours of S3 versus S7 were more different than S11 versus S7 in slope, with a larger interval at onset point in terms of pitch height. In this regard, the differences in across-category deviants could be detected earlier, thus eliciting a shorter MMN peak latency in contrast to within-category deviants (Yu et al., 2017). This finding adds a further line of evidence supporting that pitch type is associated with MMN peak latency (Chandrasekaran et al., 2007b; Yu et al., 2017).

Re-categorization of Participants' Musicianship in Tests of Pitch Processing

From the methodological perspective, the results mentioned above require us to re-consider the categorization of participants' musical experience. First, the current study emphasizes the importance of characterizing participants in light of their musical experience, because individuals with musical experience tend to outperform their non-musician counterparts in this field of research. Second, there have been various choices to select non-musicians without a conventional standard. For instance, non-musicians' musical practice ranges differently from 0 to 3 years in previous studies (e.g., Alexander et al., 2005; Wong et al., 2007; Maggu et al., 2018). Besides, individuals with non-professional musical training may be unsuitably regarded as non-musicians (Shen and Froud, 2016, 2019), and even reporting musical background has been occasionally neglected in some studies of lexical tone processing (Hao, 2012; Morett, 2019; Qin et al., 2019). The neural evidence provided by the current study showed that similar to professional musicians, amateur musicians as non-music majors with around 4-year musicianship are strengthened in acoustic processing of lexical tones. Those listeners with a limited duration of musical experience (e.g., 3 years) might have already been affected regarding their pitch-detection abilities; hence, the criteria for screening non-musicians in the future should be stricter than in the past. Moreover, although we used the comparative approach to analyze lexical tone processing by amateurs and experts from previous literature, it is recommended to directly recruit one more group of professional musicians so as to systematically research the effects of magnitudes of musical expertise on speech processing.

The present results verified that amateur musical experience modulates categorical perception of lexical tones for native adults (i.e., enhanced within-category but comparable across-category lexical tone processing) though they have preexisting long-term tone language experience. In accordance with previous studies (Gordon et al., 2015; Zhao and Kuhl, 2015a,b), the current study supports the conceptual framework of OPERA-e by highlighting that the perceptual demands required for musical practice benefit the neural systems that are crucial for speech perception (Patel, 2014). Findings also echo those of previous studies indicating that music can be applied to prompt language skills in both normal and clinical populations due to the facilitative effects from music impacting on language (Won et al., 2010; Herholz and Zatorre, 2012; Petersen et al., 2015; Gfeller, 2016).

Since Duncan et al. (2009) identified that the mean amplitudes of the ERP components link with the volume of neural resources engaging in brain activities, future studies are advised to continue researching hemispheric processing of lexical tones by amateur musicians[2]. Note that for the analysis of lateralization among future studies, variance of neural data should be reduced by handedness, given that there is a strong bias of handedness on cerebral lateralization and left-handers may show anomalous dominance patterns (Cai and Van der Haegen, 2015; Plante et al., 2015). Moreover, conducive to generating some ecological impacts, these findings might also encourage individuals to engage in music in either formal or informal ways, thus aiding their aesthetic development as well as helping protect against cognitive decline (Román-Caballero et al., 2018). In addition, as validated by Näätänen et al. (2004) and Pakarinen et al. (2007), the MMNs obtained via the multifeature passive oddball paradigm were equal in amplitude to those via the traditional MMN paradigm. However, it should be treated with caution when calculating MMNs, because different endogenous ERPs would be generated by this multifeature passive oddball paradigm. Due to the potential for physical confounds existing among auditory stimuli, MMNs can be obtained by subtraction from a given sound when it is a standard to the exact same sound when it is a deviant (Schröger and Wolff, 1996). Future studies are encouraged to tap this paradigm in experiments. Lastly, as an important auditory ERP component, MMN can provide an objective marker to measure the abilities of amateur musicians to discriminate lexical tones.

CONCLUSION

In summary, the current study explored cortical plasticity among adult amateur musicians, taking advantage of neurophysiological MMN indices. Although participants were native speakers of Mandarin, the results of the MMN mean amplitude indicated that the abilities to process acoustic information by amateur musicians were enhanced in terms of categorical perception of Mandarin lexical tones. Higher sensitivity for pitch shifts across domains confirmed that speech perception can be modulated by amateur musical experience in adulthood, and music associates with language even only an amateur level of musical expertise is reached by listeners. This indicated that the advantageous effects of music on speech processing are not restricted to a small group of professional musicians but extend to a large population of amateur musicians. In addition, a shorter latency was evoked by across-category deviants than that by within-category deviants, suggesting that these two types of information can be processed concurrently at the pre-attentive cortical stage; more precisely, the processing of phonological information is earlier than that of acoustic information, even for amateur musicians whose acoustic processing was strengthened for lexical tones.

DATA AVAILABILITY STATEMENT

The original contributions generated for this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Review Board at the School of Foreign Languages of Hunan University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JZ and XC contributed to the conception of the study. JZ conducted the experiments and drafted the manuscript. XC and YY contributed to the revision of the manuscript. All authors have approved the final version of the manuscript.

FUNDING

This research was supported by Social Science Foundation of Ministry of Education of China (20YJC740041) and Hunan Provincial Social Science Foundation of China (19YBQ112).

[2] Following Xi et al. (2010), we compared amplitudes of F3 and F4 and added this as an extra within-subject factor Electrode to probe into the hemispheric pattern, which indicated that the main effect or interactions of Electrode with other factors were not significant (ps > 0.1). Despite the low spatial resolution of ERP, the result as a preliminary sign manifested that the perception of Mandarin lexical tones is supported by neither one specific area nor a single hemisphere (Gandour et al., 2004; Witteman et al., 2011; Price, 2012; Si et al., 2017; Liang and Du, 2018; Shao and Zhang, 2020).

REFERENCES

Aaltonen, O., Eerola, O., Lang, A. H., Uusipaikka, E., and Tuomainen, J. (1994). Automatic discrimination of phonetically relevant and irrelevant vowel parameters as reflected by mismatch negativity. *J. Acoust. Soc. Am.* 96, 1489–1493. doi: 10.1121/1.410291

Abramson, A. S. (1978). Static and dynamic acoustic cues in distinctive tones. *Lang. Speech* 21, 319–325. doi: 10.1177/002383097802100406

Alexander, J. A., Wong, P. C. M., and Bradlow, A. R. (2005). "Lexical tone perception in musicians and non-musicians," in *Proceedings of the Interspeech 2005*, (Lisbon: ISCA Archive), 397–400.

Ayotte, J., Peretz, I., and Hyde, K. (2002). Congenital amusia—a group study of adults afflicted with a music-specific disorder. *Brain* 125, 238–251. doi: 10.1093/brain/awf028

Barrett, K. A., and Fulfs, J. M. (1998). Effect of gender on the mismatch negativity auditory evoked potential. *J. Am. Acad. Audiol.* 9, 444–451.

Besson, M., Chobert, J., and Marie, C. (2011a). Language and music in the musician brain. *Lang. Linguist. Compass* 5, 617–634. doi: 10.1111/j.1749-818x.2011.00302.x

Besson, M., Chobert, J., and Marie, C. (2011b). Transfer of training between music and speech: common processing, attention, and memory. *Front. Psychol.* 2:94. doi: 10.3389/fpsyg.2011.00094

Best, C. T. (1994). "The emergence of native-language phonological influences in infants: a perceptual assimilation model," in *The Development of Speech Perception: The Transition From Speech Sounds to Spoken Words*, eds J. C. Goodman and H. C. Nusbaum (Cambridge, MA: MIT Press), 167–224.

Best, C. T. (1995). "A direct-realist view of cross-language speech perception," in *Speech Perception And Linguistic Experience: Issues In Cross-Language Research*, ed. W. Strange (Timonium, MD: York Press), 171–204.

Best, C. T., Goldstein, L. M., Nam, H., and Tyler, M. D. (2016). Articulating what infants attune to in native speech. *Ecol. Psychol.* 28, 216–261. doi: 10.1080/10407413.2016.1230372

Best, C. T., and Tyler, M. D. (2007). "Nonnative and second-language speech perception: commonalities and complementarities," in *Second Language Speech Learning: The Role Of Language Experience In Speech Perception And Production*, eds J. Munro and O. S. Bohn (Amsterdam: John Benjamins), 13–34. doi: 10.1075/lllt.17.07bes

Boersma, P., and Weenink, D. (2019). *Praat: Doing Phonetics by Computer (Version 6.1.05) [Computer program]*. Available online at: http://www.praat.org/ (accessed October 16, 2019)

Cai, Q., and Van der Haegen, L. (2015). What can atypical language hemispheric specialization tell us about cognitive functions? *Neurosci. Bull.* 31, 220–226. doi: 10.1007/s12264-014-1505-5

Chandrasekaran, B., Krishnan, A., and Gandour, J. T. (2007a). Experience dependent neural plasticity is sensitive to shape of pitch contours. *Neuroreport* 18, 1963–1967. doi: 10.1097/wnr.0b013e3282f213c5

Chandrasekaran, B., Krishnan, A., and Gandour, J. T. (2007b). Mismatch negativity to pitch contours is influenced by language experience. *Brain Res.* 1128, 148–156. doi: 10.1016/j.brainres.2006.10.064

Chandrasekaran, B., Krishnan, A., and Gandour, J. T. (2009). Relative influence of musical and linguistic experience on early cortical processing of pitch contours. *Brain Lang.* 108, 1–9. doi: 10.1016/j.bandl.2008.02.001

Chao, Y. R. (1947). *Cantonese primer*. Cambridge, MA: Harvard University Press.

Chao, Y. R. (1948). *Mandarin Primer*. Cambridge, MA: Harvard University Press.

Chao, Y. R. (1968). *A grammar of spoken Chinese*. Berkeley, CA: University of California Press.

Chen, F., and Peng, G. (2018). Lower-level acoustics underlie higher-level phonological categories in lexical tone perception. *J. Acoust. Soc. Am.* 144, EL158–EL164. doi: 10.1121/1.5052205

Chen, F., Peng, G., Yan, N., and Wang, L. (2017). The development of categorical perception of Mandarin tones in four- to seven-year-old children. *J. Child Lang.* 44, 1413–1434. doi: 10.1017/s0305000916000581

Chen, J. K., Chuang, A. Y., McMahon, C., Hsieh, J. C., Tung, T. H., and Li, L. P. (2010). Music training improves pitch perception in prelingually deafened children with cochlear implants. *Pediatrics* 125, e793–e800. doi: 10.1542/peds.2008-3620

Chen, S., Zhu, Y., and Wayland, R. (2017). Effects of stimulus duration and vowel quality in cross-linguistic categorical perception of pitch directions. *PLoS One* 12:e0180656. doi: 10.1371/journal.pone.0180656

Chen, S., Zhu, Y., Wayland, R., and Yang, Y. (2020). How musical experience affects tone perception efficiency by musicians of tonal and non-tonal speakers? *PLoS One* 15:e0232514. doi: 10.1371/journal.pone.0232514

Cho, T., and Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *J. Phon.* 27, 207–229. doi: 10.1006/jpho.1999.0094

Cooper, A., and Wang, Y. (2012). The influence of linguistic and musical experience on Cantonese word learning. *J. Acoust. Soc. Am.* 131, 4756–4769. doi: 10.1121/1.4714355

Darwin, C. (1871). *The Descent Of Man And Selection In Relation To Sex*. New York, NY: Cambridge University Press, doi: 10.5962/bhl.title.110063

Dittinger, E., Barbaroux, M., D'Imperio, M., Jäncke, L., Elmer, S., and Besson, M. (2016). Professional music training and novel word learning: From faster semantic encoding to longer-lasting word representations. *J. Cogn. Neurosci.* 28, 1584–1602. doi: 10.1162/jocn_a_00997

Drayna, D., Manichaikul, A., de Lange, M., Snieder, H., and Spector, T. (2001). Genetic correlates of musical pitch recognition in humans. *Science* 291, 1969–1972. doi: 10.1126/science.291.5510.1969

Duncan, C. C., Barry, R. J., Connolly, J. F., Fischer, C., Michie, P. T., Näätänen, R., et al. (2009). Event-related potentials in clinical research: guidelines for eliciting, recording, and quantifying mismatch negativity. P300, and N400. *Clin. Neurophysiol.* 120, 1883–1908. doi: 10.1016/j.clinph.2009.07.045

Electrical Geodesics (2006). *Net Station Viewer And Waveform Tools Tutorial, S-MAN-200-TVWR-001*. Eugene, OR: Electrical Geodesics.

Finney, D. J. (1971). *Probit Analysis*, 3rd Edn. Cambridge: Cambridge University Press.

Francis, A. L., Ciocca, V., and Ng, B. K. C. (2003). On the (non)categorical perception of lexical tones. *Percept. Psychophys.* 65, 1029–1044. doi: 10.3758/bf03194832

Fry, D. B., Abramson, A. S., Eimas, P. D., and Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Lang. Speech* 5, 171–189. doi: 10.1177/002383096200500401

Fujisaki, H., and Kawashima, T. (1971). A model of the mechanisms for speech perception-quantitative analysis of categorical effects in discrimination. *Annu. Rep. Eng. Res. Inst. Facult. Eng. Univ. Tokyo* 30, 59–68.

Gandour, J. (1983). Tone perception in Far Eastern languages. *J. Phon.* 11, 149–175. doi: 10.1016/s0095-4470(19)30813-7

Gandour, J., and Harshman, R. (1978). Crosslanguage differences in tone perception: a multidimensional scaling investigation. *Lang. Speech* 21, 1–33. doi: 10.1177/002383097802100101

Gandour, J., Tong, Y., Wong, D., Talavage, T., Dzemidzic, M., Xu, Y., et al. (2004). Hemispheric roles in the perception of speech prosody. *Neuroimage* 23, 344–357. doi: 10.1016/j.neuroimage.2004.06.004

Gfeller, K. (2016). Music-based training for pediatric CI recipients: a systematic analysis of published studies. *Eur. Ann. Otorhinolaryngol. Head Neck Dis.* 133, S50–S56. doi: 10.1016/j.anorl.2016.01.010

Gordon, R. L., Fehd, H. M., and McCandliss, B. D. (2015). Does music training enhance literacy skills? A meta-analysis. *Front. Psychol.* 6:1777. doi: 10.3389/fpsyg.2015.01777

Hallé, P. A., Chang, Y. C., and Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *J. Phon.* 32, 395–421. doi: 10.1016/s0095-4470(03)00016-0

Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *J. Phon.* 40, 269–279. doi: 10.1016/j.wocn.2011.11.001

Herholz, S. C., and Zatorre, R. J. (2012). Musical training as a framework for brain plasticity: behavior, function, and structure. *Neuron* 76, 486–502. doi: 10.1016/j.neuron.2012.10.011

Huang, W. T., Liu, C., Dong, Q., and Nan, Y. (2015a). Categorical perception of lexical tones in Mandarin-speaking congenital amusics. *Front. Psychol.* 6:829. doi: 10.3389/fpsyg.2015.00829

Huang, W. T., Nan, Y., Dong, Q., and Liu, C. (2015b). Just-noticeable difference of tone pitch contour change for Mandarin congenital amusics. *J. Acoust. Soc. Am.* 138, EL99–EL104. doi: 10.1121/1.4923268

Hutka, S., Bidelman, G. M., and Moreno, S. (2015). Pitch expertise is not created equal: cross-domain effects of musicianship and tone language experience on neural and behavioural discrimination of speech and music. *Neuropsychologia* 71, 52–63. doi: 10.1016/j.neuropsychologia.2015.03.019

Ikezawa, S., Nakagome, K., Mimura, M., Shinoda, J., Itoh, K., Homma, I., et al. (2008). Gender differences in lateralization of mismatch negativity in dichotic listening tasks. *Int. J. Psychophysiol.* 68, 41–50. doi: 10.1016/j.ijpsycho.2008.01.006

Jiang, C., Hamm, J. P., Lim, V. K., Kirk, I. J., and Yang, Y. (2012). Impaired categorical perception of lexical tones in Mandarin-speaking congenital amusics. *Mem. Cogn.* 40, 1109–1121. doi: 10.3758/s13421-012-0208-2

Joanisse, M. F., Zevin, J. D., and McCandliss, B. D. (2006). Brain mechanisms implicated in the preattentive categorization of speech sounds revealed using fMRI and a short-interval habituation trial paradigm. *Cereb. Cortex* 17, 2084–2093. doi: 10.1093/cercor/bhl124

Kasai, K., Nakagome, K., Iwanami, A., Fukuda, M., Itoh, K., Koshida, I., et al. (2002). No effect of gender on tonal and phonetic mismatch negativity in normal adults assessed by a high-resolution EEG recording. *Brain Res. Cogn. Brain Res.* 13, 305–312. doi: 10.1016/s0926-6410(01)00125-2

Kraus, N., and Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nat. Rev. Neurosci.* 11, 599–605. doi: 10.1038/nrn2882

Kraus, N., and White-Schwoch, T. (2017). Neurobiology of everyday communication: what have we learned from music. *Neuroscientist* 23, 287–298. doi: 10.1177/1073858416653593

Kubovy, M., Cohen, D. J., and Hollier, J. (1999). Feature integration that routinely occurs without focal attention. *Psychon. Bull. Rev.* 6, 183–203. doi: 10.3758/BF03212326

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nat. Rev. Neurosci.* 5, 831–843. doi: 10.1038/nrn1533

Lee, C. Y., Lekich, A., and Zhang, Y. (2014). Perception of pitch height in lexical and musical tones by English-speaking musicians and nonmusicians. *J. Acoust. Soc. Am.* 135, 1607–1615. doi: 10.1121/1.4864473

Levitin, D. J., and Menon, V. (2003). Musical structure is processed in "language" areas of the brain: a possible role for Brodmann Area 47 in temporal coherence. *Neuroimage* 20, 2141–2152. doi: 10.1016/j.neuroimage.2003.08.016

Li, X., and Chen, Y. (2015). Representation and processing of lexical tone and tonal variants: evidence from the mismatch negativity. *PLoS One* 10:e0143097. doi: 10.1371/journal.pone.0143097

Liang, B., and Du, Y. (2018). The functional neuroanatomy of lexical tone perception: an activation likelihood estimation meta-analysis. *Front. Neurosci.* 12:495. doi: 10.3389/fnins.2018.00495

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol.* 54, 358–368. doi: 10.1037/h0044417

Liu, C. (2013). Just noticeable difference of tone pitch contour change for English- and Chinese-native listeners. *J. Acoust. Soc. Am.* 134, 3011–3020. doi: 10.1121/1.4820887

Liu, L., Ong, J. H., Tuninetti, A., and Escudero, P. (2018). One way or another: evidence for perceptual asymmetry in pre-attentive learning of non-native contrasts. *Front. Psychol.* 5:162. doi: 10.3389/fpsyg.2018.00162

Luck, S. J., and Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology* 54, 146–157. doi: 10.1111/psyp.12639

Luo, H., Ni, J. T., Li, Z. H., Li, X. O., Zhang, D. R., Zeng, F. G., et al. (2006). Opposite patterns of hemisphere dominance for early auditory processing of lexical tones and consonants. *Proc. Natl. Acad. Sci. U.S.A.* 103, 19558–19563. doi: 10.1073/pnas.0607065104

Ma, J., Chen, X., Wu, Y., and Zhang, L. (2017). Effects of age and sex on voice onset time: evidence from Mandarin voiceless stops. *Logoped. Phoniatr. Vocol.* 43, 56–62. doi: 10.1080/14015439.2017.1324915

Maggu, A. R., Lau, J. C. Y., Waye, M. M. Y., and Wong, P. C. M. (2021). Combination of absolute pitch and tone language experience enhances lexical tone perception. *Sci. Rep.* 11:1485. doi: 10.1038/s41598-020-80260-x

Maggu, A. R., Wong, P. C. M., Antoniou, M., Bones, O., Liu, H., and Wong, F. C. K. (2018). Effects of combination of linguistic and musical pitch experience on subcortical pitch encoding. *J. Neurolinguistics* 47, 145–155. doi: 10.1016/j.jneuroling.2018.05.003

Magne, C., Schön, D., and Besson, M. (2006). Musician children detect pitch violations in both music and language better than nonmusician children: behavioral and electrophysiological approaches. *J. Cogn. Neurosci*. 18, 199–211. doi: 10.1162/jocn.2006.18.2.199

Marie, C., Kujala, T., and Besson, M. (2012). Musical and linguistic expertise influence pre-attentive and attentive processing of non-speech sounds. *Cortex* 48, 447–457. doi: 10.1016/j.cortex.2010.11.006

Marques, C., Moreno, S., Castro, S. L., and Besson, M. (2007). Musicians detect pitch violation in a foreign language better than non-musicians: behavioral and electrophysiological evidence. *J. Cogn. Neurosci.* 19, 1453–1463. doi: 10.1162/jocn.2007.19.9.1453

Miller, J., and Eimas, P. (1996). Internal structure of voicing categories in early infancy. *Percept. Psychophys.* 58, 1157–1167. doi: 10.3758/bf03207549

Moreno, S. (2009). Can music influence language and cognition? *Contemp. Music Rev.* 28, 329–345. doi: 10.1080/07494460903404410

Moreno, S., Marques, C., Santos, A., Santos, M., Castro, S. L., and Besson, M. (2008). Musical training influences linguistic abilities in 8-year-old children: more evidence for brain plasticity. *Cereb. Cortex* 19, 712–723. doi: 10.1093/cercor/bhn120

Morett, L. M. (2019). The influence of tonal and atonal bilingualism on children's lexical and non-lexical tone perception. *Lang. Speech* 63, 221–241. doi: 10.1177/0023830919834679

Moulines, E., and Laroche, J. (1995). Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Commun.* 16, 175–205. doi: 10.1016/0167-6393(94)00054-e

Näätänen, R. (2001). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology* 38, 1–21. doi: 10.1111/1469-8986.3810001

Näätänen, R., Gaillard, A. W., and Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychol. (Amst)* 42, 313–329. doi: 10.1016/0001-6918(78)90006-9

Näätänen, R., Paavilainen, P., Rinne, T., and Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clin. Neurophysiol.* 118, 2544–2590. doi: 10.1016/j.clinph.2007.04.026

Näätänen, R., Pakarinen, S., Rinne, T., and Takegata, R. (2004). The mismatch negativity (MMN): towards the optimal paradigm. *Clin. Neurophysiol.* 115, 140–144. doi: 10.1016/j.clinph.2003.04.001

Nan, Y., Liu, L., Geiser, E., Shu, H., Gong, C. C., Dong, Q., et al. (2018). Piano training enhances the neural processing of pitch and improves speech perception in Mandarin-speaking children. *Proc. Natl. Acad. Sci. U.S.A.* 115, E6630–E6639. doi: 10.1073/pnas.1808412115

Nan, Y., Sun, Y., and Peretz, I. (2010). Congenital amusia in speakers of a tone language: association with lexical tone agnosia. *Brain* 133, 2635–2642. doi: 10.1093/brain/awq178

Neisser, U. (1967). *Cognitive psychology*. New York, NY: Appleton—Century—Crofts.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113. doi: 10.1016/0028-3932(71)90067-4

Ong, J. H., Wong, P. C. M., and Liu, F. (2020). Musicians show enhanced perception, but not production, of native lexical tones. *J. Acoust. Soc. Am.* 148, 3443–3454. doi: 10.1121/10.0002776

Pakarinen, S., Takegata, R., Rinne, T., Huotilainen, M., and Näätänen, R. (2007). Measurement of extensive auditory discrimination profiles using the mismatch negativity (MMN) of the auditory event-related potential (ERP). *Clin. Neurophysiol.* 118, 177–185. doi: 10.1016/j.clinph.2006.09.001

Pantev, C., Roberts, L. E., Schulz, M., Engelien, A., and Ross, B. (2001). Timbre-specific enhancement of auditory cortical representations in musicians. *Neuroreport* 12, 169–174. doi: 10.1097/00001756-200101220-00041

Partanen, E., Torppa, R., Pykäläinen, J., Kujala, T., and Huotilainen, M. (2013). Children's brain responses to sound changes in pseudo words in a multifeature paradigm. *Clin. Neurophysiol.* 124, 1132–1138. doi: 10.1016/j.clinph.2012.12.005

Patel, A. D. (2008). *Music, Language, And The Brain*. Oxford: University Press.

Patel, A. D. (2014). Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis. *Hear. Res.* 308, 98–108. doi: 10.1016/j.heares.2013.08.011

Peng, G., Zheng, H. Y., Gong, T., Yang, R. X., Kong, J. P., and Wang, W. S. Y. (2010). The influence of language experience on categorical perception of pitch contours. *J. Phon.* 38, 616–624. doi: 10.1016/j.wocn.2010.09.003

Peretz, I., Ayotte, J., Zatorre, R. J., Mehler, J., Ahad, P., Penhune, V. B., et al. (2002). Congenital amusia: a disorder of fine-grained pitch discrimination. *Neuron* 33, 185–191. doi: 10.1016/s0896-6273(01)00580-3

Peretz, I., Gosselin, N., Tillmann, B., Cuddy, L., Gagnon, B., Trimmer, G. C., et al. (2008). On-line identification of congenital amusia. *Music Percept.* 25, 331–343. doi: 10.1525/mp.2008.25.4.331

Petersen, B., Weed, E., Sandmann, P., Brattico, E., Hansen, M., Sørensen, S. D., et al. (2015). Brain responses to musical feature changes in adolescent cochlear implant users. *Front. Hum. Neurosci.* 9:7. doi: 10.3389/fnhum.2015.00007

Plante, E., Almryde, K., Patterson, D. K., Vance, C. J., and Asbjørnsen, A. E. (2015). Language lateralization shifts with learning by adults. *Laterality* 20, 306–325. doi: 10.1080/1357650x.2014.963597

Posedel, J., Emery, L., Souza, B., and Fountain, C. (2012). Pitch perception, working memory, and second-language phonological production. *Psychol. Music* 40, 508–517. doi: 10.1177/0305735611415145

Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage* 62, 816–847. doi: 10.1016/j.neuroimage.2012.04.062

Putkinen, V., Tervaniemi, M., and Huotilainen, M. (2019). Musical playschool activities are linked to faster auditory development during preschool-age: a longitudinal ERP study. *Sci. Rep.* 9:11310. doi: 10.1038/s41598-019-47467-z

Qin, Z., Tremblay, A., and Zhang, J. (2019). Influence of within-category tonal information in the recognition of Mandarin-Chinese words by native and non-native listeners: an eye-tracking study. *J. Phon.* 73, 144–157. doi: 10.1016/j.wocn.2019.01.002

Qin, Z., Zhang, C., and Wang, W. S. Y. (2021). The effect of Mandarin listeners' musical and pitch aptitude on perceptual learning of Cantonese level-tones. *J. Acoust. Soc. Am.* 149, 435–446. doi: 10.1121/10.0003330

Román-Caballero, R., Arnedo, M., Triviño, M., and Lupiáñez, J. (2018). Musical practice as an enhancer of cognitive function in healthy aging - a systematic review and meta-analysis. *PLoS One* 13:e0207957. doi: 10.1371/journal.pone.0207957

Rousseau, J. J. (1781/1993). *Essai Sur L'origine des Langues*. Paris: Flammarion.

Sadakata, M., and Sekiyama, K. (2011). Enhanced perception of various linguistic features by musicians: a cross-linguistic study. *Acta Psychol.* 138, 1–10. doi: 10.1016/j.actpsy.2011.03.007

Schellenberg, E. G. (2004). Music lessons enhance IQ. *Psychol. Sci.* 15, 511–514. doi: 10.1111/j.0956-7976.2004.00711.x

Schön, D., Magne, C., and Besson, M. (2004). The music of speech: music training facilitates pitch processing in both music and language. *Psychophysiology* 41, 341–349. doi: 10.1111/1469-8986.00172.x

Schröger, E., and Wolff, C. (1996). Mismatch response of the human brain to changes in sound location. *NeuroReport* 7, 3005–3008. doi: 10.1097/00001756-199611250-00041

Shao, J., and Zhang, C. (2020). Dichotic perception of lexical tones in Cantonese-speaking congenital amusics. *Front. Psychol.* 11:1411. doi: 10.3389/fpsyg.2020.01411

Shen, G., and Froud, K. (2016). Categorical perception of lexical tones by English learners of Mandarin Chinese. *J. Acoust. Soc. Am.* 140, 4396–4403. doi: 10.1121/1.4971765

Shen, G., and Froud, K. (2019). Electrophysiological correlates of categorical perception of lexical tones by English learners of Mandarin Chinese: an ERP study. *Biling. Lang. Cogn.* 22, 253–265. doi: 10.1017/s136672891800038x

Shen, X. S., and Lin, M. (1991). A perceptual study of Mandarin tones 2 and 3. *Lang. Speech* 34, 145–156. doi: 10.1177/002383099103400202

Si, X., Zhou, W., and Hong, B. (2017). Cooperative cortical network for categorical processing of Chinese lexical tone. *Proc. Natl. Acad. Sci. U.S.A.* 114, 12303–12308. doi: 10.1073/pnas.1710752114

Sluming, V., Barrick, T., Howard, M., Cezayirli, E., Mayes, A., and Roberts, N. (2002). Voxel-based morphometry reveals increased gray matter density in Broca's area in male symphony orchestra musicians. *Neuroimage* 17, 1613–1622. doi: 10.1006/nimg.2002.1288

Spencer, H. (1857). The origin and function of music. *Fraser's Mag.* 56, 396–408.

Strait, D. L., and Kraus, N. (2014). Biological impact of auditory expertise across the life span: musicians as a model of auditory learning. *Hear. Res.* 308, 109–121. doi: 10.1016/j.heares.2013.08.004

Tang, W., Xiong, W., Zhang, Y., Dong, Q., and Nan, Y. (2016). Musical experience facilitates lexical tone processing among Mandarin speakers: behavioral and neural evidence. *Neuropsychologia* 91, 247–253. doi: 10.1016/j.neuropsychologia.2016.08.003

Tillmann, B., Burnham, D., Nguyen, S., Grimault, N., Gosselin, N., and Peretz, I. (2011). Congenital amusia (or tone-deafness) interferes with pitch processing in tone languages. *Front. Psychol.* 2:120. doi: 10.3389/fpsyg.2011.00120

Tsolaki, A., Kosmidou, V., Hadjileontiadis, L., Kompatsiaris, I. Y., and Tsolaki, M. (2015). Brain source localization of MMN. P300 and N400: Aging and gender differences. *Brain Res.* 1603, 32–49. doi: 10.1016/j.brainres.2014.10.004

Vuust, P., Brattico, E., Seppänen, M., Näätänen, R., and Tervaniemi, M. (2012). The sound of music: differentiating musicians using a fast, musical multi-feature mismatch negativity paradigm. *Neuropsychologia* 50, 1432–1443. doi: 10.1016/j.neuropsychologia.2012.02.028

Vuvan, D. T., Nunes-Silva, M., and Peretz, I. (2015). Meta-analytic evidence for the non-modularity of pitch processing in congenital amusia. *Cortex* 69, 186–200. doi: 10.1016/j.cortex.2015.05.002

Wang, W. S. Y. (1973). The Chinese language. *Sci. Am.* 228, 50–63.

Wang, W. S. Y. (1976). Language change. *Ann. N. Y. Acad. Sci.* 208, 61–72. doi: 10.1111/j.1749-6632.1976.tb25472.x

Wang, X., Ossher, L., and Reuter-Lorenz, P. A. (2015). Examining the relationship between skilled music training and attention. *Conscious. Cogn.* 36, 169–179. doi: 10.1016/j.concog.2015.06.014

Wang, Y., Jongman, A., and Sereno, J. A. (2001). Dichotic perception of Mandarin tones by Chinese and American listeners. *Brain Lang.* 78, 332–348. doi: 10.1006/brln.2001.2474

Wang, Y. X., Yang, X. H., and Liu, C. (2017). Categorical perception of Mandarin Chinese tones 1–2 and tones 1–4: effects of aging and signal duration. *J. Speech Lang. Hear. Res.* 60, 3667–3677. doi: 10.1044/2017_JSLHR-H-17-0061

Witteman, J., van Ijzendoorn, M. H., van de Velde, D., van Heuven, V. J. J. P., and Schiller, N. O. (2011). The nature of hemispheric specialization for linguistic and emotional prosodic perception: a meta-analysis of the lesion literature. *Neuropsychologia* 49, 3722–3738. doi: 10.1016/j.neuropsychologia.2011.09.028

Won, J. H., Drennan, W. R., Kang, R. S., and Rubinstein, J. T. (2010). Psychoacoustic abilities associated with music perception in cochlear implant users. *Ear Hear.* 31, 796–805. doi: 10.1097/aud.0b013e3181e8b7bd

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., and Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nat. Neurosci.* 10, 420–422. doi: 10.1038/nn1872

Wu, H., Ma, X., Zhang, L., Liu, Y., Zhang, Y., and Shu, H. (2015). Musical experience modulates categorical perception of lexical tones in native Chinese speakers. *Front. Psychol.* 6:436. doi: 10.3389/fpsyg.2015.00436

Xi, J., Zhang, L., Shu, H., Zhang, Y., and Li, P. (2010). Categorical perception of lexical tones in Chinese revealed by mismatch negativity. *Neuroscience* 170, 223–231. doi: 10.1016/j.neuroscience.2010.06.077

Xu, Y., Gandour, J. T., and Francis, A. L. (2006). Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *J. Acoust. Soc. Am.* 120, 1063–1074. doi: 10.1121/1.2213572

Yang, X., Yu, Y., Chen, L., Sun, H., Qiao, Z., Qiu, X., et al. (2016). Gender differences in pre-attentive change detection for visual but not auditory stimuli. *Clin. Neurophysiol.* 127, 431–441. doi: 10.1016/j.clinph.2015.05.013

Yang, Y., and Chen, X. (2019). Within-organ contrast in second language perception: the perception of Russian initial /r-l/ contrast by Chinese learners. *J. Acoust. Soc. Am.* 146, EL117–EL123. doi: 10.1121/1.5120549

Yang, Y., Chen, X., and Xiao, Q. (2020). Cross-linguistic similarity in L2 speech learning: evidence from the acquisition of Russian stop contrasts by Mandarin speakers. *Second Lang. Res.* doi: 10.1177/0267658319900919. [Epub ahead of print].

Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

Yu, K., Li, L., Chen, Y., Zhou, Y., Wang, R., Zhang, Y., et al. (2019). Effects of native language experience on Mandarin lexical tone processing in proficient second language learners. *Psychophysiology* 56, 1–20. doi: 10.1111/psyp.13448

Yu, K., Wang, R., Li, L., and Li, P. (2014). Processing of acoustic and phonological information of lexical tones in Mandarin Chinese revealed by mismatch negativity. *Front. Hum. Neurosci.* 8:729. doi: 10.3389/fnhum.2014.00729

Yu, K., Zhou, Y., Li, L., Su, J., Wang, R., and Li, P. (2017). The interaction between phonological information and pitch type at pre-attentive stage: an ERP study of lexical tones. *Lang. Cogn. Neurosci.* 32, 1164–1175. doi: 10.1080/23273798.2017.1310909

Zhang, C., Shao, J., and Huang, X. (2017). Deficits of congenital amusia beyond pitch: evidence from impaired categorical perception of vowels in Cantonese-speaking congenital amusics. *PLoS One* 12:e0183151. doi: 10.1371/journal.pone.0183151

Zhang, L., Xi, J., Xu, G., Shu, H., Wang, X., and Li, P. (2011). Cortical dynamics of acoustic and phonological processing in speech perception. *PLoS One* 6:e20963. doi: 10.1371/journal.pone.0020963

Zhang, Y., Kuhl, P., Imada, T., Kotani, M., and Tohkura, Y. (2005). Effects of language experience: neural commitment to language-specific auditory patterns. *Neuroimage* 26, 703–720. doi: 10.1016/j.neuroimage.2005.02.040

Zhao, T. C., and Kuhl, P. K. (2015a). Effect of musical experience on learning lexical tone categories. *J. Acoust. Soc. Am.* 137, 1452–1463. doi: 10.1121/1.4913457

Zhao, T. C., and Kuhl, P. K. (2015b). Higher-level linguistic categories dominate lower-level acoustics in lexical tone processing. *J. Acoust. Soc. Am.* 138, EL133–EL137. doi: 10.1121/1.4927632

Zheng, Y., and Samuel, A. G. (2018). The effects of ethnicity, musicianship, and tone language experience on pitch perception. *Q. J. Exp. Psychol.* 71, 2627–2642. doi: 10.1177/1747021818757435

Conflict of Interest: 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 Neuroscience

ORIGINAL RESEARCH
published: 23 July 2021
doi: 10.3389/fnins.2021.648724

Effectiveness of Melodic Intonation Therapy in Chinese Mandarin on Non-fluent Aphasia in Patients After Stroke: A Randomized Control Trial

Xiao-Ying Zhang[1,2,3,4,5*†], *Wei-Yong Yu*[1,6†], *Wen-Jia Teng*[1,5], *Meng-Yang Lu*[1,5], *Xiao-Li Wu*[1,7], *Yu-Qi Yang*[1,7], *Chen Chen*[8], *Li-Xu Liu*[1,7], *Song-Huai Liu*[1,5] *and Jian-Jun Li*[1,2,3,4*]

[1] *School of Rehabilitation Medicine, Capital Medical University, Beijing, China,* [2] *China Rehabilitation Science Institute, Beijing, China,* [3] *Beijing Key Laboratory of Neural Injury and Rehabilitation, Beijing, China,* [4] *Center of Neural Injury and Repair, Beijing Institute for Brain Disorders, Beijing, China,* [5] *Music Therapy Center, Department of Psychology, China Rehabilitation Research Center, Beijing, China,* [6] *Department of Imaging, China Rehabilitation Research Center, Beijing, China,* [7] *Department of Neurorehabilitation, China Rehabilitation Research Center, Beijing, China,* [8] *Department of Music Education, Xinghai Conservatory of Music, Guangzhou, China*

Edited by:
Mary Miu Yee Waye,
The Chinese University of Hong Kong, China

Reviewed by:
Corene Patricia Thaut,
University of Toronto, Canada
Xiaoyu Xia,
The Military General Hospital of Beijing PLA, China

***Correspondence:**
Xiao-Ying Zhang
389608661@qq.com
Jian-Jun Li
crrclijj@163.com;
lintaozixyz_1228@hotmail.com
orcid.org/0000-0002-8441-7537

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience

Received: *01 January 2021*
Accepted: *10 June 2021*
Published: *23 July 2021*

Citation:
Zhang X-Y, Yu W-Y, Teng W-J, Lu M-Y, Wu X-L, Yang Y-Q, Chen C, Liu L-X, Liu S-H and Li J-J (2021) Effectiveness of Melodic Intonation Therapy in Chinese Mandarin on Non-fluent Aphasia in Patients After Stroke: A Randomized Control Trial. Front. Neurosci. 15:648724. doi: 10.3389/fnins.2021.648724

Melodic intonation therapy (MIT) positively impacts the speech function of patients suffering from aphasia and strokes. Fixed-pitch melodies and phrases formulated in MIT provide the key to the target language to open the language pathway. This randomized controlled trial compared the effects of music therapy-based MIT and speech therapy on patients with non-fluent aphasia. The former is more effective in the recovery of language function in patients with aphasia. Forty-two participants were enrolled in the study, and 40 patients were registered. The participants were randomly assigned to two groups: the intervention group (n = 20; 16 males, 4 females; 52.90 ± 9.08 years), which received MIT, and the control group (n = 20; 15 males, 5 females; 54.05 ± 10.81 years), which received speech therapy. The intervention group received MIT treatment for 30 min/day, five times a week for 8 weeks, and the control group received identical sessions of speech therapy for 30 min/day, five times a week for 8 weeks. Each participant of the group was assessed by a Boston Diagnostic Aphasia Examination (BDAE) at the baseline (t1, before the start of the experiment), and after 8 weeks (t2, the experiment was finished). The Hamilton Anxiety Scale (HAMA) and Hamilton Depression Scale (HAMD) were also measured on the time points. The best medical care of the two groups is the same. Two-way ANOVA analysis of variance was used only for data detection. In the spontaneous speech (information), the listening comprehension (right or wrong, word recognition, and sequential order) and repetitions of the intervention group were significantly higher than the control group in terms of the cumulative effect of time and the difference between groups after 8 weeks. The intervention group has a significant time effect in fluency, but the results after 8 weeks were not significantly different from those in the control group. In terms of naming, the intervention group was much better than the control group in spontaneous naming. Regarding object naming, reaction naming, and sentence completing, the intervention group showed a strong time accumulation effect. Still, the results after 8 weeks were not significantly different from those in the control group. These results indicate that, compared with speech therapy, MIT based on music therapy is a more effective musical activity and is effective and valuable for the recovery of speech function in patients with non-fluent aphasia. As a more professional non-traumatic treatment method, MIT conducted by qualified music therapists requires deeper cooperation between doctors and music therapists to improve rehabilitating patients with aphasia. The Ethics Committee of the China Rehabilitation Research Center approved this study (Approval No. 2020-013-1 on April 1, 2020) and was registered with the Chinese Clinical Trial Registry (Registration number: Clinical Trials ChiCTR2000037871) on September 3, 2020.

Keywords: stroke, non-fluent aphasia, melodic intonation therapy, Chinese Mandarin, music therapy

INTRODUCTION

Stroke constitutes one of the leading causes of long-term disability worldwide (Benjamin et al., 2017). Of the major neurological deficits, language function disorder is the main symptoms for stroke-related impairments, which are defined as aphasia in the clinic. Aphasia is a kind of acquired loss or impairment of the ability to communicate by language following brain damage, which is usually in the left hemisphere (Wade et al., 1986). Aphasia is part of the most common complications that occurred in one-third after stroke, about 21–38% of the patients are correlated with different degrees of symptoms in stroke survivors (Dickey et al., 2010). Aphasia is often subdivided into fluent and non-fluent aphasia. Non-fluent aphasia generally results from a stroke in the left frontotemporal regions and is characterized by slow, effortful speech (Meulen et al., 2016). It mainly presents an oral expression barrier, with relatively good comprehension, and difficulty in understanding grammatical words, order words, sentences, retelling, naming, reading, and writing in varying degrees (Fazio et al., 2009). Non-fluent aphasia mainly includes Broca's aphasia, complete aphasia, and so on.

Melodic intonation therapy (MIT) is a formalized impairment-based approach of language rehabilitation that uses melodic and rhythmic elements of intoning phrases and words to assist in speech recovery in patients with Broca's aphasia (Albert et al., 1973). MIT was developed by a group of neurologic researchers in the early 1970s and, now, is a hierarchically structured treatment program identified by the American Academy of Neurology as an effective form of output-focused language therapy (Helm-Estabrooks and Albert, 1991; Assessment, 1994). The basic rationale for MIT emphasizes the use of rhythmic musical elements to engage language-capable regions of the undamaged right hemisphere (Helm-Estabrooks and Albert, 2004). Considering the dominant role of music processing in the right hemisphere, MIT uses the comprehensive characteristics of music, rhythm, and speech output, and uses the proprioceptive input of the left hand to participate in the control of the sensory motor network and oral output (Schuppert et al., 2000; Gentilucci and Dalla Volta, 2008; Norton et al., 2009). Among the musical parts in MIT, the intoned-speech technique is a musical stylization of the normal speech prosody using a few pitches, usually only two or four, separated by a third or a fourth, and a simple rhythm, quarter or eighth notes (Sparks, 2008), formulated into a short melody to represent the trained phrases on a slow tempo.

MIT was originally applied in the English-speaking patients (Norton et al., 2009). In recent years, there are several literatures that reported that non-English MIT were applied in clinical aphasia populations (Cortese et al., 2015; Tabei et al., 2016), including non-English linguistic patients such as Italian (Cortese et al., 2015), Japanese (Tabei et al., 2016), Romanian (Popovici, 1995), Persian (Bonakdarpour et al., 2003), French (Zumbansen et al., 2014), Dutch (Van der Lugt-van Wiechen and Verschoor, 1987), and Caucasian (Breier et al., 2010) with comparable clinical results. However, most of these studies were case studies, minimal sample studies, meta-analysis, and mechanism researches in chronic aphasia (Van der Meulen et al., 2012), and no large sample studies focused on East Asian languages, especially for Chinese Mandarin. The literature has shown now that MIT is more effective than normal speech therapy in different language chronic aphasia. A group study of 11 chronic non-fluent English-speaking aphasic patients examined by Wan et al. (2014) reported an improved communicative effectiveness and verbal fluency after MIT, and associated with structural changes in the white matter underlying the right inferior frontal gyrus. In Italian and French-speaking patients (Zumbansen et al., 2014; Cortese et al., 2015), the MIT group showed a significant improvement after 16 weeks and also has the same effect in spontaneous speech at the 6-months follow-up. Japanese is the best close to Chinese Mandarin in Eastern Asian language family. In Japanese MIT, Tabei et al. (2016) reported a marked improvement in following an intensive 9-day training on one Japanese MIT. Following MIT-J, the arcuate fasciculus of a part of the right hemisphere was improved by increased neural processing efficiency. Chen et al. (2020) reported that 17 patients with non-fluent aphasia in Chinese Mandarin had a significant improvement in the score of Western Aphasia Battery (WAB) scale after MIT treatment. Therefore, it can be seen that, although MIT is effective in clinical interventions in East Asian languages, especially for Mandarin Chinese (MIT-C), a large sample evidence is still needed.

It is reported that the number of new stroke cases that occurred in China was about 2.6–4.7 million in 2019 (Wang et al., 2019), ranking the first in the world (Johnson et al., 2019). Therefore, the MIT in Chinese for patients with aphasia after stroke is particularly necessary. Unlike the multisyllable pronunciation in the Western language, Chinese mandarin is monosyllabic pronunciation, and also, one syllable has four tones, each of which represents a different meaning. According to the regularity of melody and rhythm of speech, syllabic pitch is a relative-pitch system using musical notes and a series of "peculiar symbols" that would represent the relative pitch and relative duration of each spoken syllable of an utterance (Chow and Brown, 2018). In this basic terms of pronunciation rule, MIT is more suitable with short melodies in Chinese character of word and sound. Zhang et al. reported in a case study that according to the three levels of language rehabilitation, the content of Chinese MIT can be divided into three steps (Zhang et al., 2016): (1) 1–3-words sentence, (2) 4–6-word sentence, (3) 7-word sentence and 7 above (Zhang et al., 2016). This randomized controlled trial (RCT) is to observe the behavioral efficacy of existing treatment paradigm of MIT-C in clinical intervention for Chinese aphasia. We used the RCT design to compare the therapeutic effects of MIT-C and speech therapy in patients with non-fluent aphasia whose mother tongue is Chinese, and to explore the specific target curative effect of the existing MIT-C clinical operation paradigm with aphasia.

SUBJECTS AND METHODS

This study was approved by the Ethics Committee of China Rehabilitation Research Center (CRRC) (approval No. 2020-013-1) on April 1, 2020 (**Supplementary File 1**), and informed consent (**Supplementary File 2**) was obtained from the participants, relatives, or guardians before commencing the study. The study trial was registered with the Chinese Clinical Trial Registry (Registration No. ChiCTR2000037871) on September 3, 2020.

Participants

Forty participants were recruited from CRRC, Beijing. The inclusion criteria were as follows: (1) diagnosed with fMRI or CT imaging, showing left ischemic stroke or hemorrhagic stroke; (2) The ninth language score on the National Institutes of Health Stroke Scale (NIHSS) (Farooque et al., 2020) is 1—mild to moderate aphasia and 2—severe aphasia. (3) meeting the diagnostic criteria for non-fluent aphasia: less active speech expression, lack of fluency in speaking, acceptable hearing ability, can give a sign of yes/no questions, willing to express, good cooperation, and emotional stability (Wang et al., 2019); (4) Aphasia for more than 15 days after stroke, hospitalized patients; (4) aged 18–70; (5) tolerance to lying therapy for more than half an hour without postural hypotension; (6) The medication and other brain metabolism enhancers are the same; physical therapy, occupational therapy, and routine care are the same. (7) None of the participants had professional musical experience. (8) Patients and their families provided written informed consent to participate in this study. The exclusion criteria were (1) severe auditory dysfunction; (2) having epilepsy, malignant arrhythmia, or other serious physical diseases; and (3) patients with mental symptoms and obvious emotional agitation. Criteria for withdrawal and termination: patients could be terminated if their condition changed, if they were discharged from the hospital, or if they voluntarily withdrew. Forty participants completed the experiment. Two participants were withdrawn from the study because they did not meet the inclusion criteria. The data of participants' characteristics are shown in **Table 1**.

TABLE 1 | Participants' characteristics in this study.

	Intervention group	Control group	*t*	*p*
Total Number	20	20		
Gender				
Male	16	15		
Female	4	5		
Age	52.90 ± 9.08	54.05 ± 10.81	0.5089	0.5845
months since injury	2.57 ± 1.74	1.96 ± 1.38	0.2677	0.865
Stoke classification				
Left cortical ischemic	10	14		
Left cortical hemorrhagic	10	6		
Non-fluent aphasia classification				
Global aphasia	9	12		
Broca's aphasia	8	7		
Transcortical mixing	3	1		

Data were expressed as a number in total number, gender, years since injury, stoke classification and non-fluent aphasia classification. Other data were expressed as the mean ± SD, and analyzed by paired t-test. Intervention group: melodic intonation therapy group; Control group: speech therapy group. P > 0.05 indicates no significant difference between the two groups.

Study Design

The study was a randomized controlled trial with a pre-test–post-test design. It included two groups: the intervention group (n = 20) and the control group (n = 20). This study adopts a double-blind design—neither the participants nor the data analyst knows which group of data is being tested and analyzed. The intervention group received melodic intonation therapy, while the control group received speech therapy. The study was conducted from April 2020 to October 2020 at CRRC. The costs involved in this trial are all funded by the 2020CZ-10 scientific research project of the Chinese Institute of Rehabilitation Sciences (CIRS). This is a national non-profit foundation plan and has been approved by the Ministry of Finance of China.

Procedure

After obtaining approval from the Scientific Research Foundation of CIRS, participants were screened by the neurorehabilitation specialists. Patients who were diagnosed as non-fluent aphasia in the ninth language score on the NIHSS are 1—mild-to-moderate aphasia and 2—severe aphasia and were referred to the Music Therapy Department at CRRC. Participants were reviewed by the researchers to identify potential interventional objectives based on the inclusion and exclusion criteria of the study. Once potential participants were identified, an invitation inform to the study was sent to their family members. The inform included the purpose, procedures, risks, benefits, confidentiality, and participants' rights. Once we acquired the consent forms, the participants were assessed by professional evaluators for the modified Boston Diagnostic Aphasia Examination (BDAE) (Fong et al., 2019) to determine non-fluent aphasia types. The clinical researchers screened patients based on BDAE scores to confirm whether they had an abnormal speech function. After the screening, computer-generated sequences (by Excel 2013, Microsoft office software, Seattle, WA, United States)

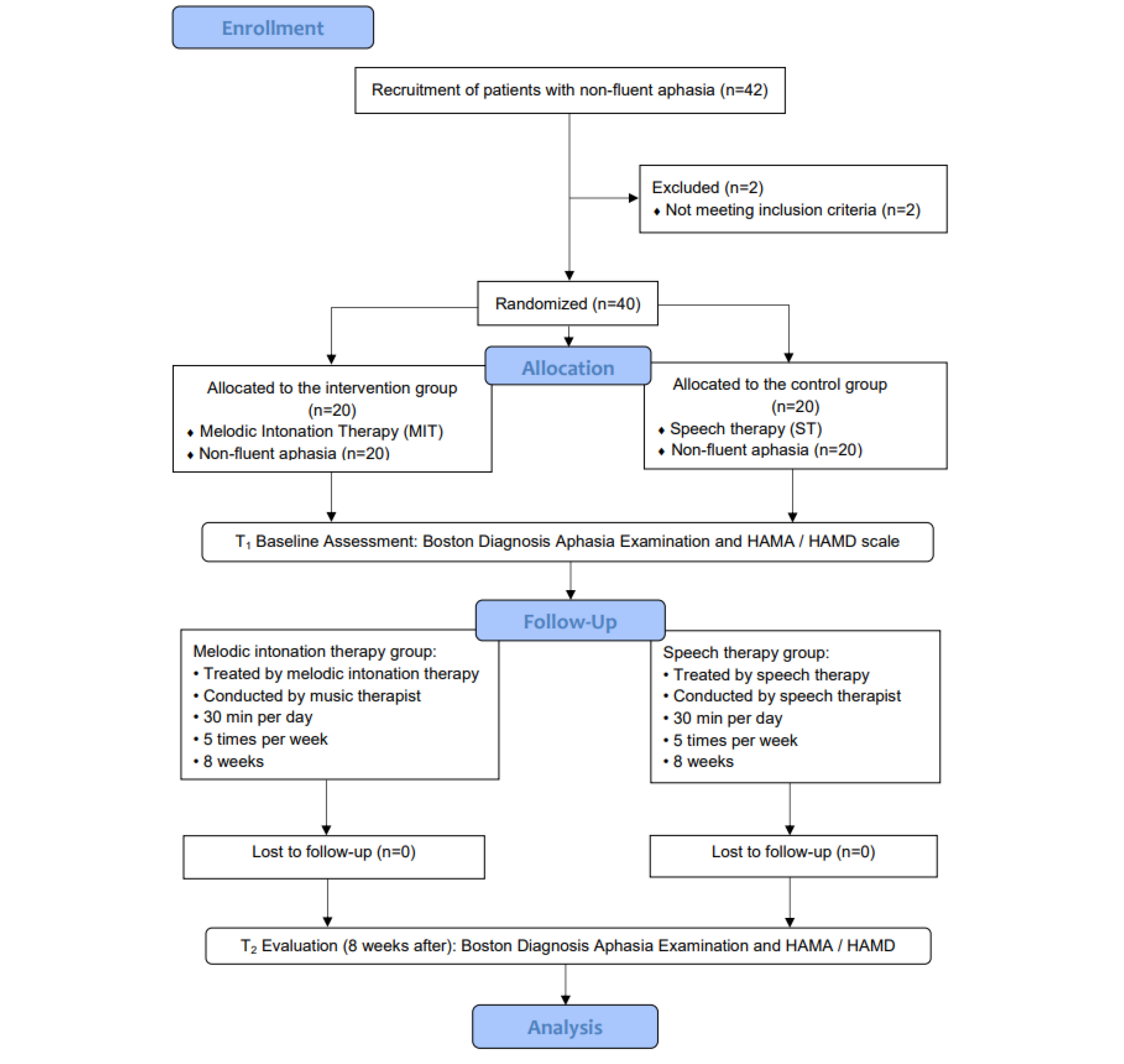

Enrollment
Recruitment of patients with non-fluent aphasia (n=42)
Excluded (n=2)
⬩ Not meeting inclusion criteria (n=2)
Randomized (n=40)
Allocation
Allocated to the intervention group (n=20)
⬩ Melodic Intonation Therapy (MIT)
⬩ Non-fluent aphasia (n=20)
Allocated to the control group (n=20)
⬩ Speech therapy (ST)
⬩ Non-fluent aphasia (n=20)
T₁ Baseline Assessment: Boston Diagnosis Aphasia Examination and HAMA / HAMD scale
Follow-Up
Melodic intonation therapy group:
• Treated by melodic intonation therapy
• Conducted by music therapist
• 30 min per day
• 5 times per week
• 8 weeks
Speech therapy group:
• Treated by speech therapy
• Conducted by speech therapist
• 30 min per day
• 5 times per week
• 8 weeks
Lost to follow-up (n=0)
Lost to follow-up (n=0)
T₂ Evaluation (8 weeks after): Boston Diagnosis Aphasia Examination and HAMA / HAMD
Analysis

FIGURE 1 | Flow diagram, consort flowchart for participants' recruitment and allocation.

were used to randomly assign the patients into the two groups. The participants in the intervention group were treated by melodic intonation therapy for 8 weeks by registered music therapists, while participants in the control group were treated with speech therapy for 8 weeks. The enrollment and allocation of participants are shown in **Figure 1**.

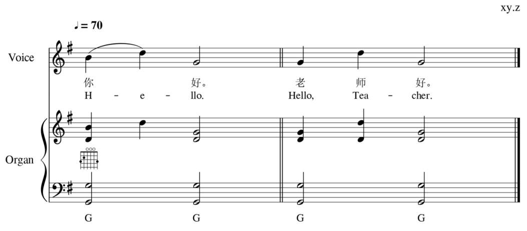
Figure 1 illustrates that 42 participants were enrolled in the study, and 2 participants withdrew from the study because of not meeting the inclusion criteria ($n = 2$). The intervention group was treated with melodic intonation therapy ($n = 20$) by a music therapist, while the control group was treated with speech therapy ($n = 20$) by speech therapists. Two times evaluation was conducted during the whole period: t1 (baseline) and t2 (after 8 weeks). The data analysis included a sample of 40 non-fluent aphasia patients.

Interventions

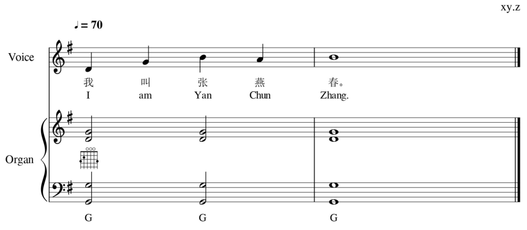
Once the participants for each group were identified, an intervention delivery schedule was developed. All patients in the intervention group received MIT training. Each patient was trained for 30 min per session, five sessions a week, for 8 weeks. The training process is carried out by music therapy professionals who have been trained in neurological music therapy (NMT) and have obtained a registered music therapist license to ensure the music professionalism of the intervention. The intervention steps of MIT strictly follow the operational steps of Chinese Mandarin MIT (Zhang et al., 2016).

According to the different three levels of speech rehabilitation, the music therapist trained the aphasia patients to intone and chant the targeted speech items, like "*sprenchsang*," (Helm-Estabrooks et al., 2013) and then fade slowly with tapping to let the patients speak out the targeted sentences in the first level. The music therapist leads the patients to sing and speak out in the same way in the second and third levels; the only difference is the length of the melodic target language (the second level is 5–9-word sentences, and the third level is 10-word sentences and above). All the melodic phrases are noted according to the natural phonic pitches of targeted Mandarin sentences; the specific 12-item implementation contents of MIT are shown in **Figure 2**. The music therapist uses a keyboard or guitar to accompany while they are singing the melody with the patients. According to the different types of damage in non-fluent aphasia, in the intervention group and the control group of this study (**Table 1**), there are 21 patients with global aphasia (*n* = 9, *n* = 12), 15 patients with Broca's aphasia (*n* = 8, *n* = 7), and 4 patients with transcortical mixed aphasia (*n* = 3, *n* = 1). According to the MIT training content in the **Supplementary Figure 2**, within 8 weeks, the training objectives for patients with global aphasia in the interventional group are the first and second levels (**Supplementary Figure 2**, items 1–6), the training objectives of the patients with Broca's aphasia in the interventional group are the second and third levels (**Supplementary Figures 2**, **1–10** items), and the patients with mixed transcortical aphasia in the interventional group are the first to third levels (**Supplementary Figure 2**, items 1–9). The effective behavioral performance of the intervention is that, when the therapist asks the target question, the patient can speak the target language at a natural speed without the melody and rhythm, and the behavior performance can last for more than 3 weeks without regression. The therapy sessions of the two groups were both 30 min per day, five times a week, for a total of eight consecutive weeks. All patients underwent routine treatment during the study period, including taking medication and other care and support.

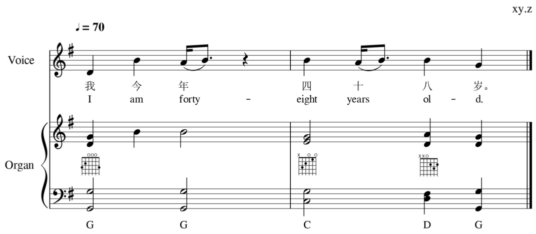
① Greetings



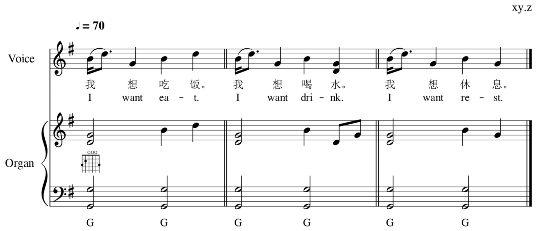
② Name



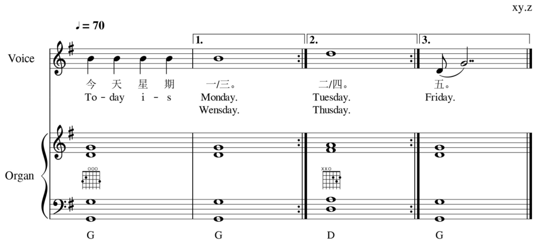
③ Age



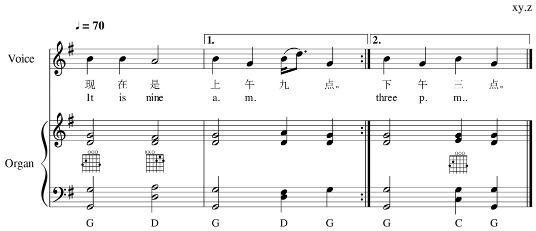
④ ADL Daily Behaviors


xy.z
♩= 70
Voice
我 想 吃 饭。我 想 喝 水。我 想 休 息。
I want ea - t. I want dri - nk. I want re - st.
Organ
G G G G G G

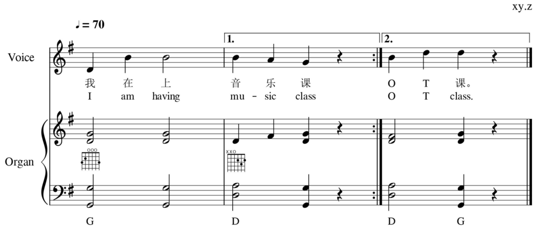
⑤ Weeks


xy.z
♩= 70
1. 2. 3.
Voice
今 天 星 期 一/三。二/四。五。
To - day i - s Monday. Tuesday. Friday.
Wensday. Thusday.
Organ
G G D G

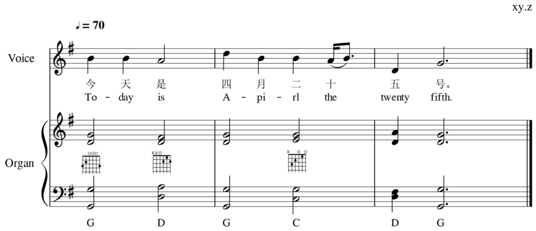
⑥ Times


xy.z
♩= 70
1. 2.
Voice
现 在 是 上 午 九 点。下 午 三 点。
It is nine a. m. three p. m..
Organ
G D G D G G C G

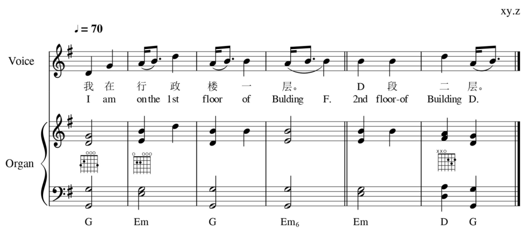
⑦ Behaviors in the calendar


xy.z
♩= 70
1. 2.
Voice
我 在 上 音 乐 课 O T 课。
I am having mu - sic class O T class.
Organ
G D D G

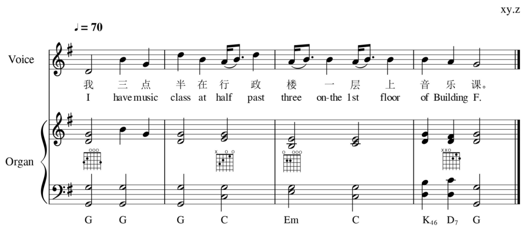
⑧ Dates


xy.z
♩= 70
Voice
今 天 是 四 月 二 十 五 号。
To - day is A - pi - rl the twenty fifth.
Organ
G D G C D G

⑨ Locations


xy.z
♩= 70
Voice
我 在 行 政 楼 一 层。 D 段 二 层。
I am onthe 1st floor of Bulding F. 2nd floor-of Building D.
Organ
G Em G Em6 Em D G

⑩ Long sentences


xy.z
♩= 70
Voice
我 三 点 半 在 行 政 楼 一 层 上 音 乐 课。
I have music class at half past three on-the 1st floor of Building F.
Organ
G G G C Em C K46 D7 G

Measurements

This is to assess the baseline of all the participants in BDAE within 30–40 min before all the therapy started. Then the participants were randomly assigned to the intervention group and the control group. The final evaluation was measured after 8 weeks. The test sessions consisted of the (1) BDAE (Fong et al., 2019) and (2) Hamilton Anxiety Scale (HAMA)/Hamilton Depression Scale (HAMD) (Spreen and Risser, 2003). During the evaluation process, no participants wore a 24-h Holter test or a 24-h ambulatory blood pressure test; no participants had unsealed tracheostomy and had difficulty expecting sputum.

Boston Diagnostic Aphasia Examination

BDAE is a measure used in the neuropsychological assessment of aphasia and is currently in its third edition (Fong et al., 2019). It evaluates language skills in aphasia based on perceptual modalities (auditory, visual, and gestural), processing functions (comprehension, analysis, and problem solving), and response modalities (writing, articulation, and manipulation). Administration time ranges from 35 to 45 min. Other tests are sometimes used by neurologists and speech language pathologists on a case-by-case basis, but BDAE is a universal language assessment scale that has been proven, with high reliability and validity, and can be applied to multiple languages. BDAE is a comprehensive, multifactorial battery designed to evaluate a broad range of language impairments that often arise as a consequence of organic brain dysfunction. The examination is designed to go beyond simple functional definitions of aphasia into the components of language dysfunctions (symptoms) that have been shown to underlie the various aphasic syndromes.

BDAE includes four dimensions: spontaneous speech, repetition, listening comprehension, and naming. Among them, spontaneous speech includes information and fluency assessment; listening comprehension includes three assessments of right and wrong questions, word recognition, and sequential instruction; naming includes four items: object naming, spontaneous speech, sentence supplement, and response naming. Finally, the above four subtables are calculated into the total score formula to obtain the aphasia quotient (AQ). Thus, this approach allows for measurement of language-related skills and abilities and neuropsychological analysis from both ideographic and nomothetic bases, as well as a comprehensive approach to the symptom configurations that relate to neuropathologic conditions (Spreen and Risser, 2003).

Hamilton Anxiety Scale and Hamilton Depression Scale

The HAMA is a scale commonly used in the clinic to assess the anxiety of patients. The HAMD is the most commonly used scale for clinical evaluation of depression. The HAMD used in this study is a 17-item score battery. Both HAMA and HAMD use a five-level scoring method of 0–4 points. The scores range from asymptomatic to extremely severe. Both of the evaluations are concise and efficient. In this study, HAMA and HAMD have clinical reference values for the positive psychological effect of patients with non-fluent aphasia before and after therapy.

Among those assessments, the BDAE tests and the HAMA/HAMD questionnaires are evaluated by experienced professionals. All the evaluators were registered research assistants who worked as health care professionals with 5 years of clinical experience. The evaluated results were on the consistency test analysis with open-label design.

Statistical Analysis

The measure data of the two groups were collected at two time points before intervention (t_1) and 8 weeks later (t_2). Taking the mean of each group and the standard deviation of the normal distribution, repeated measures of variance (two-way ANOVA) were used to observe intergroup differences, time effects, and intergroup time interaction differences. SPSS statistical software, Version 22.0 (IBM Lenovo, BJ) was used for statistical analysis. The data of 40 patients with non-fluent aphasia who completed this study were analyzed by SPSS 22.0. All the data of the intervention group and the control group were collected before (t1) and after the intervention (t2). Before analysis, basic frequencies were run on the data to screen for missing values and outliers and to establish data entry accuracy. Data were analyzed using a repeated-measures ANOVA to determine the specific effects of the interventions.

RESULTS

Effectiveness of the Boston Diagnostic Aphasia Examination Test Results in Patients With Non-fluent Aphasia: The Part of Spontaneous Speech, Repetition, and Listening Comprehension

The first part of the BDAE scale, spontaneous speech, repetition, and listening comprehension, was tested in both groups before the first session and after the last session. Two-way ANOVA was used to analyze the results of the intervention group and the control group at t1 and t2. Individual results were normalized by ruling out the difference of greater dispersion.

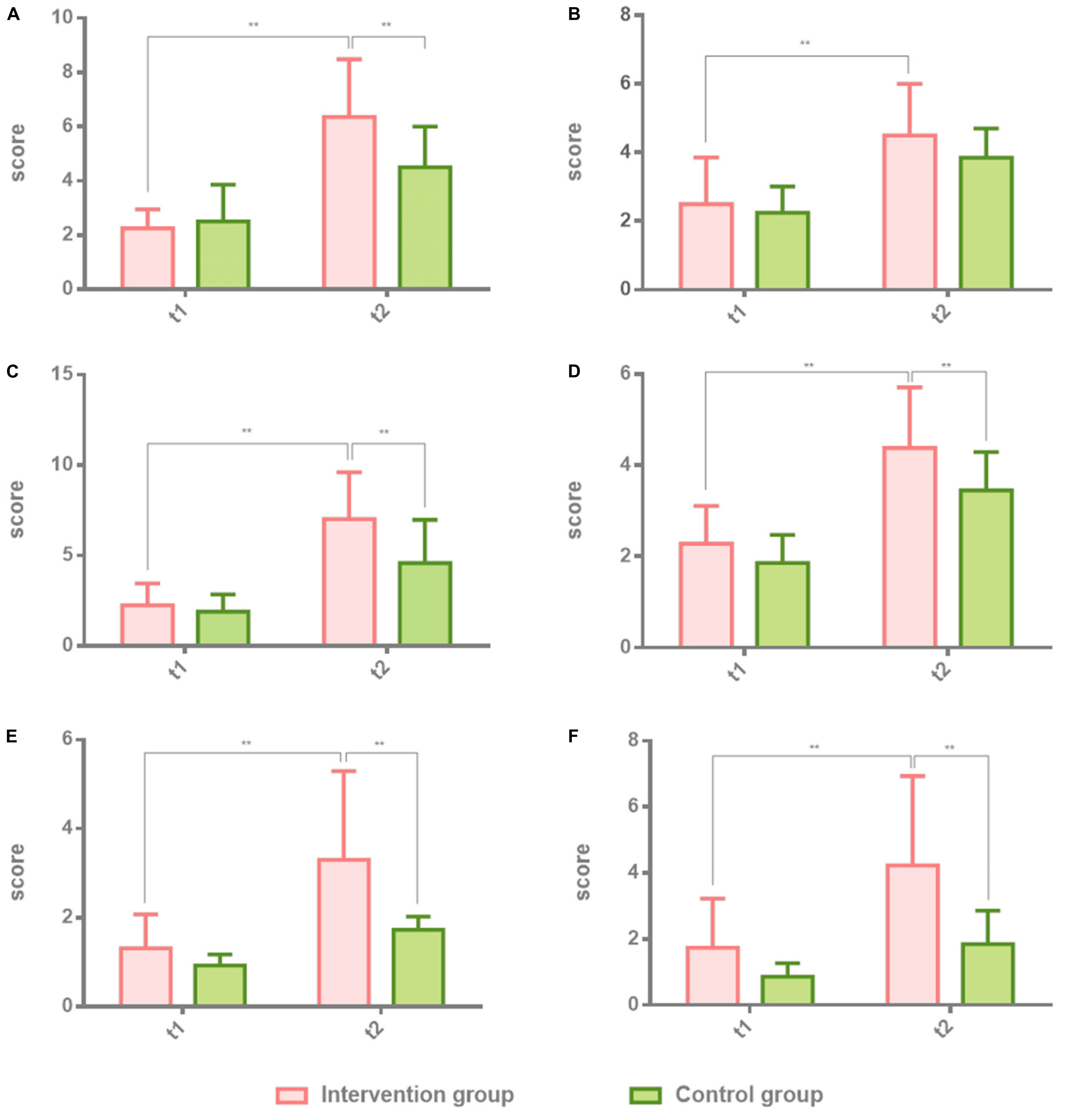


FIGURE 2 | Comparison of spontaneous speech, repetition, and listening comprehension in patients with non-fluent aphasia with the intervention group and the control group. The intervention group: melodic intonation therapy group; the control group: speech therapy group. **(A–F)** Information, fluency, repetition, true or false, word recognition, and sequential commands. Data are expressed as mean ± SD (n = 20) and analyzed by repeated-measures analysis of variance. $^{*}p < 0.05$, $^{**}p < 0.01$. t_1, baseline; t_2, after 8 weeks.

The effect of the intervention group is higher than the control group. There were significant differences in the intervention group at t2 (8 weeks after) on information (t_2 = 6.35 ± 2.13, t = 0.5775, p = 0.0002), fluency (t_2 = 4.50 ± 1.50, t = 3.975, p = 0.0019), which belongs to spontaneous speech, and repetition (t_2 = 7.01 ± 2.61, t = 3.975, p = 0.0019) in comparison with the control group. There were also significant differences in the intervention group at t2 (8 weeks after) on true or false (t_2 = 4.38 ± 1.33, t = 3.134, p = 0.0019), word recognition (t_2 = 3.30 ± 2.00, t = 0.13, p = 0.0001), and sequential commands (t_2 = 4.23 ± 2.70, t = 4.591, p = 0.0001) in comparison with the control group. **Table 2** shows the results

TABLE 2 | The results of Boston Diagnosis Aphasia Examination (BDAE) in patients with non-fluent aphasia across the study period for the intervention group and the control group.

			Intervention group (*n* = 20)	Control group (*n* = 20)	*t*	*p*
			Mean ± SD	Mean ± SD		
Spontaneous speech	Information	t_1	2.25 ± 0.70	4.40 ± 0.86	4.967	0.0001** a
		t_2	6.35 ± 2.13	6.60 ± 1.32	0.5775	0.0002** b
	Fluency	t_1	2.50 ± 1.36	2.25 ± 0.77	0.6795	0.0878
		t2	4.50 ± 1.50	3.85 ± 0.85	1.767	0.0001** b
Repetition		t1	2.25 ± 1.22	1.90 ± 0.95	0.5655	0.0001** a
		t_2	7.01 ± 2.61	4.59 ± 2.39	3.975	0.0019** b
Listening comprehension	True or False	t_1	2.28 ± 0.83	1.86 ± 0.61	1.415	0.0001** a
		t_2	4.38 ± 1.33	3.45 ± 0.84	3.134	0.0019** b
	Words recognition	t_1	1.31 ± 0.77	0.93 ± 0.24	1.106	0.0001** a
		t_2	3.30 ± 2.00	1.73 ± 0.30	4.583	0.0001** b
	Sequential commands	t_1	1.74 ± 1.49	0.85 ± 0.41	1.688	0.0001** a
		t_2	4.23 ± 2.70	1.85 ± 1.02	4.591	0.0001** b
Naming	Objective naming	t_1	0.66 ± 0.36	1.27 ± 0.48	1.89	0.3064
		t_2	2.36 ± 1.72	2.22 ± 0.92	0.4337	0.0001** b
	Spontaneous naming	t_1	0.16 ± 0.11	0.12 ± 0.75	0.4707	0.0042** a
		t_2	0.50 ± 0.44	0.22 ± 0.11	3.698	0.0001** b
	Sentences completing	t_1	0.16 ± 0.12	0.19 ± 0.05	0.4052	0.8865
		t_2	0.50 ± 0.43	0.46 ± 0.14	0.6079	0.0001** b
	Reaction naming	t_1	0.14 ± 0.11	0.15 ± 0.05	0.0598	0.087
		t_2	0.62 ± 0.50	0.41 ± 0.12	2.512	0.0001** b
Aphasia Quation (AQ)		t_1	26.87 ± 9.65	27.85 ± 4.02	0.236	0.0088** a
		t_2	67.47 ± 22.99	50.71 ± 7.22	4.036	0.0001** b

*Intervention group: melodic intonation therapy group; Control group: speech therapy group. Data were expressed as mean ± SD (n = 20), and analyzed by repeated measures analysis of variance. **P < 0.01, remarkable significance; *p < 0.05, significance. Superscript a represents difference factor at the same time between groups, and superscript b represents difference effect of time factor in inter-group.*

of the two groups. **Figure 2** shows the comparison results of the two groups.

Effectiveness of the Boston Diagnostic Aphasia Examination in Patients With Non-fluent Aphasia: Naming and Aphasia Quotient

The second part of the BDAE scale, naming and aphasia quotient (AQ), was tested in both groups before the first and after the last session. Two-way ANOVA was used to analyze the results of the intervention group and the control group at t1 and t2. Individual results were normalized by ruling out the difference of greater dispersion. The effect of the intervention group is higher than the control group. There were significant differences in the intervention group at t2 (8 weeks after) on objective naming (t_2 = 2.36 ± 1.72, t = 0.4337, p = 0.0001), spontaneous naming (t_2 = 0.50 ± 0.44, t = 3.698, p = 0.0001), sentence completing (t_2 = 0.50 ± 0.43, t = 0.6079, p = 0.0001), reaction naming (t_2 = 0.62 ± 0.50, t = 2.512, p = 0.0001), which belongs to naming. There were also significant differences in the intervention group at t2 (8 weeks after) on the AQ (t_2 = 67.47 ± 22.99, t = 4.036, p = 0.0001) in comparison with the control group. **Table 3** shows the results of the two groups. **Figure 3** shows the comparison results of the two groups.

Effectiveness of the Hamilton Anxiety Scale and Hamilton Depression Scale in Patients With Non-fluent Aphasia

A main effect of time was found for the HAMA and the HAMD. The effect of the intervention group is higher than the control group. There was a significant difference at t2 (8 weeks after) on HAMD (t_2 = 8.95 ± 1.97, F = 5.63, p = 0.0202) in the intervention group in comparison with the control group. There was no significant difference at t2 (8 weeks after) on HAMA (t_2 = 8.6 ± 2.68, F = 2.054, p = 0.1559) in the intervention group in comparison with the control group. A significant difference

TABLE 3 | HAMA and HAMD questionnaire results.

		Intervention group (*n* = 20)		Control group (*n* = 20)		F	p
		Mean	*SD*	Mean	*SD*		
HAMA	t_1	12.15	3.16	12.75	2.47	0.1528	0.697
	t_2	8.6	2.68	9.65	1.8	2.054	0.1559
HAMD	t_1	16.15	2.52	16.45	2.25	3.028	0.0859
	t_2	8.95	1.97	10.9	1.64	5.63	0.0202*

**p < 0.05 significant differences; (a) difference factor at the same time between groups. (b) difference effect of time factor in inter-group.*

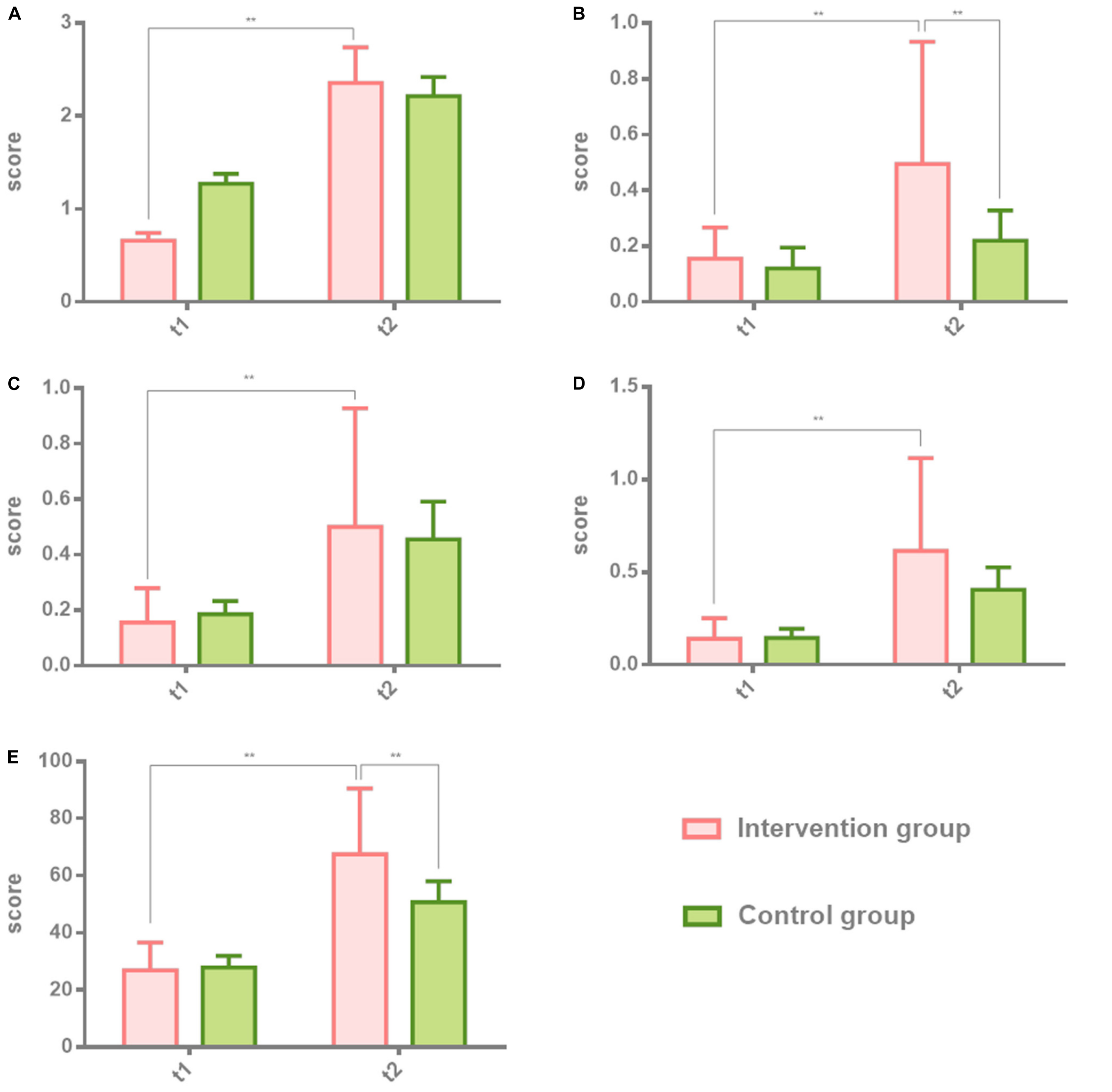


FIGURE 3 | Comparison of naming and aphasia quotient (AQ) in patients with non-fluent aphasia with the intervention group and the control group. The intervention group: melodic intonation therapy group; the control group: speech therapy group. **(A–E)** Objective naming, spontaneous naming, sentence completing, reaction naming, and AQ. Data are expressed as mean ± SD (n = 20) and analyzed by repeated-measures analysis of variance. *p < 0.05, **p < 0.01. t_1, baseline; t_2, after 8 weeks.

was observed on HAMD at the t2 time point. **Supplementary Table** shows the results of the two groups. **Figure 4** shows the comparison results of the two groups.

DISCUSSION

In this study, melodic intonation therapy and speech therapy were used as the therapeutic method in patients with aphasia. Compared with the previously reported method that used traditional speech therapy (Wan et al., 2014) as the therapeutic way to treat the patients with aphasia, this study was based on a long-time clinical work with aphasia patients that found that melodic intonation therapy has a positive effect in the clinic. The researcher of this study and the professional music therapists tried to use melodic-inducted speech output as the melodic auditory stimulation to intervene in the patients with aphasia

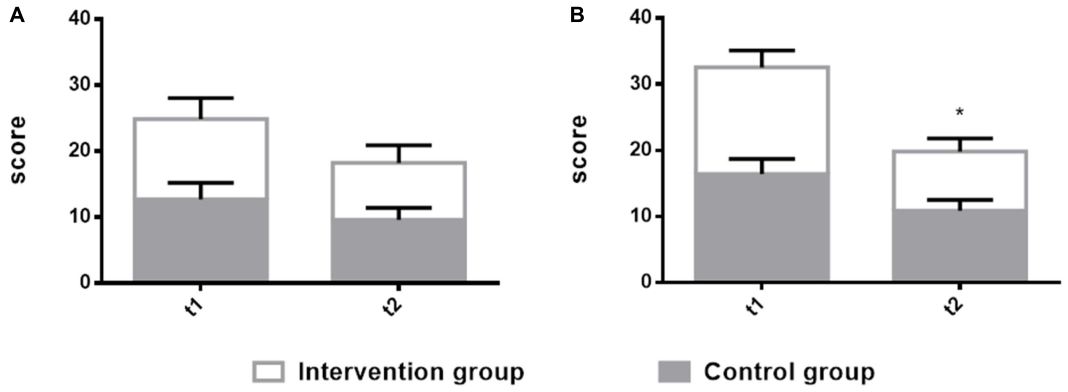


FIGURE 4 | Comparison of the Hamilton Anxiety Scale (HAMA) and the Hamilton Depression Scale (HAMD) in patients with non-fluent aphasia between the two groups with melodic intonation therapy and speech therapy. The intervention group: melodic intonation therapy group; the control group: speech therapy group. **(A)** HAMA; **(B)** HAMD. Data are expressed as mean ± SD (n = 20) and analyzed by repeated-measures analysis of variance. *p < 0.05, **p < 0.01. t_1, baseline; t_2, after 8 weeks.

and obtained unexpected results in comparison with the single-speech therapy group.

Spontaneous Speech, Repetition, and Listening Comprehension

Speech therapy is a common training method for speech disorders in patients with aphasia (Albert et al., 1973). However, due to the limitations of the method and a single-training mode, it is actually difficult for patients to adhere to or results are slow. Due to the different mechanisms, MIT uses "singing" to guide and uses the method of chanting a melody formula language to guide patients with aphasia from "singing" to speak. It is generally believed that the right hemisphere is better at processing short melody information (Sihvonen et al., 2017). Therefore, for patients with stroke on the left, the right hemisphere plays a compensatory role in oral output when "singing." MIT is, therefore, effective.

In BDAE, spontaneous speech, repetition, and listening comprehension are the test criteria for speech and listening ability. Spontaneous speech includes information, fluency, grammar, and paraphasia. Listening comprehension includes true or false questions, word recognition, and successive instructions. Repetition is a separate item, which includes 10 sentences of different lengths, includes words, short sentences, and long sentences; in the examination of this item, the patient only needs to imitate. The three sub-examinations of the BDAE scores were used to assess language comprehension and imitation in aphasia patients with visual and auditory cues. In the imitation of listening comprehension and speaking, MIT has a clearer immediate effect than speech therapy because it uses the mechanism of singing of music. The patient sings the target language while imitating the tone, which is more direct and effective than single vocabulary auditory stimulation. Therefore, in this study, the intervention group undertaking MIT had significant time cumulative effects and intergroup effects in terms of spontaneous language information, oral expression, imitation repetition, true or false judgment, hearing word recognition, and instruction execution than the control group undertaking speech therapy alone. In terms of improvement of spontaneous speech fluency, patients in the MIT group showed higher speech fluency after 8 weeks of cumulative treatment. However, compared with the control group, speech therapy also had similar effects on spontaneous speech fluency. It can be seen that the improvement of fluency is more due to the accumulation of effective treatment time. In other words, among the hospitalized non-fluent aphasia patients who received the same medical care, the music therapy–MIT group has better speech recovery effects in listening comprehension, repetition, and spontaneous speech than patients who received speech therapy and has a similar effect in terms of improving fluency.

Naming and Aphasia Quotient

The BDAE's naming test on abstract thinking is divided into four dimensions: objective naming, spontaneous naming, sentence completing, and reaction naming. Patients in the intervention group receiving MIT performed more prominently in spontaneous naming. After 8 weeks of intervention, they showed a significant cumulative effect of time. Compared with the control group, the intervention group showed a larger difference between groups. This is closely related to the repeated use of "singing" for intervention in the MIT treatment. Singing is a whole-brain activity; when the patient participates in singing, the cognitive information network related to the song will be activated (Merrett et al., 2014). Because the song contains more information, the listening experience of singing with an accompanying instrument is more complicated, not only for the language in the trained items but also for the language in other untrained items (Meulen et al., 2016). MIT can also activate more spontaneous naming responses. In terms of object naming, sentence completion, and reaction naming, the MIT group was also significantly different from the speech therapy group, but there was no obvious contrast effect in the cumulative effect after 8 weeks. Therefore, compared with MIT, speech therapy has consistent efficacy in non-spontaneous naming.

In conclusion, in the aphasia quotient performance of the two groups of patients through the BDAE test, compared with the control group using speech therapy, the overall score of the intervention group has an obvious improvement either in the time effect after 8 weeks or in the comparison between the two groups.

Hamilton Anxiety Scale and Hamilton Depression Scale

In this study, the score of the HAMD of the intervention group was lower than that of the control group, and there is no significant difference in the score of HAMA, which means a more positive subjective experience of the music group and has a decrease feeling in depression. Patients in the intervention group reported decreasing feelings in expressive difficulty, and the falling ratings are accompanied by improved speech functionality. In the control group, ratings also declined, but it did not show significant differences compared with the intervention group. In the HAMD measurements, both of the groups reported a decrease in depressive symptoms, but patients' score of the intervention group was lower than that of the control group. Each patient in the intervention group underwent an individual music therapy session that promoted interaction and positive experience. Singing interventions may reduce the depressive stress of the patients. Through familiar songs, singing, and accompaniment with music therapists, patients' sense of satisfied and happy feelings will enhance during a singing session. Therefore, music therapy may have a positive effect on aphasia patients. Having an enjoyable musical training course not only motivates them to increase participation but also brings important emotional experience. The family and guardian of aphasia patients from the intervention group reported that the patients felt reconnected with life, either by singing more, or by exploring and listening to familiar songs, and had a greater sense of participation in music activities in life, as well as the link between music, health, and quality of life. This training method, based on music and singing, has a higher participation rate, is easier, simpler, and more effective for patients to comply.

The Core of Melodic Intonation Therapy Intervention

Melodic guidance at MIT can be divided into two parts: in the first part, melody guides the language, the second, the musical language stimulation. In the process of melody in guiding the target language, the pitch of the melody comes from the natural pronunciation of Mandarin Chinese. For example, the three-tone "you" in Mandarin can be imitated by the interval of "sol–dol" (G-C). The patient began to sing and slowly generalized into speaking. These fixed-pitch formulaic melody languages range from 2 to 5 short sentences to 7–10 long sentences (**Supplementary Materials**), that is, to create lyrics of daily life language with a fixed melody, teach patients to sing, and then slowly get out of the melody, and the pitch becomes speaking. This is an entirely different approach from speech therapy intervention. According to the concept of "*sprechsang*," first proposed by Sparks scholars in 1974 (Albert et al., 1973), the melody in MIT is between "singing" and "speaking." In this study, the core intervention technology of MIT complies with the core principle of "*sprechsang*," but the innovation lies in its application in Mandarin Chinese. The formation of pitch melody completely simulates the laws of Chinese phonetics.

In the second part, after MIT, if the patient can imitate the pitch but cannot imitate the Chinese character sound, they will use MUSTIM to sing familiar songs to guide the words that cannot be expressed verbally. The choice of the song is not blind. When the music therapist chooses the song, the lyrics will include the vocabulary of the target language. For example, when the patient is guided to say "drink water," and the patient cannot complete it, the therapist will lead the patient to sing a song with the sound of "he" and "shui," such as "Wanquan River Clear and Clear" (with lyrics "River water"-"drink water") or "Love the country and the beauty more" (including the lyrics "drink the same water"). After the familiar cognitive melody is guided, the patient is guided back to the model singing of the formulaic melody, so that the patient can imitate and say it. This is why aphasia patients in the MIT group performed particularly prominently in the repetition items.

When the patient sings each short melody or song, it is accompanied by a guitar or piano and other harmonic instruments. Music therapists use musical instruments with human voices to guide patients to chant, which enriches patients' auditory experience in auditory input. Due to the interaction of the left and right hemisphere networks when the brain processes language information, when the auditory center receives multiple stimulations, they jointly activate the output of emotion, memory, and spoken language. Therefore, in this study, the MIT group in the BDAE score has improved listening comprehension, repetition, spontaneous speech, and naming.

Limitations

One limitation was the limited sample, as previously detailed. Two participants dropped out of the study, which may have caused the variance in group allocation. If a blank, the control group was added to observe self-healing, and the comparison might have been more accurate. This study only recruited 40 patients. If larger-sized studies are conducted in the future, the therapeutic outcomes could be more precisely observed. Besides, the participants with three different types of non-fluent aphasia are included into the trial. Although they belong to non-fluent aphasia, they also belong to different subtypes. If more samples can be included in future studies and different subtypes of aphasia are classified and compared, the effect comparison of the two methods will be clearer.

Implications for Clinical Practice

In previous reports in the literature, clinicians usually recommend speech therapy to train the patients with aphasia but neglected that the function of song singing played an important role in speech output. Although MIT has been proposed and used in the 1970s, it is often used by speech therapists. Given that professionals with a musical background, that is, music therapists, will have a more professional understanding of music or songs, and the operability of the musical instrument, the MIT performed by the music therapist will provide multiple auditory stimulation to the patients to activate more potential brain networks and better restore language ability. Through this study, we confirmed the positive effect of the MIT performed by a music therapist in 20 aphasia patients. All the participants in the intervention group were more active in every aspect of AQ than the control group, which provided a more effective way for speech recovery of aphasic patients. Music therapists with professional backgrounds provide multiple auditory stimuli with instrumental accompaniment and fixed-pitch melody formulaic language during the treatment process, which are all necessary conditions for the implementation of MIT.

CONCLUSION

The MIT performed by music therapists has a more obvious effect on improving the language function of patients with non-fluency aphasia. Therefore, it is recommended that clinicians and professional music therapists work together to make the clinical treatment effect more remarkable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

This study was approved by the Ethics Committee of China Rehabilitation Research Center (Approval No. 2020-013-1 in April 1, 2020), and was registered with the Chinese Clinical Trial Registry (Registration number: Clinical Trials ChiCTR2000037871) on September 3th, 2020. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

W-JT and M-YL supported music therapy—MIT treatment in the intervention group. W-YY supported the assessment. X-LW, Y-QY, and L-XL supported the patients' allocation. J-JL was the corresponding author. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the China Rehabilitation Research Center (CRRC). This research received a grant from the Scientific Research Project 2020CZ-10 of the Chinese Institute of Rehabilitation Science. It is a national non-profit foundation program and was approved by the Ministry of Finance of China. The funding sources had no role in the study conception and design, data analysis or interpretation, manuscript writing, or deciding to submit this manuscript for publication.

ACKNOWLEDGMENTS

We thank our colleagues from the Music Therapy Center, Department of Psychology, CRRC; Department of Neurorehabilitation, CRRC; Chinese Institute of Rehabilitation Science; and Center of Neural Injury and Repair and Beijing Key Laboratory of Neural Injury and Rehabilitation, CRRC who provided technical support had greatly assisted in the research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.648724/full#supplementary-material

Supplementary File 1 | Aphasia Participants' characteristics.

Supplementary File 2 | Melodic intonation therapy musical pieces.

Supplementary File 3 | Hospital Ethics Documentation.

REFERENCES

Albert, M. L., Sparks, R. W., and Helm, N. A. (1973). Melodic intonation therapy for aphasia. *Arch. Neurol.* 29, 130–131. doi: 10.1001/archneur.1973.00490260074018

Assessment, A. A. N. (1994). melodic intonation therapy. *Neurology* 44, 566–568.

Benjamin, E. J., Blaha, M. J., Chiuve, S. E., Cushman, M., Das, S. R., Deo, R., et al. (2017). Heart disease and stroke statistics—2017 update: a report from the American Heart Association. *Circulation* 135, e146–e603.

Bonakdarpour, B., Eftekharzadeh, A., and Ashayeri, H. (2003). Melodic intonation therapy in Persian aphasic patients. *Aphasiology* 17, 75–95. doi: 10.1080/729254891

Breier, J., Randle, S., Maher, L., and Papanicolaou, A. (2010). Changes in maps of language activity activation following Melodic Intonation Therapy using magnetoencephalography: two case studies. *J. Clin. Exp. Neuropsychol.* 32, 309–314. doi: 10.1080/13803390903029293

Chen, C., Xie, Y., Wu, C. W., Gui, P. J., Wu, H., and Zhang, B. (2020). Influence of melodic intonation therapy on speech apraxia in patients with early non-fluent aphasia after stroke. *J. Clin. Interv. Med.* 05, 502–505. doi: 10.3969/j.Issn.1671-4695

Chow, I., and Brown, S. (2018). A musical approach to speech melody. *Front. Psychol.* 3:247. doi: 10.3389/fpsyg.2018.00247

Cortese, M., Riganello, C., Arcuri, F., Pignataro, L. M., and Buglione, I. (2015). Rehabilitation of aphasia: application of melodic-rhythmic therapy to Italian language. *Front. Hum. Neurosci.* 9:520. doi: 10.3389/fnhum.2015.00520

Dickey, L., Kagan, A., Lindsay, M. P., Fang, J., Rowland, A., and Black, S. (2010). Incidence and profile of inpatient stroke-induced aphasia in Ontario, Canada. *Arch. Phys. Med. Rehabil.* 91, 196–202. doi: 10.1016/j.apmr.2009.09.020

Farooque, U., Lohano, A., Kumar, A., Karimi, S., Yasmin, F., Bollampally, V. C., et al. (2020). Validity of national institutes of health stroke scale for severity of stroke to predict mortality among patients presenting with symptoms of stroke. *Cureus* 12:e10255. doi: 10.7759/cureus.10255

Fazio, P., Cantagallo, A., and Craighero, L. (2009). Encoding of human action in Broca's area. *Brain* 132, 1980–1988. doi: 10.1093/brain/awp118

Fong, M. W. M., Van Patten, R., and Fucetola, R. P. (2019). The Factor Structure of the Boston Diagnostic Aphasia Examination, Third Edition. *J. Int. Neuropsychol. Soc.* 25, 772–776. doi: 10.1017/s1355617719000237

Gentilucci, M., and Dalla Volta, R. (2008). Spoken language and arm gestures are controlled by the same motor control system. *Q. J. Exp. Psychol.* 61, 944–957. doi: 10.1080/17470210701625683

Helm-Estabrooks, N., Albert, M., and Nicholas, M. (2013). *Manual of Aphasia and Aphasia Therapy*. Austin: Pro-Ed.

Helm-Estabrooks, N., and Albert, M. L. (1991). *Manual of Aphasia Therapy*. Austin: Pro-Ed.

Helm-Estabrooks, N., and Albert, M. L. (2004). *Manual of Aphasia and Aphasia Therapy*. Austin: PRO-ED Publishers.

Johnson, C. O., Nguyen, M., Roth, G. A., Nichols, E., Alam, T., and Abate, D. (2019). Global, regional, and national burden of stroke, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* 18, 439–458. doi: 10.1016/S1474-4422(19)30034-1

Merrett, D. L., Peretz, I., and Wilson, S. J. (2014). Neurobiological, Cognitive, and Emotional Mechanisms in Melodic Intonation Therapy. *Front. Hum. Neurosci.* 8:401. doi: 10.3389/fnhum.2014.00401

Meulen, I., Sandt-Koenderman, M., and Heijenbrok, M. (2016). Melodic Intonation Therapy in Chronic Aphasia_ Evidence from a Pilot Randomized Controlled Trial. *Front. Hum. Neurosci.* 10:533. doi: 10.3389/fnhum.2016.00533

Norton, A., Zipse, L., Marchina, S., and Schlaug, G. (2009). Melodic intonation therapy: shared insights on how it is done and why it might help. *Ann. N. Y. Acad. Sci.* 1169, 431–436. doi: 10.1111/j.1749-6632.2009.04859.x

Popovici, M. (1995). Melodic intonation therapy in the verbal decoding of aphasics. *Rom. J. Neurol. Psychiatry* 33, 57–97.

Schuppert, M., Münte, T. F., Wieringa, B. M., and Altenmüller, E. (2000). Receptive amusia: evidence for cross-hemispheric neural networks underlying music processing strategies. *Brain* 123, 546–559. doi: 10.1093/brain/123.3.546

Sihvonen, A. J., Särkämö, T., Leo, V., Tervaniemi, M., Altenmüller, E., and Soinila, S. (2017). Music-based interventions in neurological rehabilitation. *Lancet Neurol.* 16, 648–660. doi: 10.1016/s1474-4422(17)30168-0

Sparks, R. W. (2008). "Melodic intonation therapy," in *Language Intervention Strategies in Aphasia and Related Neurogenic Communication Disorders*, ed. R. Chapey (Baltimore: Lippincott Williams &Wilkins), 837–851.

Spreen, O., and Risser, A. H. (2003). *Assessment of Aphasia*. Uk: Oxford University Press.

Tabei, K. I., Satoh, M., Nakano, C., Ito, A., Shimoji, Y., Kida, H., et al. (2016). Improved neural processing efficiency in a Chronic Aphasia patient following Melodic Intonation Therapy: a neuropsychological and functional MRI study. *Front. Neurol.* 9:148. doi: 10.3389/fneur.2016.00148

Van der Lugt-van Wiechen, K., and Verschoor, J. (1987). *Melodic Intonation Therapy, Nederlandse Bewerking*. Rotterdam: Stichting Afasie Rotterdam.

Van der Meulen, I., Van de Sandt-Koenderman, W. M. E., and Ribbers, G. M. (2012). Melodic Intonation Therapy: present controversies and future opportunities. *Arch. Phys. Med. Rehabil.* 93, 46–52. doi: 10.1016/j.apmr.2011.05.029

Wade, D. T., Hewer, R. L., David, R. M., and Enderby, P. M. (1986). Aphasia after stroke: natural history and associated deficits. *J. Neurol. Neurosurg. Psychiatr.* 49, 11–16. doi: 10.1136/jnnp.49.1.11

Wan, C. Y., Zheng, X., Marchina, S., Norton, A., and Schlaug, G. (2014). Intensive therapy induces contralateral white matter changes in chronic stroke patients with Broca's aphasia. *Brain Lang.* 9, 1–7. doi: 10.1016/j.bandl.2014.03.011

Wang, L. D., Liu, J. M., Yang, Y., Peng, B., and Wang, Y. L. (2019). Stroke prevention in China still faces huge challenges-Summary of "Stroke Prevention Report 2018 in China". *China Circ. J.* 34, 105–119. doi: 10.3969/j.issn.1000-3614.2019.02.001

Zhang, X. Y., Liu, S. H., and Wang, C. X. (2016). A case study of melody pronunciation therapy combined with therapeutic singing in the treatment of motor aphasia after stroke. *Chin. J. Stroke* 11, 761–765.

Zumbansen, A., Peretz, I., and Hébert, S. (2014). The combination of rhythm and pitch can account for the beneficial effect of Melodic Intonation Therapy on connected speech improvements in Broca's aphasia. *Front. Hum. Neurosci.* 8:592. doi: 10.3389/fnhum.2014.00592

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

ORIGINAL RESEARCH
published: 04 October 2021
doi: 10.3389/fpsyg.2021.712753

Musicianship Influences Language Effect on Musical Pitch Perception

*William Choi**

Academic Unit of Human Communication, Development, and Information Sciences, The University of Hong Kong, Hong Kong, SAR China

Given its practical implications, the effect of musicianship on language learning has been vastly researched. Interestingly, growing evidence also suggests that language experience can facilitate music perception. However, the precise nature of this facilitation is not fully understood. To address this research gap, I investigated the interactive effect of language and musicianship on musical pitch and rhythmic perception. Cantonese and English listeners, each divided into musician and non-musician groups, completed the Musical Ear Test and the Raven's 2 Progressive Matrices. Essentially, an interactive effect of language and musicianship was found on musical pitch but not rhythmic perception. Consistent with previous studies, Cantonese language experience appeared to facilitate musical pitch perception. However, this facilitatory effect was only present among the non-musicians. Among the musicians, Cantonese language experience did not offer any perceptual advantage. The above findings reflect that musicianship influences the effect of language on musical pitch perception. Together with the previous findings, the new findings offer two theoretical implications for the OPERA hypothesis—bi-directionality and mechanisms through which language experience and musicianship interact in different domains.

Keywords: OPERA, pitch, tone, rhythm, language-to-music transfer

OPEN ACCESS

Edited by:
Caicai Zhang,
The Hong Kong Polytechnic University, Hong Kong, SAR China

Reviewed by:
Claude Alain,
Rotman Research Institute (RRI), Canada
Akshay Maggu,
Hofstra University, United States
Linjun Zhang,
Beijing Language and Culture University, China

****Correspondence:***
William Choi
willchoi@hku.hk

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

Received: *21 May 2021*
Accepted: *13 September 2021*
Published: *04 October 2021*

Citation:
Choi W (2021) Musicianship Influences Language Effect on Musical Pitch Perception. Front. Psychol. 12:712753. doi: 10.3389/fpsyg.2021.712753

INTRODUCTION

Long-term musical experience facilitates speech perception (Pfordresher and Brown, 2009; Bidelman et al., 2010). This effect, known as *music-to-language transfer*,[1] largely undergirds theoretical models of cross-domain plasticity (Patel, 2011, 2012, 2014; Krishnan et al., 2012). Interestingly, there is emerging evidence of *language-to-music transfer* (Bidelman et al., 2013; Chen et al., 2016; Zhang et al., 2020). These studies generally showed that tone language experience enhanced musical pitch perception among non-musicians. However, these novel findings could not situate well in the OPERA hypothesis as it was designed for music-to-language transfer. Also, the OPERA hypothesis does not embody the interaction between musicianship and language experience, presumably because very few studies systematically manipulated both variables together (Cooper and Wang, 2012; Ngo et al., 2016; Maggu et al., 2018b). Apart from pitch, rhythm is also a common acoustic feature of music and speech (Zhang et al., 2020). As such, Patel (2012)

[1] In the current study, cross-domain transfer is said to occur when experience with perceptual attributes (e.g., tones) in one domain facilitates the sensitivity to perceptual attributes in a different domain (e.g., musical pitch). For example, language-to-music transfer refers to the facilitatory effect of language experience on music perception. This definition does not reject the idea that the facilitation is mediated by subcortical sensitivity to shared acoustic features (i.e., periodicity).

called for future studies to extend the OPERA hypothesis from pitch to rhythm. To broaden the OPERA hypothesis, the current study examined the interactive effects of musicianship and language experience on musical pitch and rhythmic perception.

The OPERA hypothesis theorizes how long-term musical experience increases neuronal sensitivity to perceptual attributes in the language domain, most notably tones (Patel, 2011). In the hypothesis, a clear conceptual distinction was made between perceptual attributes (e.g., tones and musical pitch) and acoustic features (e.g., periodicity). According to Patel (2011), music-to-language transfer will occur only when five conditions are met—Overlap, Precision, Emotion, Repetition, and Attention. Regarding Overlap, although different perceptual attributes (tones and musical pitch) are processed differently at the cortical level, the processing of their acoustic feature (i.e., periodicity) recruit overlapping subcortical networks. For Precision, music must require more nuanced processing than speech. For Emotion, strong positive emotion must be brought about by musical activities. In terms of Repetition, there must be a frequent repetition of the musical activities. For Attention, the musical activities must require focused attention. When the above conditions are met, musical experience will enhance neuronal precision in the subcortical area shared by music and language. Enhanced subcortical processing of the acoustic feature (i.e., periodicity) will in turn facilitate the processing of the linguistic perceptual attribute (i.e., tones).

Music-to-Language Transfer

There is mounting cross-sectional evidence that musicianship facilitates tone perception in different tasks, e.g., discrimination, identification, sequence recall, and word learning. Concerning discrimination, English musicians discriminated Mandarin tones more accurately than did English non-musicians (Alexander et al., 2005). On the one hand, this result implied that musicianship facilitated English listeners' tone discrimination. On the other hand, the perceptual facilitation might be speech general rather than specific to tones. In a later study, Italian musicians, Italian non-musicians, and Italian learners of Mandarin were presented with monosyllabic Mandarin word sequences with tonal and segmental violations (Delogu et al., 2010). Compared with the non-musicians, the musicians only detected tonal variations more accurately. This suggested that the music-to-language transfer was specific to tones. In a more recent study, English listeners heard pairs of Mandarin phrases, half of which contained a syllable with a deviant tone (i.e., the f0 level of the syllable was increased by 10%) (Zheng and Samuel, 2018). Compared with the English non-musicians, the English musicians were better able to detect the tonal differences. This indicated that music-to-language transfer was not limited to isolated words. As the above studies only used Mandarin tones, it remained unclear whether music-to-language transfer applied to more complex tone systems such as Cantonese (for a review of Cantonese tonal complexity, see Yip, 2002; Gu et al., 2007). In a Cantonese tone discrimination task, English musicians outperformed English non-musicians in half of the possible Cantonese tonal contexts (Choi, 2020). Despite the subtle differences between the tone discrimination studies, they generally provided evidence of music-to-language transfer. Remarkably, this transfer was not limited to Mandarin.

Music-to-language transfer also applied to tone identification and sequence recall. Following a brief familiarization of Mandarin tones, English musicians identified the Mandarin tones more accurately than did English non-musicians (Alexander et al., 2005; Lee et al., 2014). Critically, some tone identification studies reported the lack of correlation between musical pitch identification (i.e., absolute pitch) and Mandarin tone identification tasks (Lee and Lee, 2010; Lee et al., 2011, 2014). Does this lack of correlation indicate the absence of music-and-language relationship? In the only study which included English musicians and non-musicians, the musicians showed superior performance on Mandarin tone identification (Lee et al., 2014). Thus, the lack of correlation should not be taken to indicate the absence of music-to-language transfer. Instead, it merely reflected that the music-to-language transfer was not because the musicians had employed the perceptual mechanism of absolute pitch for Mandarin tone identification. In particular, enhanced neural encoding of periodicity might underlie a perceptual advantage on Mandarin tones (Patel, 2011, 2014). Going beyond identification, a recent study compared English musicians and non-musicians on their ability to recall Cantonese tone sequences (Choi, 2020). The English musicians outperformed the non-musicians on recalling contour tone sequences, indicating the presence of music-to-language transfer at the higher perceptual levels. Here, higher perceptual levels refer to the relative levels at which the perceptual operations are more complex than basic perceptual operations (e.g., forming phonological representations vs. judging the loudness of two beeps).

Concerning the higher perceptual levels, music-to-language transfer was also evident in tone-word learning. In a Mandarin tone-word learning experiment, English musicians and non-musicians were classified as successful (95% accuracy or above for two consecutive sessions) or less successful (less than 5% improvement for four consecutive sessions) learners (Wong and Perrachione, 2007). While only 22% of the non-musicians reached the successful criterion, as many as 88% of the musicians were classified as successful learners. Despite its small sample size (n = 17), the study provided initial evidence that music-to-language transfer applied to tone-word learning. With a more adequate sample size (n = 54), a later study compared English musicians, English non-musicians, Thai musicians, and Thai non-musicians on Cantonese tone word learning (Cooper and Wang, 2012). After training, the English musicians identified the tone words more accurately than did the English non-musicians. This convincingly reflected that music-to-language transfer was potent at the linguistic level, i.e., formation and recall of phonological-semantic links.

Aside behavioral evidence, there is ample neural evidence of music-to-language transfer (e.g., Wong et al., 2007; Bidelman et al., 2010; cf. Maggu et al., 2018a). At the subcortical level, English musicians showed stronger fundamental frequency-following response (FFR) to Mandarin tonal changes than English non-musicians (Wong et al., 2007). In a later study, English musicians even encoded two sections of the Mandarin rising tone more robustly than did Mandarin listeners

(Bidelman et al., 2010). The above findings situated well in the OPERA hypothesis—musical experience strengthens the subcortical neural network shared by music and language; and the enhancement of the subcortical plasticity was leveraged for tone perception (Patel, 2011, 2014).

Language-to-Music Transfer

Originally devised to account for music-to-language transfer, the OPERA hypothesis did not explicitly articulate about bidirectionality (Patel, 2011, 2012, 2014; see Asaridou and McQueen, 2013). Recall the Precision condition—for language-to-music transfer to occur, language must entail more precise pitch processing than music. However, Patel (2014) has argued that music requires finer pitch distinctions than language does—one semitone difference is perceptually salient in musical notes but not in lexical tones (Peretz and Hyde, 2003; Zatorre and Baum, 2012). Pertaining to the Emotion condition, Asaridou and McQueen (2013) believed that emotional reinforcement of speaking a tone language was hardly comparable to that of musical activities. As such, the authors reasoned that the OPERA hypothesis was not very, if at all, predictive of language-to-music transfer.

Interestingly, there is growing behavioral evidence on language-to-music transfer (Wong et al., 2012; Asaridou and McQueen, 2013; Bidelman et al., 2013). Lexically, tone languages (e.g., Cantonese and Mandarin) place a heavier demand on pitch than do non-tonal languages (e.g., Dutch, English, French, and Japanese) (Cutler, 2012). Relative to non-tonal language listeners, tone language listeners consistently showed superior performance on musical pitch perception tests. In the Online Identification Test of Congenital Amusia, Cantonese listeners outperformed English and French listeners on musical pitch perception (Wong et al., 2012). Even when non-verbal intelligence and working memory were controlled, Cantonese listeners outperformed English non-musicians on self-designed musical pitch memory and discrimination tasks (Bidelman et al., 2013). This further indicated that tone language experience enhanced not only basic auditory sensitivity but also complex music perception. Besides Cantonese listeners, there were similar findings from other tonal populations, e.g., Mandarin listeners. In the Montreal Battery of Evaluation of Amusia, Mandarin listeners discriminated pitch more accurately than did Dutch listeners (Chen et al., 2016). In the melody subtest of the well-validated Musical Ear Test, Mandarin listeners scored higher than Japanese listeners (Zhang et al., 2020). Collectively, the above studies have suggested that speaking a tone language sharpens musical pitch sensitivity.

Beyond behavioral advantages, language-to-music transfer also enhances the neural encoding of musical pitch (Bidelman et al., 2010, 2011). Bidelman et al. (2010) compared English musicians, English non-musicians, and Mandarin non-musicians on their FFR to musical pitch interval and Mandarin tone. Relative to the English non-musicians, the Mandarin non-musicians showed a higher pitch tracking accuracy on musical pitch interval. In line with the OPERA hypothesis, this result suggested that tone language experience enhanced the subcortical encoding of musical pitch (Patel, 2011, 2014). Could this enhanced neural encoding explain the behavioral advantage enjoyed by tone language speakers on musical pitch perception? In a later study, Bidelman et al. (2011) tested English musicians, English non-musicians, and Mandarin non-musicians on behavioral and neural perception of musical pitch. While the Mandarin non-musicians showed stronger FFR than English non-musicians, the former did not outperform the latter on behavioral musical pitch discrimination. This seemed to indicate that although tone language experience enhanced the subcortical processing of musical pitch, this neural enhancement did not yield any behavioral perceptual advantage. However, the results should be interpreted with caution given (a) the small sample size (n = 11 per group) and (b) the preponderance of studies showing that Cantonese/Mandarin non-musicians outperformed Dutch/English/French/Japanese non-musicians on behavioral measures of musical pitch perception (Wong et al., 2012; Asaridou and McQueen, 2013; Bidelman et al., 2013; Chen et al., 2016; Zhang et al., 2020). It remained unclear whether enhanced neural encoding of musical pitch could explain the behavioral advantage on musical pitch perception. However, this does not underscore the collective neural evidence that tone language experience enhanced the subcortical processing of musical pitch (Bidelman et al., 2010, 2011).

Interactive Effects of Language and Musicianship on Speech and Music Perception

Although cross-domain transfer was well supported by empirical evidence, its exact nature has seldom been explored. Regarding music-to-language transfer, only few studies examined the interaction between tone language experience and musicianship (Cooper and Wang, 2012; Maggu et al., 2018b). Cooper and Wang (2012) investigated whether the combination of both tone language experience and musicianship would offer extra advantage above either experience. Specifically, they compared the Cantonese tone word learning proficiencies between Thai musicians, Thai non-musicians, English musicians, and English non-musicians. Resonating previous studies on music-to-language transfer, the English musicians had a greater learning success than the English non-musicians. However, music-to-language transfer was not observed among the Thai listeners. Intriguingly, the Thai musicians even tended to perform poorer than the Thai non-musicians. The authors attributed this non-additive effect to an internal conflict between linguistic and music perceptual mechanisms. In a related study, English musicians also outperformed English non-musicians on Thai tone word learning (Maggu et al., 2018b). Similar to the earlier finding, the Mandarin musicians tended to perform poorer than the Mandarin non-musicians. Interestingly, the study also included double tone language (i.e., Cantonese-Mandarin bilingual) groups. Compared with the Mandarin listeners, the Cantonese-Mandarin bilingual listeners did not exhibit any perceptual advantage. In other words, speaking an additional tone language did not provide any extra benefit on tone word learning. Taken together, the available studies showed that tone language experience influenced the effect of musicianship on tone word learning (Cooper and Wang, 2012; Maggu et al., 2018b).

Although language-to-music transfer was well supported by empirical evidence, its exact nature was not fully explored (Wong et al., 2012; Asaridou and McQueen, 2013; Bidelman et al., 2013; Chen et al., 2016; Zhang et al., 2020). In the context of music perception, it remains unclear whether and how language experience and musicianship interact. Most of the available studies only manipulated the language variable, and the lack of musician groups rendered them impossible to test the interaction (e.g., Wong et al., 2012; Chen et al., 2016; Zhang et al., 2020). One study attempted to manipulate both language (Cantonese and English) and musicianship variables, but the musician group only contained English musicians (Bidelman et al., 2013). In a similar vein, the lack of Cantonese musicians made it impossible to systematically test the interaction between musicianship and language experience. A recent study compared Cantonese musicians and Cantonese non-musicians on FFR to Cantonese tones and musical pitch (Maggu et al., 2018a). Regarding music-to-language transfer, the Cantonese musicians showed stronger FFR to musical pitch than the Cantonese non-musicians. The authors concluded that the combination of Cantonese language experience and musicianship offered extra perceptual advantage on musical pitch than did either experience. However, the authors also acknowledged that the lack of English musicians and non-musicians in their study rendered it impossible to fully test the music and language interaction. Also, the previous study only included a subcortical measure, i.e., FFR, which did not always correlate with behavioral musical pitch perception (Maggu et al., 2018a; Yu and Zhang, 2018). So, it remained unclear as to how such an interaction would manifest behaviorally. Limitations aside, this study provided preliminary evidence of the interaction of music and language on musical pitch perception (Maggu et al., 2018a).

Among the studies on language-to-music transfer, one study systematically manipulated both musicianship and language experience together (Ngo et al., 2016). Vietnamese and English listeners, each split into musician and non-musician groups, were assessed with the Cochran-Weiss-Shanteau index of expertise and the Montreal Battery of Evaluation of Amusia. Importantly, the interaction between language experience and musicianship was not significant. More surprisingly, the main effect of language experience was not significant too on both musical tests. While the lack of interaction might be possible, the lack of language-to-music transfer seemed unusual given substantial previous evidence (Bidelman et al., 2013; Chen et al., 2016; Wong et al., 2012; Zhang et al., 2020). Critically, Ngo et al. (2016) only recruited eight participants per group. This very small sample size might have rendered the statistical power too small to detect any effects. Also, their Vietnamese listeners grew up in the U.S. and only half of them reported having achieved native Vietnamese proficiency. Given the above limitations, the current study re-examined the interaction between language experience and musicianship with a larger (31 participants per group) and representative (native Cantonese speakers born and raised in Hong Kong; native English speakers born and raised in the United States) sample. Given the preliminary evidence that Cantonese musicians had stronger FFR to musical pitch than Cantonese non-musicians, I anticipated an interaction between language experience and musicianship on musical pitch perception (Maggu et al., 2018a). Specifically, musicianship was expected to amplify the language-to-music transfer.

The OPERA Hypothesis and Rhythmic Perception

As mentioned previously, rhythm is another acoustic feature shared by music and speech. Musically, rhythm represents an ordered alteration of long and short notes regardless of the absolute duration of each note. Similarly, speech rhythm represents the timing of successive vowel and consonant sequences (Hayes, 1989). Speakers of different languages use rhythm differently. Based on rhythmical properties, languages are typically categorized as stress-timed (e.g., English), mora-timed (e.g., Japanese), and syllable-timed (e.g., Cantonese) (Pike, 1945; Ladefoged, 1975). In stress-timed languages, unstressed syllables are often compressed to fit in the constant interval between stressed syllables (Nespor et al., 2011). As such, successive intervals between vowels vary rigorously (i.e., high vocalic interval variability) in these languages (Ladefoged and Johnson, 2011). In syllable-timed languages, syllables have highly similar durations, rendering the vocalic interval relatively constant (i.e., low vocalic interval variability). In mora-timed languages, contrastive vowel length characterizes mora, a syllabic sub-unit which organizes speech (Otake et al., 1993). As such, stress-timed and mora-timed languages have higher vocalic interval variabilities than syllable-timed languages (Nespor and Vogel, 1986; Warner and Arai, 2001; Grabe and Low, 2002). Cross-language differences aside, rhythm (like pitch) is a common feature of music and speech. However, the OPERA hypothesis has seldom been discussed in relation to rhythmic perception (Patel, 2012).

In a follow-up paper on refining the OPERA hypothesis, Patel (2012) raised the possibility that the OPERA hypothesis might apply to rhythmic perception. For music-to-language transfer, there was behavioral and neural evidence of enhanced speech rhythm sensitivity among musicians (Marie et al., 2011; Cason et al., 2015; Magne et al., 2016; Choi, 2021b). This suggested that the OPERA hypothesis also applied to rhythmic perception, at least unidirectionally (i.e., music-to-language). Similar evidence on language-to-music transfer was scarce. Two studies investigated the effect of language experience on rhythmic perception (Wong et al., 2012; Zhang et al., 2020). In the more recent study, Zhang et al. (2020) tested Mandarin and Japanese listeners with the Musical Ear Test (Wallentin et al., 2010). The Japanese listeners outperformed the Mandarin listeners on the rhythm subtest, presumably because Japanese had a higher vocalic interval variability than Mandarin. Together with prior evidence on music-to-language transfer in rhythmic perception, this finding implied that the transfer was bidirectional (Marie et al., 2011; Cason et al., 2015; Magne et al., 2016).

Being a stress-timed language, English has a higher vocalic interval variability than Cantonese (Ladefoged, 1975; Grabe and Low, 2002). Thus, it was reasonable to hypothesize that English listeners would outperform Cantonese listeners on rhythmic perception. Counterintuitively, Wong et al. (2012) reported that English listeners and Cantonese listeners performed similarly on rhythmic perception. Concerning rhythmic perception, this finding did not support language-to-music transfer at least among English vs. Cantonese listeners. Critically, methodological issues necessitate a re-examination of this preliminary conclusion. Firstly, a ceiling effect was shown on the rhythmic measure, probably because the congenital amusia screening test was too easy for typical listeners (Peretz et al., 2008). Secondly, despite the role of non-verbal intelligence in auditory perception, such measure had not been controlled (even in the study by Bidelman et al., 2013; Tang et al., 2016; Choi, 2020; Zhang et al., 2020). Going beyond these methodological limitations, the present study adopted the Musical Ear Test, the rhythmic subtest of which did not show any ceiling effect on speakers of syllable-timed languages (Wallentin et al., 2010; Zhang et al., 2020). Similar to the aforementioned research question on musical pitch, the potential interaction between musicianship and language experience on rhythmic perception was also explored.

To broaden the OPERA hypothesis, the present study examined the interactive effects of language experience and musicianship on music perception. Of particular interest was whether and how musicianship influenced the language effects on musical pitch and rhythmic perception. Based on preliminary neural evidence, I anticipated that musicianship would amplify the language effect on musical pitch perception (Maggu et al., 2018a). In other words, Cantonese musicians were expected to outperform Cantonese non-musicians, English musicians, and English non-musicians. Regarding rhythmic perception, I expected that English musicians would outperform English non-musicians, Cantonese non-musicians, and Cantonese musicians. Given the role of non-verbal intelligence in pitch perception, it was also measured and controlled as necessary (Tang et al., 2016; Choi, 2020, 2021a).

MATERIALS AND METHODS

Participants

To abide by the social distancing rules associated with COVID-19, data collection was switched from face-to-face to online. Ethical approval was obtained from the University Human Research Ethics Committee for the research project (Ref. no. A2019-2020-0036). Thus, 62 Cantonese (24 males, 38 females), and 62 English (26 males, 34 females, and 2 undisclosed) listeners were recruited via email and Prolific,[2] respectively. Prior to data collection, all participants completed an initial online or phone screening. All Cantonese listeners reported that they (i) were living in Hong Kong, (ii) spoke Cantonese as a first language, and (iii) had normal hearing. All English listeners reported that they (iv) were living in the United States, (v) spoke English as a first language, and (vi) had normal hearing.

Based on the pre-established criteria, musicians were individuals who (a) had received 7 or more years of continuous music training and (b) could play at least one music instrument (Choi, 2020, 2021b). Non-musicians were individuals who (c) had never received more than 2 years of music training, (d) had not received any music training in the past 5 years, and (e) could not play any music instrument.

Participants were tested on an online experiment platform (Gorilla Experiment Builder)[3] (Anwyl-Irvine et al., 2020; Tsantani and Cook, 2020; Jasmin et al., 2021). They were asked to sit comfortably in a quiet environment and wear headphones. An automatic procedure ensured that the participants were using a computer but not phones or tablets. After giving written consent, the participants filled out a language and music background questionnaire (Choi et al., 2017, 2019; Choi, 2021b). Prior to the Musical Ear Test, the participants could test and adjust the sound volume to their satisfaction (Wallentin et al., 2010). Following a written description of the task, the Musical Ear Test began. Upon completion of the Musical Ear Test, the participants completed the digital short form of the Raven's 2 Progressive Matrices Clinical Edition (Raven et al., 2018). Between each task, the participants were given opportunities to take breaks at their own pace. To prevent prolonged idle time, an overall experimental time limit of 120 min was set for each participant.

To test whether the participants remained attentive throughout the study, five attention-check trials were embedded in the perceptual tasks. On each attention-check trial, two identical audio stimuli were presented (see **Supplementary Materials**). Participants then judged whether the two sounds were different. With acoustically identical stimuli, these attention-check trials could be answered easily. To be empirically stringent, only one mistake on the attention-check trials was allowed (i.e., 80% accuracy or above). As such, one Cantonese musician, one Cantonese non-musician, and one English musician were removed from the dataset.

Offline screening of the language background questionnaires showed that three English musicians and one English non-musician had learnt Cantonese or Mandarin as a second language. These participants were excluded from the dataset. Thus, the final sample consisted of 30 Cantonese musicians, 30 Cantonese non-musicians, 27 English musicians, and 30 English non-musicians.

In the final sample, all Cantonese listeners had learnt English as a second language. This is because English language education is compulsory in Hong Kong since Grade 1. Among the English listeners, only eight reported having learnt a second language. Specifically, four English musicians and four English non-musicians learnt Farsi, Hindi, Polish, Portuguese, Punjabi, Spanish, or Urdu as a second language. None of the English listeners had learnt any tone language or resided in any tone language-speaking country. The demographic, language, and music backgrounds of all participants are summarized in **Tables 1–3**. The very high mean accuracies on the attention trials suggested that the participants remained attentive during the experiment (M_{CM} = 97%, SD_{CM} = 7%; M_{CNM} = 97%,

[2] prolific.co

[3] www.gorilla.sc

TABLE 1 | Age, years of music training, onset age of music training, and non-verbal intelligence of the Cantonese and English musicians and non-musicians.

Group	Chronological age in years (*SD*)	Years of music training (*SD*)	Onset age of music training (*SD*)	Non-verbal intelligence (*SD*)
CM	23.8 (4.0)	10.8 (2.9)	7.8 (3.5)	17.4 (3.3)
CNM	24.9 (5.4)	0.4 (0.7)	11.5 (6.2)	16.6 (2.8)
EM	22.5 (4.7)	10.4 (3.2)	9.0 (3.5)	16.2 (3.8)
ENM	27.8 (6.3)	0.4 (0.7)	11.6 (2.9)	14.9 (5.3)

CM, Cantonese musician; CNM, Cantonese non-musician; EM, English musician; ENM, English non-musician.

TABLE 2 | Language background of the bilingual Cantonese and English musicians and non-musicians.

Group	Number of bilinguals (max)	L1 frequency of use (*SD*)	L2 frequency of use (*SD*)	L1 proficiency (*SD*)	L2 proficiency (*SD*)
CM	30 (30)	5.0 (0.0)	3.3 (0.9)	4.9 (0.4)	3.3 (0.8)
CNM	30 (30)	5.0 (0.0)	2.9 (0.9)	4.9 (0.3)	3.4 (0.6)
EM	4 (27)	4.9 (0.4)	2.2 (1.2)	5.0 (0.0)	2.7 (1.0)
ENM	4 (30)	4.8 (0.7)	2.0 (0.9)	5.0 (0.0)	2.6 (1.6)

Frequency ranges from 0 (never) to 5 (always); Proficiency ranges from 0 (non-proficient) to 5 (highly proficient).

SD_{CNM} = 8%; M_{EM} = 94%, SD_{EM} = 9%; M_{ENM} = 96%, SD_{ENM} = 8%).

Musical Ear Test

The Music Ear Test was adopted to assess musical pitch and rhythmic discrimination (Wallentin et al., 2010). The test was validated in previous studies and strongly correlated with other musical tests such as the Montreal Battery of Evaluation of Amusia (e.g., Wallentin et al., 2010; Chen et al., 2016). It has been vastly used to measure musical aptitude in Eastern and Western populations (e.g., Chen et al., 2016; Yates et al., 2019; Zhang et al., 2020).

The melody subtest contained 52 pairs of piano-played melodic phrases. They were presented audibly with an AX paradigm. On each trial, two melodic phrases with duration of one measure were played at 100 beats per minute. The participants then judged whether the melodic phrases were different. There were 26 “same” and 26 “different” trials, each carrying one point (i.e., maximum possible score = 52). All “different” trials contained a pitch violation. On half of the “different” trials, the pitch violation also caused a pitch contour change. The melody subtest began with two practice trials with feedback. No feedback was given on the experimental trials. To test the internal consistency of the items, a Cronbach’s alpha reliability coefficient was obtained from the performances of all participants across the 52 trials. The internal consistency of the melody subtest was satisfactory (Cronbach’s α = 0.76).

The rhythm subtest contained 52 pairs of rhythmical phrases generated by 4–11 wood block beats. It had the same procedure as the melody subtest. Presented in a randomized order, 31

TABLE 3 | Musical experience of the musicians.

Participant	Onset age (years old)	Amount of music training (years)	First instrument	Second instrument	Third instrument
CM1	5.00	15.00	Piano		
CM2	9.00	11.00	Flute		
CM3	10.00	10.00	Violin		
CM4	8.00	7.00	Tuba	Trumpet	
CM5	9.00	17.00	Oboe	Piano	
CM6	6.00	15.00	Flute	Horn	
CM7	3.00	15.00	Piano	Viola	Recorder
CM8	3.00	14.00	Piano		
CM9	5.00	14.00	Piano		
CM10	7.00	14.00	Violin		
CM11	5.00	13.00	Piano	Flute	
CM12	5.00	13.00	Piano		
CM13	5.00	13.00	Piano		
CM14	6.00	12.00	Piano		
CM15	7.00	11.00	Piano		
CM16	10.00	10.00	Violin	Piano	Others
CM17	13.00	10.00	Guitar	Bagpipe	
CM18	7.00	10.00	Piano	Clarinet	Guitar
CM19	6.00	10.00	Drums		
CM20	8.00	10.00	Piano	Drum	
CM21	7.00	9.00	Piano	Guitar	Drum
CM22	13.00	9.00	Trumpet		
CM23	20.00	9.00	Piano		
CM24	9.00	9.00	Piano		
CM25	6.00	8.00	Piano	Guitar	
CM26	8.00	8.00	Piano		
CM27	4.00	7.00	Piano	Guitar	
CM28	9.00	7.00	Clarinet		
CM29	11.00	7.00	Violin		
CM30	10.00	7.00	Piano		
EM1	11.00	17.00	Flute	Cello	
EM2	5.00	17.00	Piano	Bass	Organ
EM3	9.00	11.00	Flute	Piccolo	Others
EM4	10.00	9.00	Clarinet	Piano	Violin
EM5	6.00	7.00	Piano	Flute	Guitar
EM6	18.00	7.00	Piano		
EM7	12.00	7.00	Saxophone		
EM8	9.00	7.00	Piano		
EM9	6.00	16.00	Violin	Piano	
EM10	14.00	15.00	Piano		
EM11	5.00	14.00	Violin	Piano	
EM13	6.00	13.00	Dilruba	Harmonium	
EM14	7.00	12.00	Piano	Ukulele	
EM15	6.00	12.00	Piano		
EM16	9.00	11.00	Flute		
EM17	7.00	11.00	Piano	Clarinet	Others
EM18	11.00	11.00	Cello		
EM19	7.00	10.00	Piano	Cello	
EM21	4.00	10.00	Guitar	Piano	
EM23	8.00	10.00	Piano	Clarinet	Others

(Continued)

TABLE 3 | (Continued)

Participant	Onset age (years old)	Amount of music training (years)	First instrument	Second instrument	Third instrument
EM24	15.00	9.00	Tuba	Trombone	Others
EM25	10.00	9.00	Keyboard	Mallets	
EM26	8.00	8.00	Saxophone	Piano	
EM27	10.00	8.00	French horn	Piano	
EM28	5.00	7.00	Piano	Clarinet	
EM29	14.00	7.00	Bass		
EM30	11.00	7.00	Saxophone	Clarinet	

trials contained even subdivisions of the beat whereas 21 trials contained triplets. This resulted in varying rhythmic complexities across trials. Like the melody subtest, there were 26 "same" and 26 "different" trials and the maximum possible score was 52. On each "different" trial, there was one rhythmic change (refer to Wallentin et al., 2010, p. 189; Zhang et al., 2020, p. 387 for audio and visual illustrations). Reliability analysis showed a fair internal consistency of the rhythm subtest (Cronbach's α = 0.64).

Raven's Test

The digital short form of the Raven's 2 Progressive Matrices Clinical Edition was adopted. The digital short form contained 24 randomly selected items. On each trial, a picture with a missing pattern was presented along with five possible options. The participants then chose the option which could best complete the picture. Task administration and scoring were done according to the test manual (Raven et al., 2018). The time limit was 20 min. The same reliability analysis was conducted on the 24 items. The internal consistency was satisfactory (Cronbach's α = 0.79).

RESULTS

Preliminary Analysis

To evaluate whether the four groups matched on age and non-verbal intelligence, two-way univariate analysis of variance (ANOVAs) were conducted separately on age and non-verbal intelligence with language (Cantonese and English) and musicianship (musician and non-musician) as the between-subjects factors. Regarding age, the main effect of language was not significant, $F(1, 113) = 0.69$, $p = 0.41$. However, the main effect of musicianship, $F(1, 113) = 11.27$, $p = 0.001$, $\eta_p^2 = 0.09$, and the interaction between language and musicianship, $F(1, 113) = 4.56$, $p = 0.04$, $\eta_p^2 = 0.04$, were significant. Simple effects analysis showed that the English non-musicians were older than the Cantonese non-musicians, $F(1, 113) = 4.52$, $p = 0.04$, $\eta_p^2 = 0.04$. The English and the Cantonese musicians matched on age, $F(1, 113) = 0.83$, $p = 0.36$.

Concerning non-verbal intelligence, the main effect of language was significant, $F(1, 113) = 4.21$, $p = 0.04$, $\eta_p^2 = 0.04$, but not the main effect of musicianship, $F(1, 113) = 2.06$, $p = 0.15$. Pairwise comparison showed that the Cantonese listeners outperformed the English listeners, $F(1, 113) = 4.21$, $p = 0.04$, $\eta_p^2 = 0.04$. The interaction between language and musicianship was not significant, $F(1, 113) = 0.14$, $p = 0.71$. Given the group differences in age and non-verbal intelligence, these two variables were controlled in the main analysis.

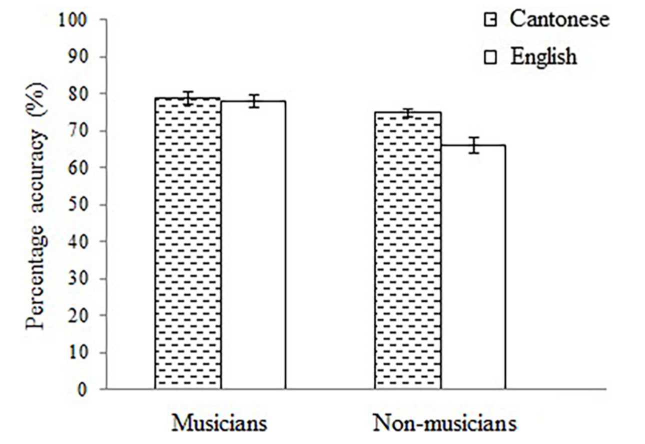


FIGURE 1 | Mean pitch score of the Cantonese musicians, Cantonese non-musicians, English musicians, and English non-musicians. The error bars represent the standard error of the mean.

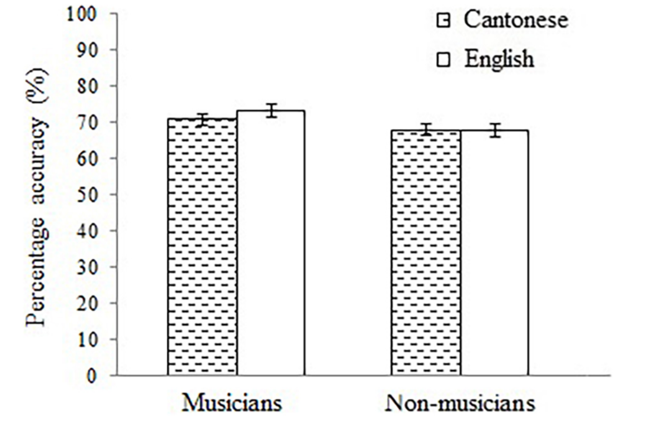


FIGURE 2 | Mean rhythm score of the Cantonese musicians, Cantonese non-musicians, English musicians, and English non-musicians. The error bars represent the standard error of the mean.

Main Analysis

To ascertain whether musicianship influenced the language effect on musical pitch and rhythmic perception, a two-way MANCOVA was conducted on pitch and rhythmic scores with language (Cantonese and English) and musicianship (musician and non-musician) as the between-subjects factors, and age and non-verbal intelligence as the covariates (see **Figure 1**). MANCOVA revealed significant main effects of language, $\Lambda = 0.91$, $F(2, 110) = 5.78$, $p = 0.004$, $\eta_p^2 = 0.10$, and musicianship, $\Lambda = 0.83$, $F(2, 110) = 11.58$, $p < 0.001$, $\eta_p^2 = 0.17$, and the interaction between language and musicianship, $\Lambda = 0.94$, $F(2, 110) = 3.61$, $p = 0.03$, $\eta_p{}^2 = 0.06$.

Concerning musical pitch perception, there were significant main effects of language, $F(1, 111) = 5.78$, $p = 0.02$, $\eta_p{}^2 = 0.05$, and musicianship, $F(1, 111) = 23.28$, $p < 0.001$, $\eta_p{}^2 = 0.17$. Consistent with previous studies, a clear language-to-music transfer was found—knowing Cantonese seemed to offer the listeners a perceptual advantage on musical pitch perception. Expectedly, long-term musical experience also facilitated musical pitch perception. Crucially, the interaction between language and musicianship was also significant, $F(1, 111) = 7.21$, $p = 0.01$, $\eta_p{}^2 = 0.06$. This hinted that the language-to-music transfer was influenced by musicianship. Indeed, simple effects analysis revealed that the Cantonese outperformed the English listeners among the non-musicians, $F(1, 111) = 12.97$, $p < 0.001$, $\eta_p{}^2 = 0.11$, but not among the musicians, $F(1, 111) = 0.04$, $p = 0.85$. This further adds that knowing Cantonese is helpful only to non-musicians.

To further elucidate the interaction, a one-way ANCOVA was conducted on pitch score with group (Cantonese musicians, Cantonese non-musicians, English musicians, and English non-musicians) as the between-subjects factor, and age and non-verbal intelligence as the covariates. The main effect of group was significant, $F(3, 111) = 11.21$, $p < 0.001$, $\eta_p{}^2 = 0.23$. Pairwise comparisons with Bonferroni adjustments showed that the English non-musicians performed poorer than the Cantonese musicians, $p < 0.001$, Cantonese non-musicians, $p = 0.003$, and English musicians, $p < 0.001$. However, the Cantonese musicians, Cantonese non-musicians, and English musicians performed similarly, ps = 0.563, 0.423, 1.00.

Regarding rhythmic perception, the main effect of musicianship was significant, $F(1, 111) = 6.20$, $p = 0.01$, $\eta_p{}^2 = 0.05$, but not the main effect of language, $F(1, 111) = 1.10$, $p = 0.30$. The interaction between musicianship and language was not significant, $F(1, 111) = 0.99$, $p = 0.32$. Expectedly, long-term music training facilitated rhythmic perception (see **Figure 2**). However, language-to-music transfer was not evident in rhythmic perception, nor was any interaction.

DISCUSSION

The present study investigated the interactive effects of language experience and musicianship on music perception. An interaction between language experience and musicianship was found on musical pitch perception—Cantonese language experience facilitated musical pitch perception among the non-musicians but not among the musicians. Regarding rhythmic perception, the musicians consistently outperformed the non-musicians. No language or interactive effects were found.

Musicianship Influences Language Effect on Musical Pitch Perception

The most crucial finding is that musicianship influences language-to-music transfer. It is known that tone language experience enhances musical pitch perception (Wong et al., 2012; Bidelman et al., 2013; Chen et al., 2016; Zhang et al., 2020). Strikingly, the present study found that language-to-music transfer occurred only among non-musicians. Given relevant musical experience, Cantonese language experience was no longer beneficial to musical pitch perception. This is reminiscent of two previous studies which investigated music-language interaction in an opposite direction, i.e., music-to-language transfer (Cooper and Wang, 2012; Maggu et al., 2018b). Like the present study, Cooper and Wang found significant interactive effects of musicianship and language experience (though on tone word learning). While tone language experience and musicianship each led to enhanced tone word learning, their beneficial effects did not add up. Specifically, musicians outperformed non-musicians only among the English listeners but not among the Thai/Mandarin listeners. This previous finding could be re-interpreted as such—musicianship aided tone perception only in the absence of tone language experience (Cooper and Wang, 2012; Maggu et al., 2018b). Though the present study focuses on an opposite direction, i.e., language-to-music transfer, a striking similarity was found—Cantonese language experience facilitated musical pitch perception only in the absence of long-term musical experience. Tentatively, these collective results indicate the potential need to add a new condition "**L**ack of relevant experience" to the OPERA hypothesis: For cross-domain transfer to occur, one must not possess long-term experience in the target domain.

With the potential condition "Lack" now identified, it would be interesting to see how it operates. Concerning music-to-language transfer, Thai/Mandarin musicians performed similarly as English non-musicians and even tended to perform poorer than Thai/Mandarin non-musicians (though not statistically significant; Cooper and Wang, 2012; Maggu et al., 2018b). Cooper and Wang (2012) ascribed this to an internal conflict between language and music systems. Specifically, the authors argued that music training drove Thai musicians to attend to the fine-grained acoustic details of Cantonese tones; whereas Thai language experience oriented them to ignore these details and rely on coarse tonal percepts. In the context of language-to-music transfer, the present study shows a seemingly different case. Nuanced analysis showed that the Cantonese musicians outperformed the English non-musicians. Also, the Cantonese musicians performed similarly as the Cantonese non-musicians and the English musicians. Unlike the previous studies which reflected an internal conflict, our musicians simply did not benefit from Cantonese language experience.

Speculatively, there were two possible causes for the above phenomenon. Music entails finer pitch distinction than does language (Patel, 2011, 2014). Also, musical pitch is more functionally relevant to music than to language. Conceivably, music training exerts stronger influence on musical pitch perception than does language experience. Possibly, musicianship might have already saturated the *perceptual capacity* for musical pitch or periodicity, so language experience had no effect on it. The term perceptual capacity is used here because it remains uncertain whether the saturation occurred at the cortical or subcortical levels. As mentioned above, the previous FFR study only included Cantonese musicians and non-musicians (Maggu et al., 2018a). Without measuring the FFR of English musicians, it was impossible to ascertain whether musicianship saturated the subcortical plasticity to periodicity which could have otherwise been enhanced by language experience. If the saturation occurs at the subcortical level, Cantonese musicians and English musicians are expected to show similar FFR on musical pitch perception. The other possible cause of the above phenomenon was that the musicians had developed a highly specialized cortical mechanism for musical pitch perception (Tervaniemi et al., 2006; Rogalsky et al., 2011). As such, the musicians needed not leverage on language experience for musical pitch perception. By contrast, the non-musicians might at least partially leverage on their language experience for musical pitch perception. For the Cantonese non-musicians, their linguistic experience in tone perception might have translated into perceptual benefits on musical pitch. This claim is supported by neural evidence that Cantonese non-musicians showed left hemispheric lateralization on both tone and musical pitch perception (Gu et al., 2013). To further verify this claim, future fMRI studies can examine whether musical pitch and tone perception recruit overlapping or separate cortical regions among Cantonese musicians, Cantonese non-musicians, English musicians, and English non-musicians.

At first glance, the present results contrasted the previous neural findings (Maggu et al., 2018a). On the one hand, Maggu et al. (2018a) reported that Cantonese musicians had stronger FFR to musical pitch than Cantonese non-musicians. On the other hand, our Cantonese musicians did not outperform the Cantonese non-musicians on behavioral musical pitch perception. Importantly, subcortical processing only underlies one of the many cognitive operations involved in behavioral perception (Holder, 1992; Law et al., 2013). While subcortical neural encoding is a *sine qua non*, behavioral perceptual ability may hinge on other cognitive operations. Indeed, there was evidence that FFR measures did not correlate with behavioral perception (English listeners; Yu and Zhang, 2018). Thus, enhanced FFR of Cantonese musicians does not necessarily indicate that they have a behavioral advantage on musical pitch perception.

Bidirectional OPERA Hypothesis: Revisiting "Precision"

The present result enriches the body of evidence on language-to-music transfer (see text footnote 1) (Wong et al., 2012; Bidelman et al., 2013; Chen et al., 2016; Zhang et al., 2020). In the Musical Ear Test, the Cantonese non-musicians discriminated musical pitch height and contour more accurately than did the English non-musicians. This is consistent with previous studies showing that tone language listeners outperformed non-tonal language listeners on musical pitch perception (Wong et al., 2012; Bidelman et al., 2013; Chen et al., 2016; Zhang et al., 2020). Collectively, the present and previous findings inform the OPERA hypothesis about bidirectionality—cross-domain transfer not only occurs from music to language, but also from language to music. As mentioned in the Introduction, the OPERA hypothesis was originally devised to account for music-to-language transfer. Nevertheless, it has good potential to account for language-to-music transfer. I describe below some potential directions on how the OPERA hypothesis could be modified to broaden its coverage.

The converging evidence of language-to-music transfer, herein and in previous studies, motivates a reconsideration of how "Precision" should be defined (Wong et al., 2012; Bidelman et al., 2013; Chen et al., 2016; Zhang et al., 2020). In the original paper of the OPERA hypothesis, Precision was defined as the extent to which "a perceiver *requires* detailed information about the patterning of that feature in order for adequate communication to occur" (Patel, 2011; p. 7). In terms of the grain size, a pitch movement of only one semitone is structurally important in music (e.g., from C to C#) but not in Cantonese tones (Chow, 2012). By contrast, Cantonese tonal variations typically involve more than three semitones (Yiu, 2013). Regarding the word "*requires*", neutralizing tonal information in Mandarin sentences did not impede comprehension among native Mandarin listeners (Patel et al., 2010). As such, Patel (2011) argued that language processing could hardly entail a high precision relative to music processing.

With its original definition of Precision, the OPERA hypothesis was not very (if at all) predictive of language-to-music transfer (Patel, 2011, 2012; Asaridou and McQueen, 2013; Patel, 2014). This was because tone perception hardly requires more precision on periodicity encoding than musical pitch perception (Patel, 2011, 2014). The present study found robust evidence that, although Cantonese tone perception required less precision than musical pitch perception, Cantonese language experience enhanced musical pitch sensitivity. Together with similar previous findings, this finding implies the need to revisit the definition of Precision. As described above, Patel (2011) viewed Precision as domain-relative, i.e., music vs. language, in which music always prevailed. Critically, the present study indicates that Precision should be re-referenced on listeners—relative to English listeners, Cantonese listeners engaged in more precise pitch perception in their first language (due to lexical tones); this precision positively transferred to the music domain. This new specification also applies potently to music-to-language transfer—musicians had more precise musical pitch perception than non-musicians; and this precision aided lexical tone perception. This new conception of Precision may help the OPERA hypothesis cover bidirectional, and more specifically, language-to-music transfer.

Absence of Language-to-Music Transfer on Rhythmic Perception

Regarding rhythmic perception, the present study found no evidence of language-to-music transfer. Originally, it was hypothesized that the English listeners would outperform the Cantonese listeners since English had a higher vocalic interval variability than Cantonese (Nespor and Vogel, 1986; Warner and Arai, 2001; Grabe and Low, 2002). However, no significant main effect of language was shown, indicating that English language experience did not lead to better performance beyond Cantonese language experience. In fact, this finding is consistent with a previous study which reported that Cantonese and English listeners performed similarly on rhythmic perception

(Wong et al., 2012). The present study has further added that task easiness does not explain the lack of group difference, because the rhythmic subtest of the Musical Ear Test showed no ceiling effect.

There are two possible explanations for the lack of language-to-music transfer in rhythmic perception. Firstly, a previous study showed that bilinguals having learnt two languages with different rhythmic properties (syllable-timed Turkish and stress-timed German) had enhanced rhythmic perception relative to those having learnt two languages with similar rhythmic properties (stress-timed German and English) (Roncaglia-Denissen et al., 2013). In the present study, a majority (86%) of the English listeners were monolinguals. However, the Cantonese listeners in the present and previous studies were all L2 English learners, meaning that they had learnt syllable-timed (i.e., Cantonese) and stress-timed (i.e., English) languages (Wong et al., 2012). It was possible that native English language experience indeed benefited the English listeners' rhythmic perception; but then this advantage was masked by the Cantonese listeners' enhanced rhythmic perception associated with bilingual experience. As in many Asian countries, English language instruction is compulsory in Hong Kong, so it would not be feasible to recruit Cantonese monolinguals to verify this hypothesis.

The other possible interpretation was that English language experience simply did not lead to better rhythmic perception. Although English has a high vocalic variability, duration is not the primary acoustic cue for English vocalic contrasts, e.g., tense vs. lax and full vs. reduced (Zhang et al., 2020). As such, language-to-music transfer was absent in the present and previous studies (Wong et al., 2012). Interestingly, Japanese listeners perceived rhythm more accurately than did Mandarin listeners (Zhang et al., 2020). The authors reasoned that Japanese had better durational sensitivities due to the presence of long and short vowel contrasts. These vowel contrasts are, however, absent in English.

FUTURE DIRECTION AND CONCLUSION

Aiming to provide a broad picture of how musicianship influenced the language effect on musical pitch perception, the present study viewed musicianship as a binary variable. In reality, musicians can be further categorized as amateur or professional musicians. Future studies may adopt a more fine-grained research design (2 language × 3 music groups) to see whether musicianship and language experience interact differently between amateur and professional musicians. Although non-verbal intelligence and age can be controlled statistically, future studies are encouraged to enhance stringency by recruiting matched subjects prior to experiment.

In conclusion, the present study identified an interactive effect of language experience and musicianship on musical pitch perception. Specifically, Cantonese language experience facilitated musical pitch perception only in the absence of long-term musical experience. With similar evidence that musicianship enhanced tone perception only in the absence of tone language experience (Cooper and Wang, 2012; Maggu et al., 2018b), a new condition "**L**ack of relevant experience" could be considered for the OPERA hypothesis. Apart from the interactive effect, the present study also found evidence of language-to-music transfer. Together with previous studies, this informs the OPERA hypothesis about bidirectionality (Wong et al., 2012; Bidelman et al., 2013; Chen et al., 2016; Zhang et al., 2020). As described previously, the OPERA hypothesis was not devised for language-to-music transfer. As such, it's current conception of Precision does not readily allow language-to-music transfer. To better account for the bidirectionality, Precision could be re-referenced on listeners (rather than domains). Clearly, this study does not speak the last word on cross-domain transfer nor music and language interaction. Future studies are needed to further inform (i) how Precision could be redefined and (ii) whether the OPERA hypothesis could evolve into the O-PEARL (Overlapping, Precision, Emotion, Attention, Repetition, *Lack*) hypothesis.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the ethical approval does not permit data sharing. Requests to access the datasets should be directed to corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Human Research Ethics Committee Education University of Hong Kong. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

WC: conceptualization, research design, data analysis, and manuscript writing and editing.

FUNDING

This study was supported by the start-up fund provided by the Faculty of Education, The University of Hong Kong.

ACKNOWLEDGMENTS

I wish to thank Shirlyn Chiu, Sze Hung Chung, and Kin Chak Lau for their assistance in data collection and entry.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.712753/full#supplementary-material

REFERENCES

Alexander, J. A., Wang, P. C. M., and Bradlow, A. R. (2005). Lexical tone perception in musicians and non-musicians. *Paper Presented 9th European Conference on Speech Communication and Technology*, Lisbon. doi: 10.21437/Interspeech.2005-271

Anwyl-Irvine, A. L., Massonnie, J., Flitton, A., Kirkham, N., and Evershed, J. K. (2020). Gorilla in our midst: an online behavioral experiment builder. *Behav. Res. Methods* 52, 388–407. doi: 10.3758/s13428-019-01237-x

Asaridou, S. S., and McQueen, J. M. (2013). Speech and music shape the listening brain: evidence for shared domain-general mechanisms. *Front. Psychol.* 4:321. doi: 10.3389/fpsyg.2013.00321

Bidelman, G. M., Gandour, J. T., and Krishnan, A. (2010). Cross-domain effects of music and language experience on the representation of pitch in the human auditory brainstem. *J. Cogn. Neurosci.* 23, 425–434. doi: 10.1162/jocn.2009.21362

Bidelman, G. M., Gandour, J. T., and Krishnan, A. (2011). Musicians and tone-language speakers share enhanced brainstem encoding but not perceptual benefits for musical pitch. *Brain Cogn.* 77, 1–10. doi: 10.1016/j.bandc.2011.07.006

Bidelman, G. M., Hutka, S., and Moreno, S. (2013). Tone language speakers and musicians share enhanced perceptual and cognitive abilities for musical pitch: evidence for bidirectionality between the domains of language and music. *PLoS One* 8:e60676. doi: 10.1371/journal.pone.0060676

Cason, N., Astesano, C., and Schön, D. (2015). Bridging music and speech rhythm: rhythmic priming and audio-motor training affect speech perception. *Acta Psychol.* 155, 43–50. doi: 10.1016/j.actpsy.2014.12.002

Chen, A., Liu, L., and Kager, R. (2016). Cross-domain correlation in pitch perception, the influence of native language. *Lang. Cogn. Neurosci.* 31, 751–760. doi: 10.1080/23273798.2016.1156715

Choi, W. (2020). The selectivity of musical advantage: musicians exhibit perceptual advantage for some but not all Cantonese tones. *Music Percept.* 37, 423–434. doi: 10.1525/mp.2020.37.5.423

Choi, W. (2021b). Towards a native OPERA hypothesis: musicianship and English stress perception. *Lang. Speech.* doi: 10.1177/00238309211049458 [Epub ahead of print].

Choi, W. (2021a). Cantonese advantage on English stress perception: constraints and neural underpinnings. *Neuropsychologia* 158:107888. doi: 10.1016/j.neuropsychologia.2021.107888

Choi, W., Tong, X., Gu, F., Tong, X., and Wong, L. (2017). On the early perceptual integrality of tones and vowels. *J. Neurolinguistics* 41, 11–23. doi: 10.1016/j.jneuroling.2016.09.003

Choi, W., Tong, X., and Samuel, A. G. (2019). Better than native: tone language experience enhances English lexical stress discrimination in Cantonese-English bilingual listeners. *Cognition* 189, 188–192. doi: 10.1016/j.cognition.2019.04.004

Chow, M. Y. (2012). *Singing the Right Tones of the Words the Principles and Poetics of Tone-melody Mapping in Cantopop*. Ph. D. dissertation. Pokfulam: The University of Hong Kong.

Cooper, A., and Wang, Y. (2012). The influence of linguistic and musical experience on Cantonese word learning. *J. Acoust. Soc. Am.* 131, 4756–4769. doi: 10.1121/1.4714355

Cutler, A. (2012). *Native Listening: Language Experience and the Recognition of Spoken Words*. Cambridge, MA: MIT Press, doi: 10.7551/mitpress/9012.001.0001

Delogu, F., Lampis, G., and Belardinelli, M. (2010). From melody to lexical tone: musical ability enhances specific aspects of foreign language perception. *Eur. J. Cogn. Psychol.* 22, 46–61. doi: 10.1080/09541440802708136

Grabe, E., and Low, E. (2002). Durational variability in speech and the rhythm class hypothesis. *Lab. Phonol.* 7, 515–546. doi: 10.1515/9783110197105.2.515

Gu, F., Zhang, C., Hu, A., and Zhao, G. (2013). Left hemisphere lateralization for lexical and acoustic pitch processing in Cantonese speakers as revealed by mismatch negativity. *Neuroimage* 83, 637–645. doi: 10.1016/j.neuroimage.2013.02.080

Gu, W., Hirose, K., and Fujisaki, H. (2007). Analysis of tones in Cantonese speech based on the command-response model. *Phonetica* 64, 29–62. doi: 10.1159/000100060

Hayes, B. (1989). The prosodic hierarchy in meter. *Phon. Phonol. Rhythm Meter* 1, 201–260. doi: 10.1016/B978-0-12-409340-9.50013-9

Holder, D. S. (1992). Electrical impedance tomography of brain function. *Brain Topogr.* 5, 87–93. doi: 10.1007/BF01129035

Jasmin, K., Sun, H., and Tierney, A. T. (2021). Effects of language experience on domain-general perceptual strategies. *Cognition* 206:104481. doi: 10.1016/j.cognition.2020.104481

Krishnan, A., Gandour, J. T., and Bidelman, G. M. (2012). Experience-dependent plasticity in pitch encoding: from brainstem to auditory cortex. *Neuroreport* 23, 498–502. doi: 10.1097/WNR.0b013e328353764d

Ladefoged, P. (1975). *A Course in Phonetics. Harcourt Brace Jovanovich*. New York, NY: Michael Rosenberg, doi: 10.1080/03740463.1982.10414901

Ladefoged, P., and Johnson, K. (2011). A Course in Phonetics, 6th Edn. Boston, MA: Wadsworth.

Law, S., Fung, R., and Kung, C. (2013). An ERP study of good production vis-a-vis poor perception of tones in Cantonese: implications for top-down speech processing. *PLoS One* 8:e54396. doi: 10.1371/journal.pone.0054396

Lee, C.-Y., and Lee, Y.-F. (2010). Perception of musical pitch and lexical tones by Mandarin-speaking musicians. *J. Acoust. Soc. Am.* 127, 481–490. doi: 10.1121/1.3266683

Lee, C.-Y., Lee, Y.-F., and Shr, C.-L. (2011). Perception of musical and lexical tones by Taiwanese-speaking musicians. *J. Acoust. Soc. Am.* 130, 526–535. doi: 10.1121/1.3596473

Lee, C.-Y., Lekich, A., and Zhang, Y. (2014). Perception of pitch height in lexical and musical tones by English-speaking musicians and nonmusiciansa. *J. Acoust. Soc. Am.* 135, 1607–1615. doi: 10.1121/1.4864473

Maggu, A. R., Wong, P. C., Liu, H., and Wong, F. C. (2018b). Experience-dependent influence of music and language on lexical pitch learning Is not additive. *Proc. Int. Speech* 2018, 3791–3794. doi: 10.21437/Interspeech.2018-2104

Maggu, A. R., Wong, P. C., Antoniou, M., Bones, O., Liu, H., and Wong, F. C. (2018a). Effects of combination of linguistic and musical pitch experience on subcortical pitch encoding. *J. Neurolinguist.* 47, 145–155. doi: 10.1016/j.jneuroling.2018.05.003

Magne, C., Jordan, D. K., and Gordon, R. L. (2016). Speech rhythm sensitivity and musical aptitude: ERPs and individual differences. *Brain Lang.* 15, 13–19. doi: 10.1016/j.bandl.2016.01.001

Marie, C., Magne, C., and Besson, M. (2011). Musicians and the metric structure of words. *J. Cogn. Neurosci.* 23, 294–305. doi: 10.1162/jocn.2010.21413

Nespor, M., Shukla, M., and Mehler, J. (2011). "Stress-timed vs. syllable-timed languages," in *The Blackwell Companion to Phonology*, eds M. Oostendorp, C. J. Ewen, E. Hume, and K. Rice (Hoboken, NJ: John Wiley & Sons). doi: 10.1002/9781444335262.wbctp0048

Nespor, M., and Vogel, I. (1986). *Prosodic Phonology*. Dordrecht: Foris.

Ngo, M. K., Vu, K. P., and Strybel, T. Z. (2016). Effects of music and tonal language experience on relative pitch performance. *Am. J. Psychol.* 129, 125–134. doi: 10.5406/amerjpsyc.129.2.0125

Otake, T., Hatano, G., Cutler, A., and Mehler, J. (1993). Mora or syllable? speech segmentation in Japanese. *J. Mem. Lang.* 32, 258–278. doi: 10.1006/jmla.1993.1014

Patel, A. (2011). Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. *Front. Psychol.* 2:142. doi: 10.3389/fpsyg.2011.00142

Patel, A. (2012). The OPERA hypothesis: assumptions and clarifications. *Ann. N. Y. Acad. Sci.* 1252, 124–128. doi: 10.1111/j.1749-6632.2011.06426.x

Patel, A. (2014). The evolutionary biology of musical rhythm: was Darwin wrong? *PLoS Biol.* 12:e1001821. doi: 10.1371/journal.pbio.1001821

Patel, A., Xu, Y., and Wang, B. (2010). The role of F0 variation in the intelligibility of Mandarin sentences. *Proceedings Speech Prosody* 2010, 11–14.

Peretz, I., Gosselin, N., Tillmann, B., Cuddy, L. L., Gagnon, B., Trimmer, C. G., et al. (2008). On-line identification of congenital amusia. *Music Percept.* 25, 331–343. doi: 10.1525/mp.2008.25.4.331

Peretz, I., and Hyde, K. L. (2003). What is specific to music processing? Insights from congenital amusia. *Trends Cogn. Sci.* 7, 362–367. doi: 10.1016/s1364-6613(03)00150-5

Pfordresher, P. Q., and Brown, S. (2009). Enhanced production and perception of musical pitch in tone language speakers. *Atten. Percept. Psychophys.* 71, 1385–1398. doi: 10.3758/APP.71.6.1385

Pike, K. L. (1945). *The Intonation of American English*. Ann Arbor, CA: University of Michigan Press .

Raven, J., Rust, J., Chan, F., and Zhou, X. (2018). *Raven's 2 Progressive matrices, Clinical Edition (Raven's 2)*. London: Pearson.

Rogalsky, C., Rong, F., Saberi, K., and Hickok, G. (2011). Functional anatomy of language and music perception: temporal and structural factors investigated using functional magnetic resonance imaging. *J. Neurosci.* 31, 3843–3852. doi: 10.1523/JNEUROSCI.4515-10.2011

Roncaglia-Denissen, M. P., Schmidt-Kassow, M., Heine, A., Vuust, P., and Kotz, S. A. (2013). Enhanced musical rhythmic perception in Turkish early and late learners of German. *Front. Psychol.* 4:645. doi: 10.3389/fpsyg.2013.00645

Tang, W., Xiong, W., Zhang, Y. X., Dong, Q., and Nan, Y. (2016). Musical experience facilitates lexical tone processing among Mandarin speakers: behavioral and neural evidence. *Neuropsychologia* 91, 247–253. doi: 10.1016/j.neuropsychologia.2016.08.003

Tervaniemi, M., Szameitat, Kruck, S., Schroger, E., Alter, K., De Baene, W., et al. (2006). From air oscillations to music and speech: functional magnetic resonance imaging evidence for fine-tuned neural networks in audition. *J. Neurosci.* 26, 8647–8652. doi: 10.1523/JNEUROSCI.0995-06.2006

Tsantani, M., and Cook, R. (2020). Normal recognition of famous voices in developmental prosopagnosia. *Sci. Rep.* 10:19757. doi: 10.1038/s41598-020-76819-3

Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., and Vuust, P. (2010). The musical ear test, a new reliable test for measuring musical competence. *Learn. Individ. Differ.* 20, 188–196. doi: 10.1016/j.lindif.2010.02.004

Warner, N., and Arai, T. (2001). The role of the mora in the timing of spontaneous Japanese speech. *J. Acoust. Soc. Am.* 109, 1144–1156. doi: 10.1121/1.1344156

Wong, P. C., Skoe, E., Russo, N. M., Dees, T., and Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nat. Neurosci.* 10, 420–422. doi: 10.1038/nn1872

Wong, P. C. M., Ciocca, V., Chan, A. H. D., Ha, L. Y. Y., Tan, L. H., and Peretz, I. (2012). Effects of culture on musical pitch perception. *PLoS One* 7:e33424. doi: 10.1371/journal.pone.0033424

Wong, P. C. M., and Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Appl. Psycholinguist.* 28, 565–585. doi: 10.1017/S0142716407070312

Yates, K. M., Moore, D. R., Amitay, S., and Barry, J. G. (2019). Sensitivity to melody, rhythm, and beat in supporting speech-in-noise perception in young adults. *Ear Hear.* 40, 358–367. doi: 10.1097/AUD.0000000000000621

Yip, M. (2002). *Tone*. Cambridge: Cambridge University Press.

Yiu, S. S. Y. (2013). "Cantonese tones and musical intervals," in *Proceedings of the International Conference on Phonetics of the Languages in China 2013 (ICPLC 2013)*, ed. W. S. Lee (Hong Kong: the Organizers of ICPLC 2013 at the Department of Chinese), 155–158.

Yu, L., and Zhang, Y. (2018). Testing native language neural commitment at the brainstem level: a cross-linguistic investigation of the association between frequency-following response and speech perception. *Neuropsychologia* 109, 140–148. doi: 10.1016/j.neuropsychologia.2017.12.022

Zatorre, R. J., and Baum, S. R. (2012). Musical melody and speech intonation: singing a different tune. *PLoS Biol.* 10:e1001372. doi: 10.1371/journal.pbio.1001372

Zhang, L., Xie, S., Li, Y., Shu, H., and Zhang, Y. (2020). Perception of musical melody and rhythm as influenced by native language experience. *J. Acoust. Soc. Am.* 147, EL385–EL390. doi: 10.1121/10.0001179

Zheng, Y., and Samuel, A. G. (2018). The effects of ethnicity, musicianship, and tone language experience on pitch perception. *Q. J. Exp. Psychol.* 71, 2627–2642. doi: 10.1177/1747021818757435

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

ORIGINAL RESEARCH
published: 14 October 2021
doi: 10.3389/fpsyg.2021.713949

Categorical Perception of Mandarin Pitch Directions by Cantonese-Speaking Musicians and Non-musicians

Si Chen[1,2]*, *Yike Yang*[1] and *Ratree Wayland*[3]

[1] Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, SAR China, [2] Hong Kong Polytechnic University-Peking University Research Centre on Chinese Linguistics, Hong Kong, SAR China, [3] Department of Linguistics, University of Florida, Gainesville, FL, United States

Purpose: This study is to investigate whether Cantonese-speaking musicians may show stronger CP than Cantonese-speaking non-musicians in perceiving pitch directions generated based on Mandarin tones. It also aims to examine whether musicians may be more effective in processing stimuli and more sensitive to subtle differences caused by vowel quality.

Methods: Cantonese-speaking musicians and non-musicians performed a categorical identification and a discrimination task on rising and falling continua of fundamental frequency generated based on Mandarin level, rising and falling tones on two vowels with nine duration values.

Results: Cantonese-speaking musicians exhibited a stronger categorical perception (CP) of pitch contours than non-musicians based on the identification and discrimination tasks. Compared to non-musicians, musicians were also more sensitive to the change of stimulus duration and to the intrinsic F_0 in pitch perception in pitch processing.

Conclusion: The CP was strengthened due to musical experience and musicians benefited more from increased stimulus duration and were more efficient in pitch processing. Musicians might be able to better use the extra time to form an auditory representation with more acoustic details. Even with more efficiency in pitch processing, musicians' ability to detect subtle pitch changes caused by intrinsic F_0 was not undermined, which is likely due to their superior ability to process temporal information. These results thus suggest musicians may have a great advantage in learning tones of a second language.

Keywords: categorical perception, musical experience, intrinsic F_0, stimulus duration, tone

OPEN ACCESS

Edited by:
Chao-Yang Lee,
Ohio University, United States

Reviewed by:
Chang Liu,
University of Texas at Austin,
United States
Linjun Zhang,
Beijing Language and Culture
University, China

****Correspondence:***
Si Chen
sarah.chen@polyu.edu.hk

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

Received: *24 May 2021*
Accepted: *15 September 2021*
Published: *14 October 2021*

Citation:
Chen S, Yang Y and Wayland R (2021)
Categorical Perception of Mandarin
Pitch Directions by
Cantonese-Speaking Musicians and
Non-musicians.
Front. Psychol. 12:713949.
doi: 10.3389/fpsyg.2021.713949

INTRODUCTION

Pitch, the perceptual correlate of fundamental frequency (F_0), plays an important role in both music and language. Musicians and tone language speakers have in common enhanced sensitivity to small pitch changes associated with meaningful units, namely melodies for musicians and words for speakers of a tone language. While the effects of musicianship and experience with lexical tones have been separately studied among native and non-native speakers of tone languages, respectively, the effects of musical experience on lexical tone perception among speakers of lexical tone languages have been rarely investigated. To fill this research gap, the current study investigates the effects of musical training and categorical perception (CP) of tones among tonal speakers. Potential factors such as stimulus duration and vowel quality are also explored.

This study aims to better understand the relationship between musical experience and linguistic processing by comparing native Cantonese musicians and non-musicians on their categorization of falling and rising pitch continua, representative of Mandarin tones. The continua are realized on a high [i] and a low vowel [a], associated with a high and a low intrinsic F_0, respectively (Whalen and Levitt, 1995). Unlike Mandarin, Cantonese contrasts six lexical tone categories in open syllables (Xu and Mok, 2012). In addition, while the four Mandarin tones contrast with each other in terms of direction of F_0 movement (level, falling, and rising), three of the six tones in Cantonese tones differ in pitch height. Therefore, Cantonese listeners place more emphasis on average height of F_0 contour than do Mandarin listeners when processing tones (Peng et al., 2012). This may explain why, besides Mandarin tone 2 (high-rising) vs. tone 3 (low-falling-rising), Cantonese speakers have additional difficulties differentiating Mandarin tone 1 (high-level) form Mandarin tone 4 (high-falling) (Hao, 2012). However, due to their denser tonal system, native Cantonese listeners may be more efficient in processing pitch variation than native Mandarin listeners (Lee et al., 1996; Zheng et al., 2012).

Relationship Between Music and Language

Previous behavioral research has found that musical experience improves lexical tone perception among non-native tone speakers. For example, English-speaking musicians were significantly more skilled at identifying or discriminating Mandarin tones (Alexander et al., 2005; Lee and Hung, 2008) than English-speaking non-musicians, even when only partial F_0 information was present (Lee and Hung, 2008). Additionally, novice English musicians were more accurate and faster than non-musicians in their discrimination of Thai tones (Burnham et al., 2015).

Neurophysiological data has also revealed the facilitative effects of musicality on lexical tone processing. For instance, enhanced event-related brain potentials (ERPs) associated with Mandarin tone deviants were observed in French-speaking musicians compared to non-musicians (Marie et al., 2012). Furthermore, correlations between F_0 tracking quality and the amount of musical training and performance on identification and discrimination of Mandarin syllables were positive (Wong et al., 2007). Similarly, brainstem frequency following response (FFR) to homologs of musical intervals and of lexical tones showed that pitch tracking and pitch strength were more robust for English musicians compared to English non-musicians (Bidelman et al., 2011). The above findings indicated the overlap in music and language perception, suggestive of a common perceptual substrate for the two domains (Maggu et al., 2018). To account for how musical experience may benefit speech processing, Patel (2011, 2012), and Patel (2014) proposed a neurocognitive model, OPERA, which stands for Overlap, Precision, Emotion, Repetition, and Attention. According to this model, speech processing benefits in musicians are attributed to overlapping brain networks engaged during speech and music perception.

In comparison to studies conducted on non-tone speaking musicians and non-musicians (Gottfried and Riester, 2000; Gottfried et al., 2004; Alexander et al., 2005; Gottfried, 2007; Wong and Perrachione, 2007; Lee and Hung, 2008) which largely revealed an overlap between music and language perception, the interaction between music and language among native tone speakers is not conclusive. For example, Tang et al. (2016) found increased neural activity in the discrimination of Mandarin tones (Tone 1 and Tone 2) and musical notes (C4 and G3 in piano timbre) among Mandarin-speaking musicians compared to Mandarin-speaking non-musicians. Similarly, Ong et al. (2020) found that Cantonese musicians were more accurate than non-musicians in their ability to discriminate and identify the most challenging tone pair (T23–T25). On the other hand, Thai musicians in Cooper and Wang (2012) were not superior to Thai non-musicians (or English-speaking musicians) in their ability to associate a change in pitch to a change in lexical meaning. Mok and Zuo (2012) found that, while French and English-speaking musicians outperformed non-musicians in their ability to identify Cantonese tones and their pure tone analogs, such musical advantage was not observed among native Cantonese listeners, suggesting that the linguistic and musical processing may belong to separate but overlapping domains, at least among native tone speakers.

In sum, behavioral and electrophysical studies have found that musical experience facilitates pitch perception among non-tone musicians. However, the facilitative effects of musical experience on pitch perception among native speakers of lexical tones remain inconclusive.

Musical Experience and Categorical Perception of Lexical Tones

Categorical perception, a classic paradigm in speech perception, refers to the ability to perceive and group linguistically distinct categories with equal physical changes along a continuum (Liberman et al., 1957). The results of prior research suggested that tonal continua were largely perceived categorically by native tone listeners (Abramson, 1975; Wang, 1976; Francis et al., 2003), but continuously by non-native tone listeners (e.g., Hallé et al., 2004; Peng et al., 2010).

It has been reported that musical experience modulates CP of non-speech and speech F_0 continua (Zatorre and Halpern, 1979; Howard et al., 1992; Wu et al., 2015; Zhao and Kuhl, 2015a; Chen et al., 2020), suggesting that experience with the distinguishing of pitch categories defined along an F_0 continuum in the musical domain (i.e., musical notes) may induce CP of both linguistic and non-linguistic F_0 continua. For example, Zatorre and Halpern (1979) found that perception of a continuum of major and minor thirds whose component tones were sounded simultaneously was more categorical among musicians than among non-musicians.

Similarly, Howard et al. (1992) found that the ability to label members of a computer-synthesized continuum of major to minor was more categorical among the most musical compared to the least and the moderate musical listeners. Wu et al. (2015) compared identification and discrimination of Mandarin Tone 1–Tone 4 continuum by Mandarin musicians and non-musicians. While the steepness and the location of the category boundary as well as the between-category discrimination were comparable between the two groups, within-category discrimination was found to be enhanced among the musicians. The results suggested that musicality refines low-level auditory perception without interfering with higher-level, categorical processing of lexical tonal contrasts in native tonal listeners. Recently, Chen et al. (2020) found that perception of level-to-rising and level-to-falling pitch continua was more categorical among English-speaking musicians than among English-speaking non-musicians. Zhu et al. (2021) examined CP of Mandarin Tone 2–Tone 4 continuum and its non-speech, pure tone analogs, and found that mean amplitude of the mismatch negativities (MMNs) elicited by within-category deviants was significantly larger among amateur musicians than non-musicians for both types of continua. This result suggests that musical advantage extends to auditory processing of pitch at the pre-attentive level and is not confined to professional musicians. According to Bidelman (2017), musical training may improve the abilities of representing auditory objects and matching incoming sounds to memory templates, and these improved abilities may provide musicians with advantages in CP of speech.

However, musical training does not always promote or enhance CP among either non-native or native speakers of lexical tones. For example, Zhao and Kuhl (2015a) compared the perception of Mandarin Tone 2-Tone 3 continuum among English-speaking musicians, English-speaking non-musicians, and Mandarin-speaking non-musicians and found that in contrast to native Mandarin non-musicians, English-speaking musicians and non-musicians perceived the continuum non-categorically, and while short-term perceptual training improved perception, no evidence of categorical formation among either musicians or non-musicians post training. According to Bidelman et al. (2011), “Pitch encoding from one domain of expertise may transfer to another as long as the latter exhibits acoustic features overlapping those with which individuals have been exposed to from long-term experience or training” (p. 432). Zhao and Kuhl (2015b) compared musical pitch and lexical tone discrimination among Mandarin musicians, Mandarin non-musicians, and English musicians. No difference between Mandarin musicians and non-musicians was found in their sensitivity to lexical tones or in the pattern of within-pair sensitivity to the tone pairs. The English musicians showed significantly higher overall sensitivity to lexical tones than the two Mandarin groups and exhibited a different pattern of within-pair sensitivity, indicating that the processing of musical pitch and lexical pitch might be independent in nature. In addition, Chen et al. (2020) reported that Mandarin musicians did not consistently perceive rising and falling pitch directions more categorically than Mandarin non-musicians. A plausible explanation is that perception of music and speech may implicate distinct processing mechanisms, and the processing of lexical tones makes use of other phonetic cues (e.g., duration and amplitude) besides F_0 (Liu and Samuel, 2004; Lee and Lee, 2010). Maggu et al. (2018)’s finding that Cantonese musicians did not differ from non-musicians on the brainstem encoding of lexical tones, but that they showed a more robust brainstem encoding of musical pitch as compared to non-musicians lends further support to the hypothesis that distinct mechanisms are engaged in the encoding of linguistic and musical pitch among native tone speakers.

To further probe the interaction between musical and linguistic training on pitch perception, we compared CP of Mandarin tones among Cantonese musicians and non-musicians.

The Role of Stimuli Duration and Vowel Quality in Perception of Tones

It has been found that perception of shorter vowels is more categorical than perception of longer vowels, suggesting the role of stimulus duration in CP. Duration is purported as one possible acoustic dimension affecting representation strength. According to the cue-duration hypothesis, acoustic information of consonants (e.g., formant transitions) is relatively short and less represented while information of formants in vowels is longer and better represented in auditory memory (Fujisaki and Kawashima, 1970). Moreover, the interaction between tones and the perceived vowel duration has been reported in several studies (Yu et al., 2014; Wang et al., 2017). For example, Yu et al. (2014) argues that perceived duration may be affected by F_0 slope and height. Usually, syllables with dynamic F_0 tend to be perceived as longer than those with flat F_0. These results suggest an interaction between perceived duration and tonal shapes. Duration of stimulus also plays a critical role in pitch contour perception. It has been reported that for both native Mandarin and English speakers, the strength of CP increased as stimulus duration increased. In addition, native Chinese listeners showed stronger effects from stimulus duration in terms of category boundary sharpness, between- and within- category discrimination, and peakedness compared to native English listeners (Chen et al., 2017). These results are inconsistent with the cue-duration hypothesis stating that longer stimuli will be processed with weaker CP. They also revealed an influence of tone language background to the duration effect (Chen et al., 2017).

Also, a few studies reported enhanced processing of duration, both pre-attentively and attentively among musicians. For example, Marie et al. (2012) found larger pre-attentive and attentive responses to duration deviants among native speakers of Finnish, a language with phonemic vowel length contrast, and French musicians relative to non-musicians. In another study, Chobert et al. (2014) found that both the passive and the active processing of vowel duration and voice-onset-time (VOT) deviants were enhanced in musicians compared with non-musician children. These findings suggested that linguistic and musical expertise similarly influenced the processing of pitch contour and duration in music and language, possibly because they tapped on the same pool of neural resources (Besson et al., 2011). The duration effect was mostly examined on native speakers or non-tonal speakers. Therefore, it is worth examining whether similar duration effect can be observed among musicians and non-musicians with a tone language background.

In addition to stimulus duration, intrinsic F_0 effects have been recently reported to contribute to CP (Chen et al., 2017). Intrinsic F_0 effects refer to the consistent correlation between F_0-values and vowel height (Whalen and Levitt, 1995). In speech production, high vowels are correlated with higher F_0-values and low vowels with low F_0-values. But such correlation is reversed in speech production, namely, high vowels are perceived to have a lower F_0-value when they actually share the same F_0 (Wang et al., 1976; Stoll, 1984). Yu et al. (2014) also proposed a hypothesis that if perceptual compensation occurs for high vowels, where they are perceived to have a lower F_0-value, then high vowels may in turn be perceived as longer. Chen et al. (2017) found that Mandarin and English listeners required a longer duration to perceive a tone on a low vowel than a high vowel. Chen et al. (2020) also reported that vowel quality significantly contributed to tone identification and sharpness of category boundary in English and Mandarin musicians. Vowel quality also plays a role in tone perception especially for musicians and they were better at teasing apart the factor of vowel quality that may affect the F_0 cue than non-musicians. The current study aims to examine the factor of vowel quality in tone perception by non-native tonal speakers with and without musical experience.

The Current Study

We have three research goals for the current study: (1) to investigate the effects of musical experience on CP of pitch in a non-native language by native speakers of a tone language; (2) to examine how longer stimulus duration affects auditory categorization of pitch among non-native tonal musicians and non-musicians; (3) to investigate if musicians are more sensitive to factors that may potentially influence pitch processing such as vowel quality and pitch directions than non-musicians.

METHODOLOGY

Participants

A total of 28 native speakers of Cantonese participated in the experiment. All participants speak Cantonese as the first and dominant language and they also speak English and Mandarin. Of all the participants, 14 were musicians (seven males, seven females; mean age ± SD: 23.79 ± 2.40; age range: 19–27) and the other 14 were non-musicians (seven males, seven females; mean age ± SD: 23.86 ± 2.70; age range: 19–27). The musicians began receiving formal musical training of western music instruments at an average age of 6.96 (± 1.95) years and all had regular practice of the instruments at the time of the experiment (mean years of musical experience ± SD: 16.80 ± 3.31; range: 10–23). The non-musicians did not receive any after-school musical training. To confirm that the two groups of participants had similar background of Mandarin learning, we further collected information on the onset age of Mandarin learning, years of Mandarin learning as well as self-reported proficiency of reading, writing, listening, and speaking abilities in Mandarin on a five-point scale as listed in **Table 1** (Point 1 indicates the lowest level of proficiency and point 5 indicates the highest level of proficiency). The participants reported no history of speaking, hearing, or language difficulty.

Stimuli

To examine the role of vowel quality, two sets of stimuli were created in the same way on low and high vowels [a] and [i]. A male native speaker of Mandarin with no reported speaking or hearing problem produced the Mandarin syllables [a] and [i] with the high-level Tone 1 using an Audio-Technica AT2020 microphone in a soundproof booth of the phonetics lab at the University of Florida. For each set of stimuli, the pitch contour of the original target syllable was manipulated with the pitch synchronous overlap add (PSOLA) method (Moulines and Laroche, 1995) in Praat (Boersma and Weenink, 2015). Our stimuli were all linear, and the slope and intercept parameters followed the estimates of a previous study based on a corpus of Mandarin speech (Prom-on et al., 2009). According to Prom-on et al. (2009), the slope and intercept for the rising Tone 2 in Mandarin are 93.4 and −2.2 st, respectively. Equation (1) below was used to transform st from Hertz (Hz):

$$Number\ of\ st = \frac{12}{\log_{10}2} * \log(F_{02}/F_{01}) \qquad (1)$$

where F_{01} and F_{02} represent the lower F_0 and the higher F_0, and the number of st measures the distance between F_{01} and F_{02} in Hz (Lehnert-LeHouillier, 2013). Following Xu et al. (2006), we set the F_{02}-value at 130 Hz, and calculated the F_{01}-value based on the chosen F_{02}-value and the intercept value of −2.2st.

For each vowel, nine durations were manipulated: 200, 180, 160, 140, 120, 100, 80, 60, and 40 ms. In total, there were 18 continua (9 duration ∗ 2 vowels). The stepwise onset values with different duration values for the rising tone are as shown in **Table 2** and those for the falling tone are listed in **Table 3**. **Figures 1**, **2** are examples of two rising continua with the durations of 200 and 40 ms.

Using a rising continuum as an example, we describe how steps in a continuum are calculated as follows. First, the offset

TABLE 1 | Participants' background information.

Factor	Musician	Non-musician
Age	23.79 ± 2.40	23.86 ± 2.70
Onset age of musical training	6.96 ± 1.95	N/A
Years of musical training	16.80 ± 3.31	N/A
Onset age of Mandarin learning	5.43 ± 2.26	5.79 ± 1.82
Years of Mandarin learning	7.85 ± 2.56	8.14 ± 2.20
Self-reported reading ability in Mandarin	4.57 ± 0.49	4.71 ± 0.45
Self-reported writing ability in Mandarin	4.43 ± 0.73	4.36 ± 0.48
Self-reported listening ability in Mandarin	3.64 ± 1.11	3.71 ± 0.45
Self-reported speaking ability in Mandarin	3.64 ± 0.72	3.50 ± 0.50

TABLE 2 | Onset values for each step varied by duration for linear rising pitch directions.

Duration	0.2	0.18	0.16	0.14	0.12	0.1	0.08	0.06	0.04
Step 0	196.56	187.56	178.97	170.78	162.96	155.50	148.38	141.59	135.11
Step 1	179.43	172.62	166.09	159.84	153.86	148.14	142.66	137.42	132.40
Step 2	163.12	158.31	153.69	149.26	145.02	140.94	137.04	133.30	129.72
Step 3	147.58	144.61	141.76	139.03	136.41	133.91	131.52	129.24	127.06
Step 4	132.77	131.49	130.27	129.13	128.05	127.04	126.11	125.24	124.43
Step 5	118.67	118.92	119.21	119.55	119.92	120.33	120.79	121.28	121.82
Step 6	105.23	106.89	108.57	110.28	112.01	113.78	115.57	117.39	119.24

TABLE 3 | Offset values for each step varied by duration for linear falling pitch directions.

Duration	0.2	0.18	0.16	0.14	0.12	0.1	0.08	0.06	0.04
Step 0	196.56	187.56	178.97	170.78	162.96	155.50	148.38	141.59	135.11
Step 1	179.43	172.62	166.09	159.84	153.86	148.14	142.66	137.42	132.40
Step 2	163.12	158.31	153.69	149.26	145.02	140.94	137.04	133.30	129.72
Step 3	147.58	144.61	141.76	139.03	136.41	133.91	131.52	129.24	127.06
Step 4	132.77	131.49	130.27	129.13	128.05	127.04	126.11	125.24	124.43
Step 5	118.67	118.92	119.21	119.55	119.92	120.33	120.79	121.28	121.82
Step 6	105.23	106.89	108.57	110.28	112.01	113.78	115.57	117.39	119.24

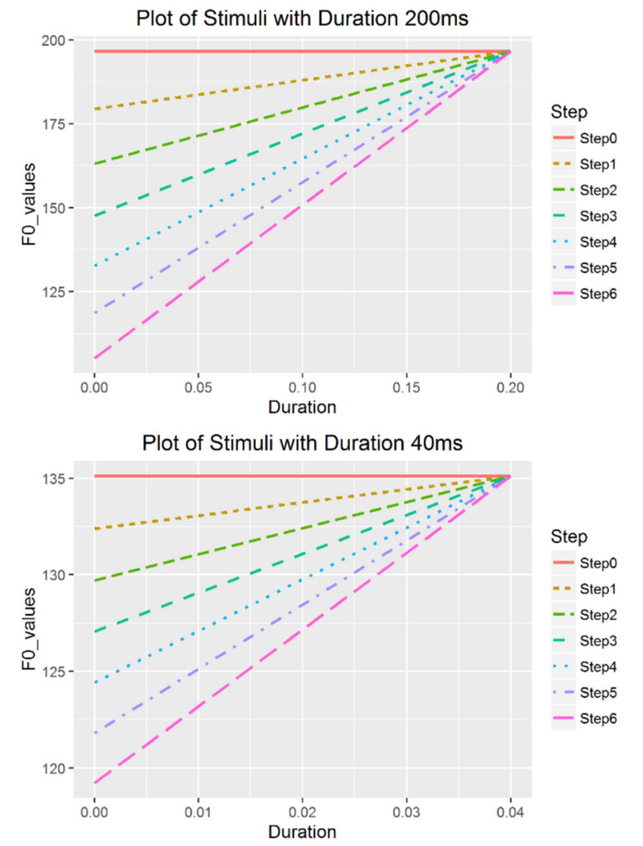


FIGURE 1 | Rising continua with duration of 200 ms (top) and 40 ms (bottom).

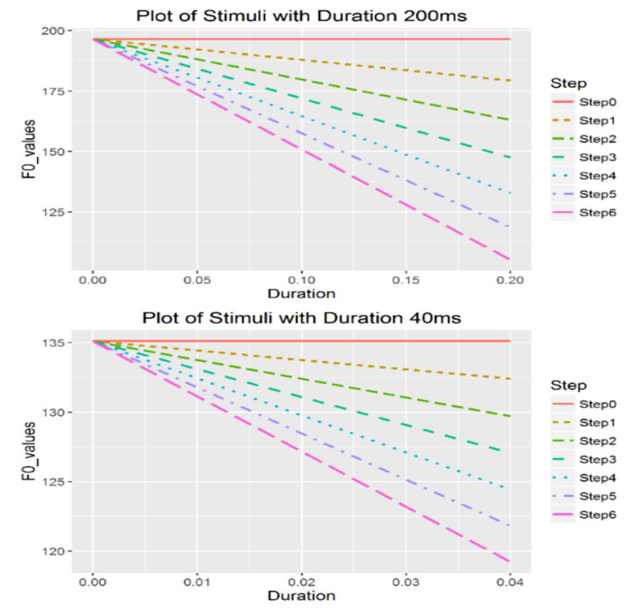


FIGURE 2 | Falling continua with duration of 200 ms (top) and 40 ms (bottom).

values for each duration were calculated with Equation (2):

$$X(t) = 93.4 * t - 2.2 \tag{2}$$

where t is the duration in s, and X(t) stands for the semitone (st)-values at the offset of the tone. For example, for the 200 ms continuum, $t = 200$ ms (0.2 s) and the onset can be calculated by setting $t = 0$ s, which is -2.2 st (or 123.02 Hz); and the offset can be calculated by setting $t = 0.2$ s, which is 16.48 st (or 196.56 Hz). Therefore, the onset-to-offset distance [i.e., $\Delta X(t)$] is 16.48 – $(-2.2) = 18.68$ st in this case. Stimuli with various onsets were created based on the calculated intercept value of -2.2 st, which was also the cutoff point. The extreme points of onset values were then determined so that the distance between the lowest onset value (-8.43 st or 105.23 Hz) and -2.2st was one third of the distance between the highest onset value (16.48 st or 196.56 Hz) and -2.2st. The highest onset was defined as the same value as the obtained offset value (196.56 Hz) and is used to generate a level tone with equal onset and offset values. After obtaining the highest and lowest onsets, steps were created between them based on the ERB scale instead of Hertz because the former reflects natural perception (Xu et al., 2006). Seven stimuli with equal perceptual distance were created for each duration value in a continuum, and the onset values were then transformed back into Hertz.

Our resynthesizing procedure is similar to Peng et al. (2010): (1) we adjusted the duration of the stimuli to the duration values in **Tables 1**, **2**; (2) we peak normalized the stimuli to the same intensity level; (3) We adjusted the pitch points according to the values in **Tables 1**, **2**. The set of stimuli used by Chen et al. (2020) was used in this current study.

For the identification task, stimuli with a rising pitch continuum were grouped in one block and those with a falling pitch continuum were grouped in another block. There were 630 stimuli (5 repetitions $*$ 7 steps $*$ 9 duration $*$ 2 syllable) in each block.

Since a one-step difference is too difficult to perceive (Francis et al., 2003), so this study used two-step difference pairs for the same-difference discrimination task. Again, the stimulus presentation was blocked by rising and falling pitch directions. Within each block, the different pairs were presented in either the forward order (0–2, 1–3, 2–4, 3–5, 4–6) or the backward order (2–0, 3–1, 4–2, 5–3, 6–4). All the same and different pairs were repeated twice in the task. Overall, there were 612 trials in each block [9 durations $*$ (7 same pairs + 10 different pairs) $*$ 2 syllables $*$ 2 repetitions].

Procedure

All the participants signed informed consent forms in compliance with a protocol approved by the Human Subjects Ethics Sub-committee at the Hong Kong Polytechnic University and participated in the experiment with a GMH C 8.100 D headset at the Speech and Language Sciences Lab of the Hong Kong Polytechnic University. An identification task and a same-difference discrimination task were implemented in E-Prime 2.0 (Schneider et al., 2012). There was one practice block and two test blocks for each task where the practice block always preceded the test block. Within each block, the order of stimulus presentation was randomized. The order of block presentation for each participant was also counterbalanced and the block order was kept identical for the musician group and the non-musician group.

Training and Practice

Prior to the actual tasks, the participants were first familiarized with the tasks and the participants were required to identify the pitch directions they heard by pressing the number key 1 for a level tone and the number key 2 for a rising or falling tone in a practice session of the identification task. In the practice session for the same-difference discrimination task, the participants listened to a pair of stimuli at a time and were asked to decide whether the stimuli had the same or different pitch direction by pressing the number key 1 for "the same" and the number key 2 for "different." Only stimuli with the longest duration value (0.2 s) at the two endpoints of each continuum were used for the practice sessions. A minimum threshold [75% (12 out of 16 trials); 71% (17 out of 24 trials) for the identification and the discrimination tasks, respectively] of correct responses was set for each practice session to make sure that the participants were able to finish the tasks and can identify or discriminate the pitch directions with either a level or steep rising/falling tone with the longest duration. Only those who passed the threshold proceeded to the experimental tasks. The reasons for including the practice session is that we need to make sure that the participants are familiar with the procedure required by the tasks and are able to press the right keys when listening to the pitch directions.

The Identification Task

In the identification task, the participants followed the same procedure as in the practice session. Once they heard a stimulus, they needed to decide whether it was a level tone or a rising/falling tone by pressing "1" for the former and "2" for the latter at a self-paced rate. The next stimulus was presented automatically after a response was given.

The Same-Different Discrimination Task

In the discrimination task, the participants listened to a pair of stimuli in each trial and were asked to decide whether the two stimuli were the same or different in terms of pitch direction at a self-paced rate. There were 34 trials of stimuli for each duration value [(7 same pairs + 10 different pairs) $*$ 2 repetitions]. The 34 trials were divided into five two-step comparison units: 0–2, 1–3, 2–4, 3–5, and 4–6, where each unit had four types of comparisons: AA, AB, BA, and BB. The same trials were included. For example, the 3–3 pair was included in both 1–3 and 3–5 units. Therefore, there were eight trials (4 types of comparison $*$ 2 repetitions) for each unit. In addition, the ISI of 500 ms was used as previous research suggests that 500 ms is the time needed to maximize the differences in between- vs. within-category discrimination (Pisoni, 1973; Xu et al., 2006; Peng et al., 2010).

The stimuli were presented automatically after a response was given. For further analyses, d-prime (d') scores were computed from raw discrimination responses with the Equation (3):

$$d' = z(H) - z(F) \tag{3}$$

where d' is the d-prime score, H is the hit rate (i.e., "different" response given for "different" trials), F is the false alarm rate (i.e., "different" response given for "same" trials) and z is z-transform (Creelman and Macmillan, 2004).

Data Analysis

There are three features of CP, namely a sharp category boundary, a discrimination peak, and prediction of discrimination from identification. The data obtained were, therefore, analyzed to examine the effects of musical background (musicians vs. non-musicians), pitch direction (rising vs. falling), vowel quality (low vs. high), and duration (nine different duration values) on these characteristics of CP.

The Identification Task

The identification data was analyzed to see if category boundary sharpness and category boundary location were affected by pitch direction type (falling vs. rising), vowel quality (low [a] vs. high [i]), stimulus duration (nine values: 40–200 in 20 ms increment) and musicianship (musicians vs. non-musicians). To achieve these goals, a generalized linear mixed model with subjects as a random effect was fitted to the data using the lme4 package (Bates et al., 2015) in R (R Core Team, 2018).

To perform the analyses, the data were divided into eight subgroups based on musical background, pitch direction, and vowel quality: FCMA, FCMI, RCMA, RCMI, FCNA, FCNI, RCNA, and RNI, where F and R represent falling and rising pitch directions, CM and CN stand for Cantonese musicians and non-musicians, and A and I for [a] and [i] syllables. For example, RCNA is the subset of data which includes rising pitch direction (R) on the syllable [a] (A) identified by Cantonese non-musicians (CN). Within each subgroup, a generalized linear mixed model was fitted with identification scores (0 or 1) as the response variable (the accuracy rate can be calculated from 0 and 1 s) and step number ($x = 0–6$) as a factor. The model is similar to a logistic regression model when only the fixed effects are considered in Equation (4).

$$\log_e\left(\frac{p_1}{1-p_1}\right) = b_0 + b_1 x \qquad (4)$$

In this equation, the coefficient b_1 stands for the sharpness of category boundary. To perform a *post-hoc* analysis on the effects of sharpness of boundary by duration, we carried out pairwise comparisons between different duration values for each of the eight subgroups. Specifically, we fit a model treating the coefficient b_1 from each pair of duration values as the same and conducted a likelihood ratio test to compare it to another model that treats them as different. Significant differences between the two models would indicate significant differences of the coefficient b_1 between two duration values. We also tested whether musical background, pitch direction and vowel quality would influence the values of b_1. In addition, we modeled the relationship between the sharpness of category boundary and duration for musicians and non-musicians.

For category boundary location, after obtaining the estimates for b_0 and b_1, $P_1 = 0.5$ was used to estimate the step number at category boundary within musician and non-musician groups, as shown in Equation (5):

$$x_{cb} = -b_0/b_1 \qquad (5)$$

Once the individual category boundary was obtained for subjects in each subgroup, a linear mixed effects model was fitted, with pitch direction, musicianship, duration, and vowel quality as fixed effects and subjects as random effects, followed by *post-hoc* analyses. Linear regression models were fitted separately for musicians and non-musicians to test the relationship between duration and category boundary.

Finally, formulas for minimum stimulus duration required to perceive a rising or a falling F_0 from a level F_0 were derived using Equation (6):

$$t = b_0 + b_1 d \qquad (6)$$

where t is the duration needed to perceive d st differences from level tones. For each duration value, the estimated step (identification rate equaled 0.5) was recorded and set as a cut-off point. Step numbers smaller than this point indicated that the stimulus was more likely to be identified as a level tone. The step number was transformed back to st-values with respect to the baseline (level tone as step 0) for each duration value. Linear mixed effects models were then fitted to obtain a relationship between cut-off st-values and duration, and it was assumed that the cut-off st-values were the smallest values for a rising or falling pitch direction to be perceived as different from a level tone. Minimum stimulus duration formulas for each of the eight subgroups were derived separately, and more general formulas for musicians and non-musicians to perceive each pitch direction were also obtained.

The Discrimination Task

d-prime's scores based on correct (hits) and incorrect (false alarms) responses were obtained in the discrimination task, and a generalized linear mixed model were fitted with duration, musical background, and vowel quality and all the two-way and three-way interactions as fixed factors and subject as a random factor.

To explore the relationship between between-and within-category discrimination, *d*-prime scores of between- and within-category discrimination for each subgroup according to its category boundary was calculated (Wu et al., 2015). The *d*-prime scores were calculated for all nine duration values. When the category boundary was <1 or >5, the *d*-prime score was not calculated, because the steps were constrained between 0 and 6 steps. Linear mixed-effects models were then fitted to examine the contribution of musical background, pitch direction, vowel quality, and duration to the *d*-prime scores, followed by *post-hoc* analyses. Pairwise comparison of duration was performed, and linear regression models were fitted to examine the relationship between discrimination scores and duration.

The peakedness of discrimination function was estimated from the difference between P_{bc} and P_{wc}. P_{bc} (between-category discrimination) is defined as P of the comparison unit corresponding to the category boundary, and P_{wc} (within-category discrimination) is defined as the average of two comparison units at the extremes of the continuum (P_{02} and P_{46}) (Pisoni, 1973). Linear models were fitted to examine whether there were any significant contributing factors. Pairwise comparison of duration with respect to peakedness was conducted, and regression models were also fitted to investigate the relationship between peakedness and duration.

Predicting Discrimination From Identification

According to Pollack and Pisoni (1971), the predicted discrimination score $P*$ can be calculated by Equation (7):

$$P^* = \left[1 + (P_A - P_B)^2\right]/2 \qquad (7)$$

where P_A and P_B are the identification scores in a comparison unit. This equation assumes that the discrimination can be solely determined by the identification of the two stimuli A and B. The correlation between the predicted and the observed discrimination scores for each comparison subgroup was calculated based on different trials of stimuli A and B by optimizing linear regression models after Fisher's z transformation. Next, the effects of musical background, pitch direction, vowel quality, and duration on the correlation was tested. The mean difference between the predicted and observed discrimination scores P – P^* were also calculated. The optimized linear model was selected based on the stepwise optimization algorithm using the function "step" (Hastie and Pregibon, 1992) to model the relationship between the distance and the tested variables.

RESULTS

Identification

A generalized linear mixed effects model was fitted to the pitch direction (i.e., level vs. rising/falling) identification data obtained from Cantonese listeners with and without musical experience. The dependent variable was response rate, and the independent variables included vowel quality, pitch direction, duration, musical experience, and their interactions. The results revealed significant main effects of vowel quality [$\chi^2_{(1)} = 26.503$, $p < 0.001$] and pitch direction [$\chi^2_{(1)} = 104.9$, $p < 0.001$] as well as a marginally significant main effect of duration [$\chi^2_{(1)} = 3.285$, $p = 0.07$]. The main effect of musical experience, however, did not reach significance [$\chi^2_{(1)} = 1.017$, $p = 0.313$]. These results indicated that a syllable with the high vowel [i] was more likely to be identified as a level tone than a syllable with the low vowel [a], and a rising syllable was more likely to be identified as a level than a falling syllable. The chance of a syllable being identified as a level tone also decreased with the increase of syllable duration. The two-way interactions were significant for vowel quality and musical experience [$\chi^2_{(1)} = 9.356$, $p = 0.002$], musical experience and pitch direction [$\chi^2_{(1)} = 21.465$, $p < 0.001$], and musical experience and duration [$\chi^2_{(8)} = 169.99$, $p < 0.001$]. *Post-hoc* tests suggested that (1) vowel quality did not influence the identification of non-musicians, but musicians were more likely to identify a high vowel /i/ as a level tone; (2) both groups of participants were more likely to identify a falling syllable as a level tone; (3) with the increase of duration, musicians tended to identify a syllable as a rising/falling tone, while non-musicians tended to identify a syllable as a level tone.

Sharpness of Category Boundary

The estimates of the coefficient b_1 (sharpness of category boundary) of all nine stimulus durations for each subgroup are plotted in **Figure 3**. From **Figure 3**, it is evident that category boundary becomes sharper as the stimulus duration increases for all subgroups. In addition, overall, the musicians showed a sharper category boundary than the non-musicians.

We extracted all the values of category boundary sharpness b_1 and treat *it* as the dependent variable. The independent variables included vowel quality, pitch direction, duration, musical experience, and their interactions. Likelihood ratio tests revealed significant effects of musical experience [$\chi^2_{(1)} = 97.687$, $p < 0.001$]; pitch direction [$\chi^2_{(1)} = 26.733$, $p < 0.001$]; and vowel quality [$\chi^2_{(1)} = 26.503$, $p < 0.001$]. These results suggested that, on average, category boundary was significantly sharper for musicians than non-musicians [$b_1 = 0.587$ vs. 0.363]; for the rising pitch than the falling pitch direction [$b_1 = 0.536$ vs. 0.414], and for the high vowel /i/ than the low vowel /a/ [$b_1 = 0.483$ vs. 0.483]. There were also significant interactions between musical experience and pitch direction [$\chi^2_{(3)} = 150.47, p < 0.001$], between musical experience and vowel quality [$\chi^2_{(3)} = 110.48$, $p < 0.001$], and between musical experience and duration [$\chi^2_{(10)} = 3686.4$, $p < 0.001$]. For both groups of listeners, follow up tests revealed higher b_1 (sharper category boundary) values for the rising than the falling pitch directions and for the high vowel [i] than the low vowel [a]. However, none of the differences reached statistical significance.

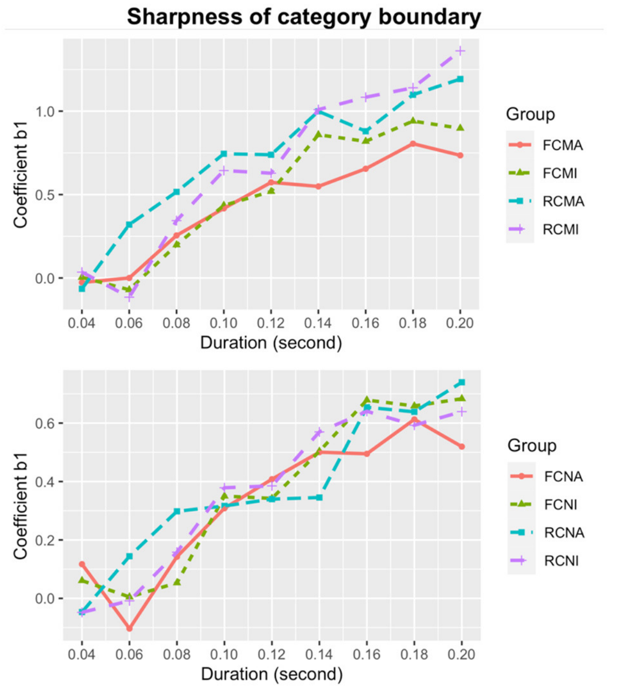


FIGURE 3 | Sharpness of category boundary for Cantonese musicians (upper) and non-musicians (lower).

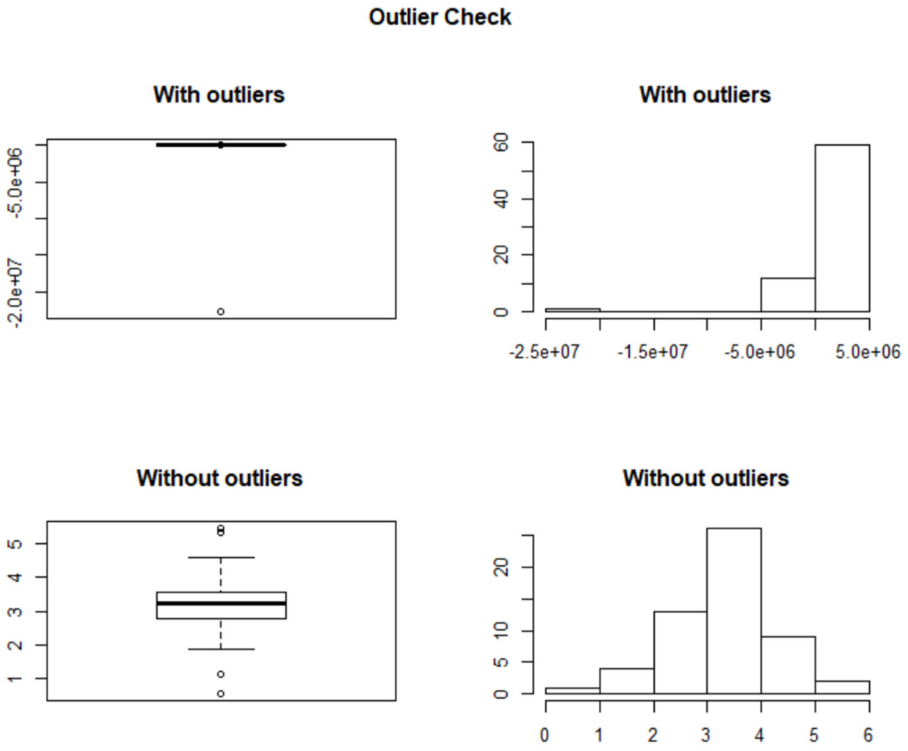


FIGURE 4 | Plots with and without outliers. The upper figures present the original data, which are severely left-skewed. The right figures represent the data after removing the outliers, which are normally distributed.

To examine the significant interaction between musicianship and stimulus duration on category boundary sharpness, regression models were separately fitted for musicians and non-musicians. For the musicians, a regression model with an extra quadratic term did not significantly improve the model compared to a linear regression model, as suggested by a likelihood ratio test [$\chi^2_{(2)} = 0.187$, $p = 0.911$]. The linear regression model in Equation (8) was thus selected and it captured the relationship between the duration and sharpness well [$F_{(1,34)} = 153.1$, $p < 0.001$; adjusted $R^2 = 0.813$]. For the non-musicians, a regression model with an extra quadratic term did not significantly differ from a linear regression model [$\chi^2_{(2)} = 2.214$, $p = 0.331$], so the linear model in Equation (9) was adopted [$F_{(1,34)} = 251.7$, $p < 0.001$; adjusted $R^2 = 0.878$].

$$b1 = 7.067 * d - 0.261 \tag{8}$$

$$b1 = 4.498 * d - 0.177 \tag{9}$$

The formulae show that the sharpness of category boundary for both Cantonese musicians and non-musicians increases with the stimulus duration, but the musicians exhibit a steeper slope and thus have a faster increment of sharpness of category boundary with increased duration.

Category Boundary Location

Category boundary for each subgroup and each duration value was calculated based on the estimated coefficients b_0 and b_1 from the generalized linear models. After removing all the outliers[1] as illustrated in **Figure 4**, we plotted the category boundary against stimulus duration for each subgroup in **Figure 5**.

In general, we see that category boundary shifted to smaller values as the stimulus duration increased. A linear regression model was fitted, where the dependent variable was category boundary, and the independent variables included musicianship, vowel quality, duration, and pitch direction. A significant effect of stimulus duration ($p < 0.001$) was found, but a marginal effect of musical experience was identified ($p = 0.068$), wherein boundary locations were smaller for non-musicians than musicians. No effects of vowel quality ($p = 0.143$) or pitch direction ($p = 0.846$) were found.

To capture the relationship between category boundary location and stimulus duration for Cantonese musicians and non-musicians, regression models were fitted to the data. For the musicians, no significant difference was observed between a regression model with only the slope and intercept terms and a regression model with an extra quadratic term, as suggested by a likelihood ratio test [$\chi^2_{(1)} = 1.793$, $p = 0.181$], so the simple model [$F_{(1,24)} = 33.77$, $p < 0.001$; adjusted $R^2 = 0.567$] was adopted [Equation (10)]. Similarly, for the Cantonese non-musicians, no significant difference was found between a simple

[1] Following the Tukey's method, outliers are defined as values above and below the $1.5 *$ IQR (interquartile). The script for detecting and removing the outliers can be found at: https://goo.gl/4mthoF.

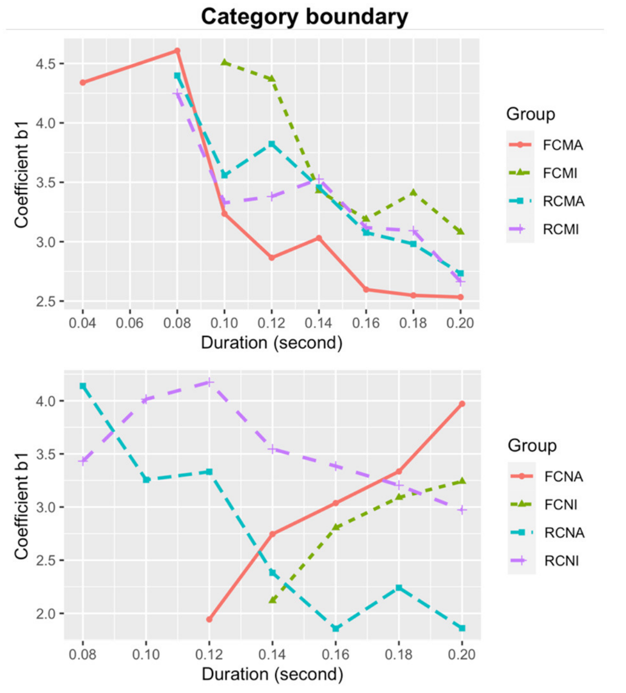


FIGURE 5 | Category boundary for Cantonese musicians and non-musicians. The upper figure shows the category boundary of the musician group and the lower one stands for that of the non-musician group.

model and a complex model [$\chi^2_{(1)} = 2.328$, $p = 0.127$], so the simple model [$F_{(1,19)} = 0.495$, $P = 0.490$; adjusted $R^2 = -0.026$] was adopted [Equation (11)].

$$cb = -12.008 * d + 5.068 \tag{10}$$

$$cb = -3.409 * d + 3.509 \tag{11}$$

We can infer from the equations that the category boundary shifted to smaller values as the stimulus duration increases, and the decrease rate is faster for the musician group.

Same-Different Discrimination Task

The raw data of the same-different discrimination task were transformed into a response of either 0 (same) or 1 (different) to fit generalized linear mixed effects models to test the main effects and interaction terms. The dependent variable was the response variable, and the independent variables included musicianship, vowel quality, duration, and pitch direction and their interactions. The results revealed significant main effects of vowel quality [$\chi^2_{(1)} = 6.451$, $p = 0.011$], pitch direction [$\chi^2_{(1)} = 583.31$, $p < 0.001$], and duration [$\chi^2_{(1)} = 2507.2$, $p < 0.001$]. The main effect of musical experience did not reach significance [$\chi^2_{(1)} = 2.377$, $p = 0.123$]. The two-way interactions were significant for vowel quality and musical experience [$\chi^2_{(2)} = 27.002$, $p < 0.001$], vowel quality and pitch direction [$\chi^2_{(2)} = 10.417$, $p = 0.005$], vowel quality and duration [$\chi^2_{(1)} = 4.867$, $p = 0.027$], musical experience and pitch direction [$\chi^2_{(1)} = 21.098$, $p < 0.001$], and music experience and duration [$\chi^2_{(1)} = 4.079$, $p = 0.043$]. *Post-hoc* tests showed the following: (1) the musicians tended to regard the low vowel [a] stimuli as being the same compared to the high vowel [i], while the non-musicians had the reversed pattern; (2) the musicians were more likely to regard the stimuli as different than the non-musicians, regardless of duration or pitch direction; (3) the low vowel [a] stimuli were more likely to be regarded as the same compared to the high vowel [i] stimuli, regardless of the pitch direction or duration; (4) the rising pitch direction was more likely to be treated as the same compared to the falling pitch direction, regardless of duration values.

Between-Category Discrimination

For each subject, the *d*-prime scores of the comparison unit corresponding to the category boundary were calculated. Linear mixed-effects models were then fitted with *d*-prime scores as the response variable, musical experience, pitch direction, vowel quality, and duration as factors, and subjects as a random effect. The main effects of musical experience [$\chi^2_{(1)} = 7.486$, $p = 0.006$], pitch direction [$\chi^2_{(1)} = 37.065$, $p < 0.001$], and duration [$\chi^2_{(1)} = 33.938$, $p < 0.001$] all reached significance. There were also significant two-way interactions between musical experience and pitch direction [$\chi^2_{(1)} = 29.447$, $p < 0.001$], musical experience and vowel quality [$\chi^2_{(1)} = 6.495$, $p = 0.039$], pitch direction and vowel quality [$\chi^2_{(1)} = 6.800$, $p = 0.033$], and duration and vowel quality [$\chi^2_{(1)} = 7.962$, $p = 0.004$]. *Post-hoc* tests suggested that musicians had higher *d*-prime scores than non-musicians regardless of pitch direction, duration, or vowel quality. Also, the falling pitch direction had higher *d*-prime scores than the rising pitch direction for musicians, while the reversed pattern was observed for non-musicians. In general, *d*-prime scores improved with the increase of duration for both musicians and non-musicians.

Next, we fitted regression models for Cantonese musicians and non-musicians to capture the relationship between between-category *d*-prime scores (dependent variable) and duration. The model for the musicians (Equation 12) was non-significant [$F_{(1,25)} = 4.412$, $p = 0.459$; adjusted $R^2 = 0.116$], and the one for non-musicians (Equation 13) was significant [$F_{(1,18)} = 8.014$, $p = 0.011$; adjusted $R^2 = 0.270$].

$$bcd = 22.710 * d - 3.073 \tag{12}$$

$$bcd = 17.616 * d - 1.975 \tag{13}$$

Within-Category Discrimination

Within-category discrimination *d*-prime scores were calculated for each subject. A linear mixed effects model was fitted where the dependent variable was *d*-prime scores, and the independent variables included musicianship, vowel quality, duration, and pitch direction. The results showed significant main effects of pitch direction [$\chi^2_{(1)} = 140.99$, $p < 0.001$; mean d′ scores = 1.441 and 0.777 for falling and rising pitch directions, respectively], and a marginally significant main effect of stimulus duration [$\chi^2_{(1)} = 3.674$, p = 0.055; mean d′ scores ranged from 0.988 to 1.240]. There was no significant effect of musical experience on the *d*-prime scores.

We then fitted regression models for Cantonese musicians and non-musicians to capture the relationship between within-category d-prime scores (dependent variable) and duration (independent variable). The regression model for the musicians as in Equation (14) was significant [$F_{(1,25)} = 6.747$, $p = 0.015$; adjusted $R^2 = 0.181$]. The model for non-musicians as in Equation (15) did not reach significance [$F(_{1,18}) = 2.67$, $p = 0.120$; adjusted $R^2 = 0.081$]. In general, the musicians had higher scores than the non-musicians, although not all pairs showed significant differences.

$$wcd = 0.812 + 5.886 * d \tag{14}$$

$$wcd = 0.305 + 5.775 * d \tag{15}$$

Peakedness

Peakedness of discrimination function was estimated by calculating the difference between P_{bc} (between-category discrimination) and P_{wc} (within-category discrimination). The dependent variable is the peakedness value, and the independent variables included musical experience, pitch direction and stimulus duration. The main effects of musical experience [$\chi^2_{(1)} = 5.828$, $p = 0.016$] and pitch direction [$\chi^2_{(1)} = 10.434$, $p = 0.001$] and stimulus duration [$\chi^2_{(1)} = 12.412$, $p < 0.001$] reached significance. The two-way interactions between musical experience and pitch direction [$\chi^2_{(1)} = 15.469$, $p < 0.001$] and between pitch direction and duration [$\chi^2_{(1)} = 3.970$, $p = 0.046$] were also significant. In general, the musicians had larger peakedness than the non-musicians regardless of vowel quality, or duration. The rising pitch direction was always greater in peakedness than the falling pitch direction. There was also a general trend of greater peakedness with the increase of duration for both groups of participants. Moreover, although musicians had larger peakedness than non-musicians for the falling pitch direction, the two groups had comparable peakedness for the rising pitch direction.

Since musical experience was shown to influence the peakedness, we fitted regression models for musicians and non-musicians separately. However, the models suggested that duration did not significantly contribute to the change in peakedness for either musicians [$F_{(1,18)} = 3.341$, $p = 0.084$; adjusted $R^2 = 0.110$] or for non-musicians [$F_{(1,6)} = 2.43$, $p = 0.170$; adjusted $R^2 = 0.170$]. The models are listed in Equations (16) (musicians) and (17) (non-musicians).

$$pk = 0.339 - 6.170 * d \tag{16}$$

$$pk = -0.704 + 6.825 * d \tag{17}$$

Predicted and Obtained Discrimination

The scores of the identification and predicted and obtained discrimination for the eight subgroups are plotted in **Figures 6–8**. Three duration values were selected for illustration: 200, 140, and 80 ms. In general, the musicians showed stronger CP than the non-musicians. For both groups, the perception became more categorical with the increase of the stimulus duration.

To confirm these observations, linear regression models were fitted where correlation values between predicted and obtained discrimination were the dependent variable, and the independent variables included musical experience, vowel quality, duration, and pitch direction. Significant main effects of musical experience [$t_{(46)} = 2.96$, $p = 0.006$], pitch direction [$t_{(46)} = 2.11$, $p = 0.044$], and duration [$t_{(46)} = 2.46$, $p = 0.013$] as well as significant two-way interactions between musical experience and pitch direction [$t_{(46)} = -3.01$, $p = 0.005$] and between musical experience and duration [$t_{(46)} = -3.20$, $p = 0.003$] were obtained. These results indicated that, overall, correlation between predicted and obtained discrimination was stronger for non-musicians than musicians and for the rising pitch direction than the falling pitch direction. In addition, correlation for the rising pitch direction obtained from both musicians and non-musicians was stronger than that obtained for the falling pitch direction obtained from musicians [FM < RM, $\chi^2_{(1)} = 6.04$, $p = 0.014$; FM < RN, $\chi^2_{(1)} = 5.75$, $p = 0.016$] but not from non-musicians [FN < RN, $\chi^2_{(1)} = 2.35$, $p = 0.125$], suggesting that identification better predicted discrimination of the rising pitch contour than the falling pitch contour, particularly among musicians.

Next, we calculated the distances between the predicted and obtained discrimination. A linear regression model suggested a significant two-way interaction between musical experience and vowel quality [$t_{(234)} = -2.30$, $p = 0.022$] and a three-way interaction among musical experience, duration and vowel quality [$t_{(234)} = -2.09$, $p = 0.021$]. *Post-hoc* tests showed that the pairs FN vs. RN [FN > RN, $\chi^2_{(1)} = 4.146$, $p = 0.042$] and MA vs. NI [MA > NI, $\chi^2_{(1)} = 4.404$, $p = 0.036$] were significant. For the three-way interaction, we only found that the distance got larger with the increase of duration, which was true for both groups and for both vowels.

Summary of the Results

In sum, musicians' perception of pitch direction in a non-native language was generally more categorical than non-musicians. Specifically, they exhibited sharper category boundary and were also more sensitive to between-category differences than non-musicians. In addition, they had greater peakedness of the discrimination functions as well as higher correlation between predicted and obtained discrimination. Surprisingly, however, the relationship between between-category discrimination and duration was significant for non-musicians but not for musicians. In contrast, the relationship between within-category sensitivity and duration was significant for musicians, but not for non-musicians. In general, the musicians had higher scores than the non-musicians. In response to our first research goal, these results strongly suggested that musicians with a tonal language background tended to have stronger CP of pitch in a non-native language.

For our second research goal on examining the effects of stimulus duration, we found that both musicians and non-musicians benefited from increased stimulus duration, but musicians were more sensitive than non-musicians to the changes in stimulus duration reflected by more changes in the sharpness of categoryand the location of category boundary.

Finally, we explored if musicians were more sensitive to factors that may potentially influence pitch processing such

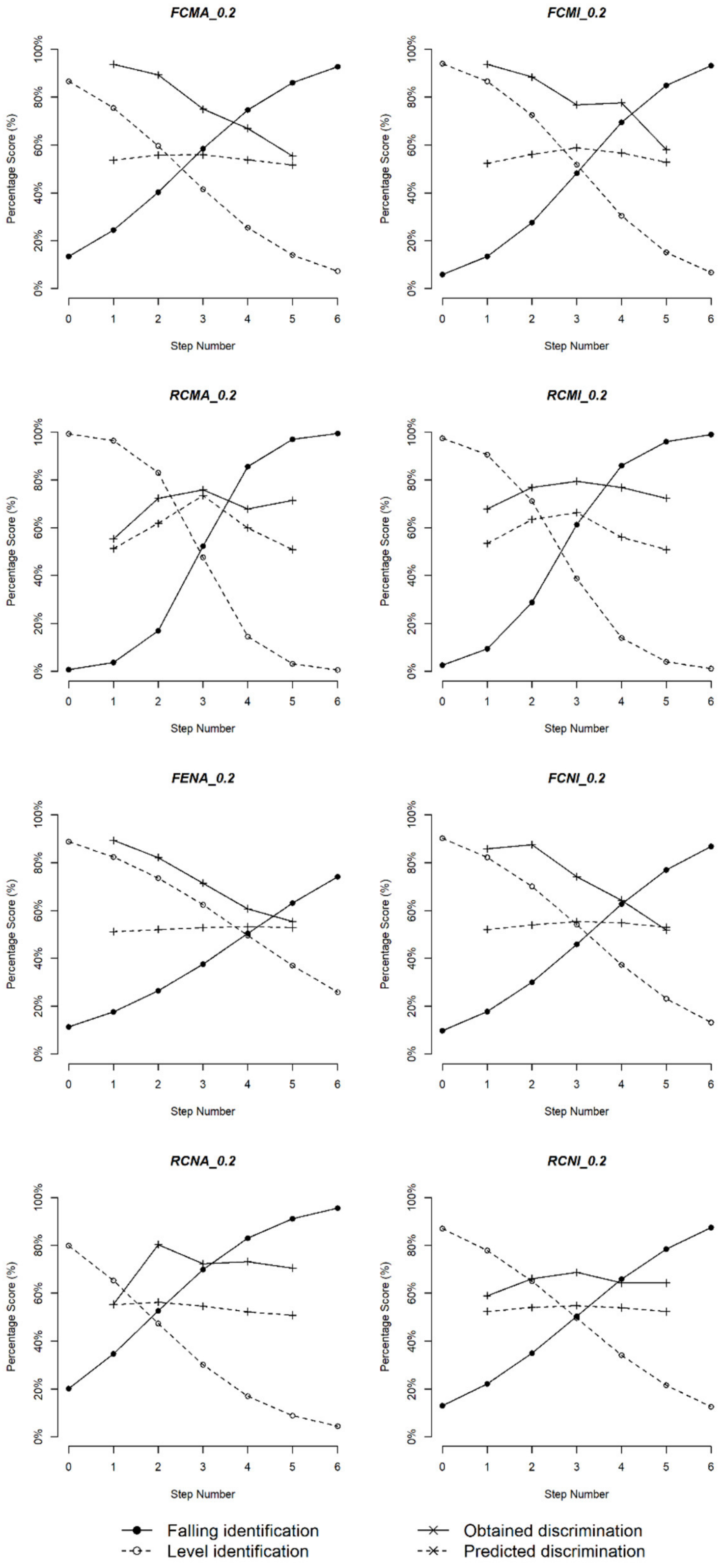


FIGURE 6 | Logistic identification functions and discrimination curves for eight subgroups of Cantonese speakers with duration 200 ms.

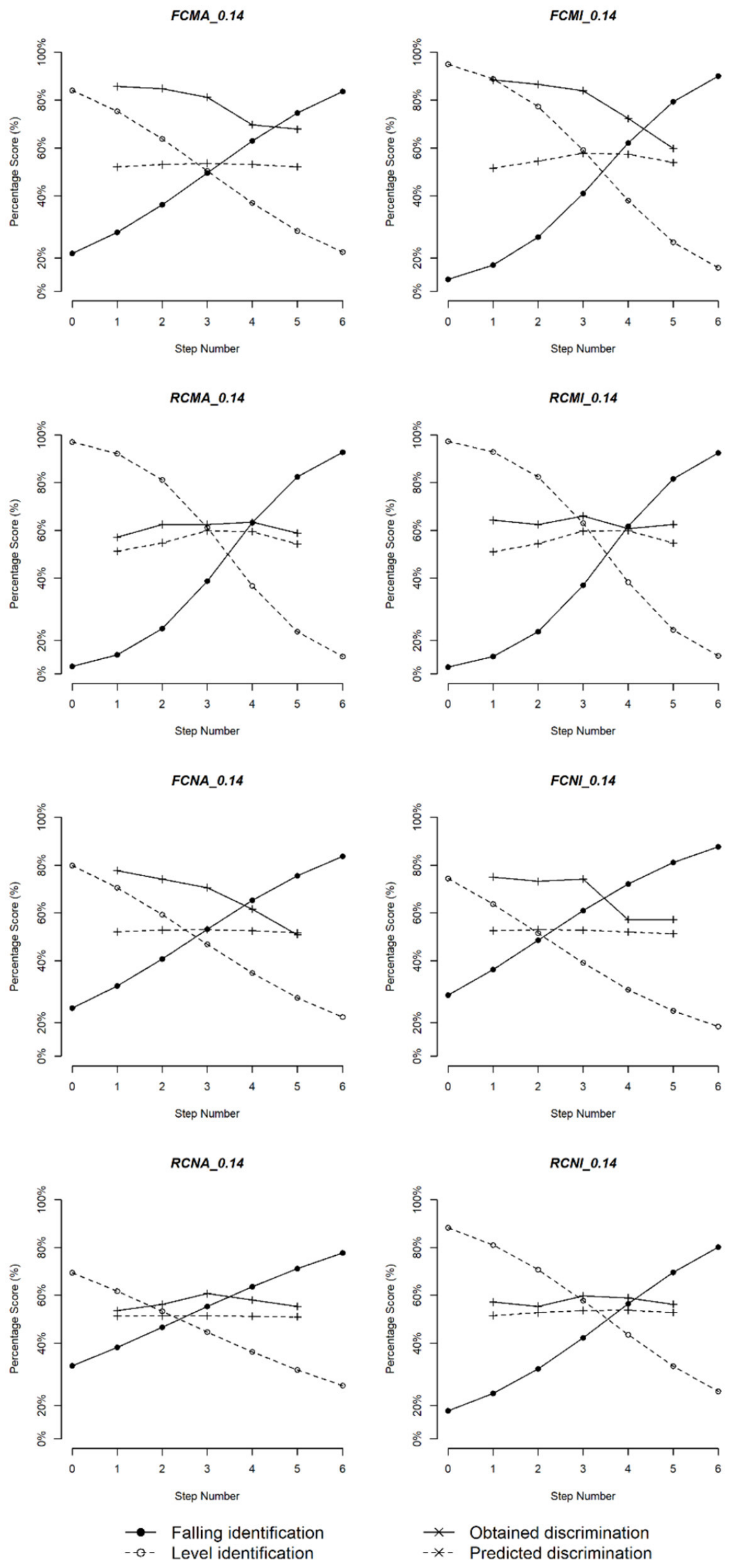


FIGURE 7 | Logistic identification functions and discrimination curves for eight subgroups of Cantonese speakers with duration 140 ms.

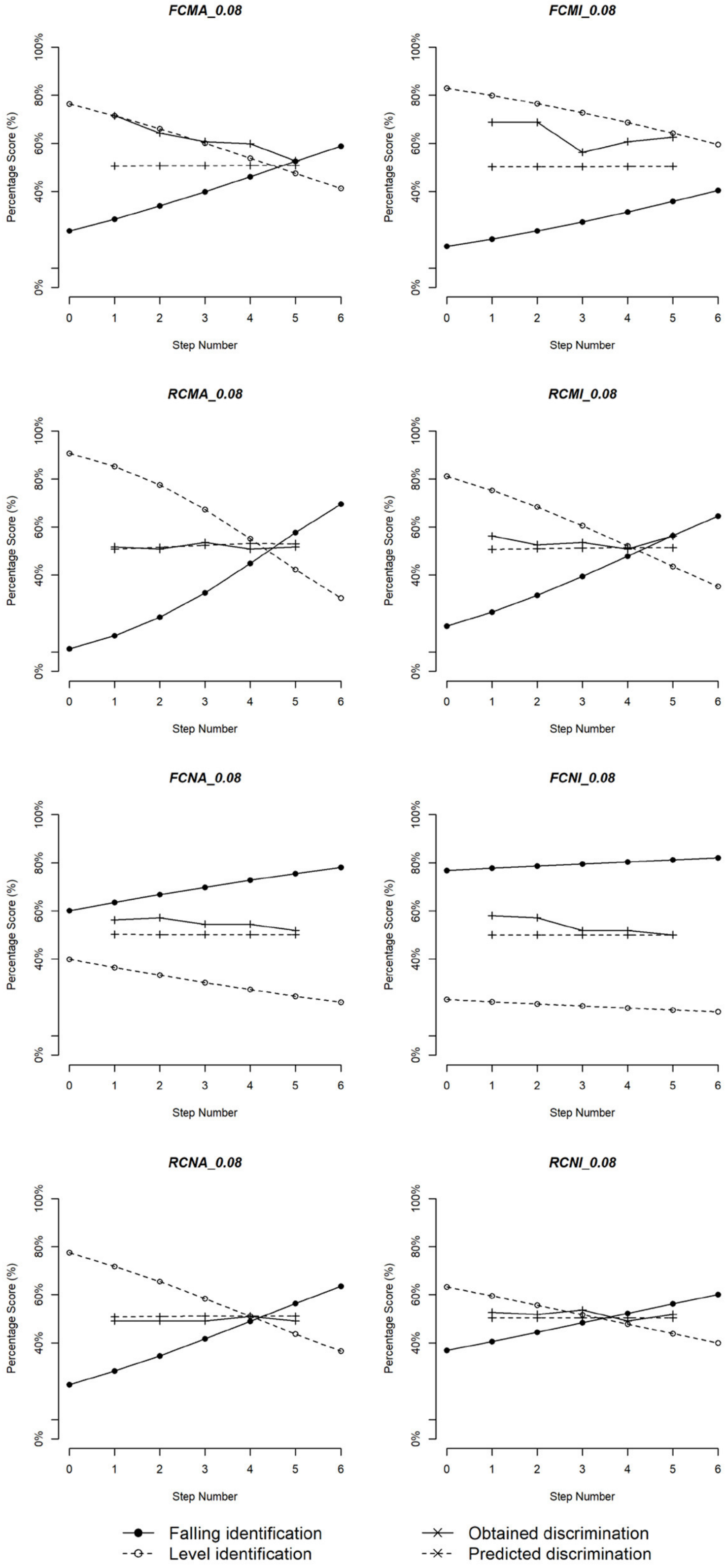


FIGURE 8 | Logistic identification functions and discrimination curves for eight subgroups of Cantonese speakers with duration 80 ms.

as vowel quality and pitch directions. We found that vowel quality did not significantly influence pitch identification among non-musicians, but musicians were more likely to hear more level tones on a high vowel /i/. In addition, both groups showed sharper category boundary for the rising than the falling pitch direction. However, the discrimination results were inconsistent. For between category discrimination, the falling pitch direction had higher *d*-prime scores than the rising pitch direction for musicians, but the reversed pattern was observed for non-musicians. For within-category discrimination, the results showed significant higher d′ scores for the falling pitch than rising pitch.

DISCUSSION AND CONCLUSION

Using a CP paradigm, this study explored the effects of musical experience, stimulus duration, vowel quality on pitch direction categorization among native Cantonese speakers. Rising and falling pitch direction continua representative of Mandarin Tone 1–2 and Tone 1–4, varying in stimulus duration and vowel height were presented to Cantonese-speaking musicians and non-musicians for identification and discrimination. Sharpness of category derived from the identification data suggested that, in general, musicians exhibited sharper category boundaries than non-musicians, an indication that their perception was more categorical. The results were similar to English-speaking musicians (Chen et al., 2020) and we did not find an obvious ceiling effect of musical experiences reported for the Mandarin-speaking musicians in their study. In addition, albeit statistically non-significance, category boundaries were sharper for the rising pitch direction than the falling pitch direction, and for the high vowel [i] than the low vowel [a] for both groups of listeners. These results suggested the universal psychoacoustic effects of pitch direction and vowel intrinsic F_0, but a unique effect of musical experience on pitch direction identification. Specifically, rising F_0 and higher intrinsic F_0 associated with a high vowel led to more precise perceptual boundaries between a level and a rising pitch contour among both groups of listeners. However, greater auditory sensitivity to both F_0 parameters (i.e., rising F_0 and higher intrinsic F_0) among musicians than non-musicians resulted in their significantly more abrupt shift in category boundaries.

In addition to sharper identification category boundary, discrimination patterns also suggested that musicians' perception of pitch direction was more categorical than that of non-musicians. Specifically, averaged across vowel height, pitch direction and duration, musicians were more sensitive to between-category differences than non-musicians. Peakedness of the discrimination function also suggested that musicians' perception of pitch direction was more categorical than that of non-musicians. That is, on average, peakedness values were larger for musicians than for the non-musicians, and for the rising than the falling pitch direction. There was also a general trend of greater peakedness with the increase of duration for both groups of participants. Moreover, although musicians had larger peakedness than non-musicians for the falling pitch direction, the two groups had comparable peakedness for the rising pitch direction. Correlation between predicted and obtained discrimination also pointed to a stronger CP among musicians as it was found that the correlation was stronger for non-musicians than musicians and for the rising pitch direction than the falling pitch direction, particularly among musicians. Regarding distances between predicted and obtained discrimination, they were numerically greater among musicians than non-musicians, but the differences did not reach statistical significance. In addition, the distance became larger as the stimulus duration increases for both groups. The effect of pitch direction was observed only among non-musicians, with greater distance for falling than rising pitch direction. In sum, against characteristics of CP, musicians' perception of pitch direction was generally more categorical than non-musicians. They exhibited sharper category boundary, greater between-category sensitivity, greater peakedness of the discrimination functions, higher correlation between predicted and obtained and greater distance between predicted and obtained discrimination.

In addition to the rising tone being inherently more salient, the possible effects of pitch range cannot be ruled out. As pointed out by an anonymous reviewer, the two Mandarin tonal pairs (T1 vs. T2, T1 vs. T4) are different in the pitch range. According to the traditional five-point scale, the tonal contour changes from point 3 to point 5 for the T2[35]/T1[55] pair, but it changes from point 5 to point 1 for the T1[55]/T4[51] pair. To control for potential confounding effects of differences in acoustic distance between members in both set of stimuli in a CP paradigm, same acoustic differences among stimuli in both tone continua are necessary. However, although the tonal continua are not natural Mandarin tonal continua, a smaller amount of F_0 fall for some members of the falling continuum (e.g., [12], [13]), may render them less "natural" and less acoustically salient than their rising counterparts in the rising continuum. This, in turn, may lead to better discrimination for the rising continuum. However, our discrimination results suggest that this might not always have been the case. For between-category discrimination, the falling pitch direction had higher *d*-prime scores than the rising pitch direction for musicians, but the reversed pattern was observed for non-musicians. For within-category discrimination, the results showed significant higher d′ scores for the falling pitch than rising pitch. These results suggest that musicality has a stronger effect on discrimination pattern than pitch range.

Our findings suggest that musicians may be at an advantage in learning tones of a second language in that they may better categorize tones in a new tonal language. Cantonese has three level tones and two rising tones, which are different from Mandarin tonal categories (one level tone and one rising tone). Although there might be some positive transfer from Cantonese tones, learners need to establish or revise the existing categories. The beneficial effects from musical training were more obvious for musicians of non-tonal speakers processing tones (Peng et al., 2010; Zhao and Kuhl, 2015a; Chen et al., 2020) than tonal speakers. Our results were inconsistent with what have been found for tone processing in the native language by musicians (Wu et al., 2015; Chen et al., 2020), suggesting that musical training may help tone processing more in a non-native tone language. It has been argued that musical training help improve auditory memory (Patel, 2011), representations of auditory objects and the mapping between new stimuli and existing memory templates (Bidelman, 2017). In addition, the sensitivity to linguistic and music pitch processing may come from a general cognitive processing (Perrachione et al., 2013). However, musical measures that were indirectly related to tones were less likely to predict tonal word learning (Bowles et al., 2016).

Furthermore, musicians were also more sensitive than non-musicians to changes in stimulus duration when identifying pitch directions. Both musicians and non-musicians benefited from increased stimulus duration, but musicians were more sensitive than non-musicians to the changes in stimulus duration reflected by greater changes in the sharpness of category. In addition, category boundary locations calculated from the identification functions showed that category boundary had a smaller value as the stimulus duration increased for both groups of listeners and musicians had a smaller value than non-musicians. Similar to the findings in Chen et al. (2020), musicians may benefit more from the extra time in that they can use the time to form a more robust auditory representation and matching sounds to internalized memory templates. Since musicians in our study are also native Cantonese speakers, they can rely on existing long-term memory of Cantonese lexical level and rising tonal categories in processing stimuli and further benefit from context-coding, matching sounds to long-term representations.

Both groups were also affected by intrinsic F_0s associated with vowel height, but only musicians appeared to be perceptually compensated for the effects. Compared to those with lower musical capacity, listeners with higher musical capacity were shown to be more capable of teasing apart acoustic cues and employ more of F_0, which is a key acoustic cue in pitch processing (Cui and Kuang, 2019). On the other hand, those with lower musical ability tend to rely on both spectral and F_0 cues. Our results thus showed that musical training affords musicians the ability to better accommodate variation in F_0 induced by factors.

While both groups showed similar effects of pitch direction in identification (i.e., sharper category boundary for the rising than the falling pitch direction), the effects of the two pitch directions were inconsistent in discrimination. The identification results were consistent with the reported difficulties that Cantonese speakers have in differentiating Mandarin Tone 1 (high-level) from Mandarin Tone 4 (high-falling) (Hao, 2012). With three contrastive level tones of different F_0 height, Cantonese speakers pay more attention to F_0 height than F_0 contour in comparison to Mandarin speakers (Peng et al., 2012). Also, although Cantonese has two rising tones (Tone 2 [25] and Tone 5 [23]), it only has a falling tone with a shallow slope (Tone 4 [21]), which may lead to their difficulties in identifying falling tones.

Though the factor of proficiency of Mandarin has been controlled between musicians and non-musicians, their exposure to Mandarin may have an effect on CP of their Mandarin lexical tones. Note that non-tonal learners of a tonal language may establish tonal categories and perceive tones in a more categorical way (Shen and Froud, 2016). However, Cantonese learners of Mandarin usually have difficulties in perceiving Mandarin tones. Cantonese speakers pay more attention to F_0 height than F_0 contours compared to Mandarin speakers due to three contrastive Cantonese level tones of different F_0 height in their tonal system (Peng et al., 2012). Also, Cantonese has two rising tones (Tone 2 [25] and Tone 5 [23]), but only a falling tone with a shallow slope (Tone 4 [21]), leading to their difficulties in identifying falling tones. Despite the learning experience of Mandarin, the identification results showed difficulties of Cantonese learners in differentiating Mandarin Tone 1 (high-level) from Mandarin Tone 4 (high-falling) (Hao, 2012). Since Cantonese has both level and rising tones, positive transfer may occur when they perceive Mandarin tones. Our data showed Cantonese musicians had sharper categorical boundary in perceiving rising tones than falling tones, though the differences were less discernible in non-musicians.

Moreover, variability in musical ability and L2 proficiency may influence tone processing. Li and DeKeyser (2017) reported that musical tonal ability is positively correlated with the accuracy rate of perception of tones and comprehension of tone words after training. Also, it has been reported that those with higher musical ability tend to separate acoustic cues better. F_0 is known as a cue most relevant to pitch processing, and listeners with higher musical ability could tease apart spectral cues and F_0 and rely on F_0 only, which may help in tone processing (Cui and Kuang, 2019). Wong and Perrachione (2007) reported that individuals' learning results were positively correlated with their musical experience. Those who received more private music lessons performed better in a Mandarin tone learning task. In addition, it has been reported that prior musical experience and musical aptitude scores predicted success in learning tonal word (Cooper and Wang, 2012) and musical aptitude scores were also positively related to tone discrimination ability (Delogu et al., 2006, 2010). In future studies, it is worth exploring how variability in musical ability and CP are related and whether more musical experience and higher capability will lead to greater sensitivity to vowel quality and more benefits from longer stimulus duration.

The variability in L2 proficiency may also influence tone processing though the results have not been consistent. It has been reported that fluent tone language speakers performed significantly better than less fluent tone language speakers and non-tone language speakers (Deutsch et al., 2009). Hao (2018) tested how proficiency level affects tone discrimination for native English speakers. Three groups of speakers were recruited, including beginning, advanced learners and those who were naïve to Mandarin. Higher accuracy were obtained by learners than those speakers naïve to Mandarin. However, both group of learners were accurate in discriminating tonal pairs except for the T2–T3 pair without significant differences. With extensive exposure to a tone language, adult listeners of non-lexical tone languages may exhibit categorical-like perception for non-native lexical tones. For example, Shen and Froud (2016) reported that native speakers of American English who are advanced learners of Mandarin perceived Mandarin tonal continua in a categorical-like manner, evidencing sharp category boundaries and prominent discrimination peaks. In fact, their discrimination performance was found to be better than that of native Mandarin listeners who had been living as university students in the US for a few years. These results suggest that CP of non-native phonetic categories can be acquired through intensive and long-term exposure. According to listeners' reports in this study, the participants were proficient non-native Mandarin speakers. It is expected that Cantonese speakers with lower proficiency or naïve to Mandarin may show weaker CP. Therefore, whether musical training may contribute more to tone processing in the initial stage of learning a tone language is worth exploring in future studies.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Human Subject Ethics Sub-committee, the Hong Kong Polytechnic University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SC and RW contributed to the conception and design of the study. YY collected data performed the statistical analysis based on code written by SC. SC, YY, and RW wrote and revised the manuscript. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

This study was supported by grants from the Department of Chinese and Bilingual Studies [88C3] and [88DW] and Faculty of Humanities [1-ZVHH], [1-ZVHJ], and [ZVNV], the Hong Kong Polytechnic University. We appreciate all the comments from the reviewers and the editor.

REFERENCES

Abramson, A. S. (1975). "The tones of Central Thai: some perceptual experiments," in *Studies in Thai linguistics in Honor of William J. Gedney,* eds J. G. Harris and J. R Chamberlain (Bangkok: Central Institute of English Language, Office of State Universities), 1–16.

Alexander, J. A., Wong, P. C. M., and Bradlow, A. R. (2005). "Lexical tone perception in musicians and non-musicians," in *Proceedings of Ninth European Conference on Speech Communication and Technology* (Barcelona), 397–400. doi: 10.21437/Interspeech.2005-271

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., Chobert, J., and Marie, C. (2011). Transfer of training between music and speech: common processing, attention, and memory. *Front. Psychol.* 2, 94. doi: 10.3389/fpsyg.2011.00094

Bidelman, G. M. (2017). Amplified induced neural oscillatory activity predicts musicians' benefits in categorical speech perception. *Neuroscience* 348, 107–113 doi: 10.1016/j.neuroscience.2017.02.015

Bidelman, G. M., Gandour, J. T., and Krishnan, A. (2011). Cross-domain effects of music and language experience on the representation of pitch in the human auditory brainstem. *J. Cogn. Neurosci.* 23, 425–434. doi: 10.1162/jocn.2009.21362

Boersma, P., and Weenink, D. (2015). *Praat: Doing Phonetics by Computer*. Available online at: http://www.praat.org/ (accessed May 01, 2020)

Bowles, A. R., Chang, C. B., and Karuzis, V. P. (2016). Pitch ability as an aptitude for tone learning. *Lang. Learn.* 66, 774–808. doi: 10.1111/lang.12159

Burnham, D., Brooker, R., and Reid, A. (2015). The effects of absolute pitch ability and musical training on lexical tone perception. *Psychol. Music* 43, 881–897. doi: 10.1177/0305735614546359

Chen, S., Zhu, Y., and Wayland, R. (2017). Effects of stimulus duration and vowel quality in cross-linguistic categorical perception of pitch directions. *PLoS ONE* 12, e0180656. doi: 10.1371/journal.pone.0180656

Chen, S., Zhu, Y., Wayland, R., and Yang, Y. (2020). How musical experience affects tone perception efficiency by musicians of tonal and non-tonal speakers? *PLoS ONE* 15, e0232514. doi: 10.1371/journal.pone.0232514

Chobert, J., François, C., Velay, J. L., and Besson, M. (2014). Twelve months of active musical training in 8-to 10-year-old children enhances the preattentive processing of syllabic duration and voice onset time. *Cereb. Cortex* 24, 956–967. doi: 10.1093/cercor/bhs377

Cooper, A., and Wang, Y. (2012). The influence of linguistic and musical experience on Cantonese word learning. *J. Acoust. Soc. Am.* 131, 4756–4769. doi: 10.1121/1.4714355

Creelman, C. D., and Macmillan, N. A. (2004). *Detection Theory: A User's Guide*. New York, NY: Psychology Press.

Cui, A., and Kuang, J. (2019). The effects of musicality and language background on cue integration in pitch perception. *J. Acoust. Soc. Amer.* 146, 4086–4096. doi: 10.1121/1.5134442

Delogu, F., Lampis, G., and Belardinelli, M. O. (2006). Music-to-language transfer effect: may melodic ability improve learning of tonal languages by native nontonal speakers? *Cogn. Process.* 7, 203–207. doi: 10.1007/s10339-006-0146-7

Delogu, F., Lampis, G., and Belardinelli, M. O. (2010). From melody to lexical tone: musical ability enhances specific aspects of foreign language perception. *Eur. J. Cogn. Psychol.* 22, 46–61. doi: 10.1080/09541440802708136

Deutsch, D., Dooley, K., Henthorn, T., and Head, B. (2009). Absolute pitch among students in an American music conservatory: association with tone language fluency. *J. Acoust. Soc. Am.* 125, 2398–2403. doi: 10.1121/1.3081389

Francis, A. L., Ciocca, V., and Ng, B. K. C. (2003). On the (non)categorical perception of lexical tones. *Percept. Psychophys.* 65, 1029–1044. doi: 10.3758/BF03194832

Fujisaki, H., and Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. *Annu. Rep. Eng. Res. Inst.* 29, 207–214.

Gottfried, T. L. (2007). "Music and language learning," in Language *Experience in Second Language Speech Learning*, eds O.-S. Bohn and M. J. Munro (Amsterdam: John Benjamins Publishing Company), 221–237. doi: 10.1075/lllt.17.21got

Gottfried, T. L., and Riester, D. (2000). Relation of pitch glide perception and Mandarin tone identification. *J. Acoust. Soc. Am.* 108, 2604. doi: 10.1121/1.4743698

Gottfried, T. L., Staby, A. M., and Ziemer, C. J. (2004). Musical experience and Mandarin tone discrimination and imitation. *J. Acoust. Soc. Am.* 115, 2545–2545. doi: 10.1121/1.4783674

Hallé, P. A., Chang, Y. C., and Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *J. Phon.* 32, 395–421. doi: 10.1016/S0095-4470(03)00016-0

Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *J. Phon.* 40, 269–279. doi: 10.1016/j.wocn.2011.11.001

Hao, Y. C. (2018). Second language perception of Mandarin vowels and tones. *Lang. Speech* 61, 135–152.

Hastie, T. J. and Pregibon, D. (1992) "Generalized linear models," in *Chapter 6 of Statistical Models*, eds J. M. Chambers and T. J. Hastie (Wadsworth & Brooks/Cole).

Howard, D., Rosen, S., and Broad, V. (1992). Major/minor triad identification and discrimination by musically trained and untrained listeners. *Music Percept*. 10, 205–220.

Lee, C.-Y., and Hung, T.-H. (2008). Identification of Mandarin tones by English-speaking musicians and non-musicians. *J. Acoust. Soc. Am.* 124, 3235–3248. doi: 10.1121/1.2990713

Lee, C.-Y., and Lee, Y.-F. (2010). Perception of musical pitch and lexical tones by Mandarin-speaking musicians. *J. Acoust. Soc. Am.* 127, 481–490. doi: 10.1121/1.3266683

Lee, Y. S., Vakoch, D. A., and Wurm, L. H. (1996). Tone perception in Cantonese and Mandarin: a cross-linguistic comparison. *J. Psycholinguist. Res.* 25, 527–542. doi: 10.1007/BF01758181

Lehnert-LeHouillier, H. (2013). "From long to short and from short to long: perceptual motivations for changes in vocalic length," in *Origins of Sound Change: Approaches to Phonologization,* ed A. C. L. Yu (Oxford University Press), 98–111. doi: 10.1093/acprof:oso/9780199573745.003.0004

Li, M., and DeKeyser, R. (2017). Perception practice, production practice, and musical ability in L2 Mandarin tone-word learning. *Stud. Second Lang. Acquisit.* 39, 593–620. doi: 10.1017/S0272263116000358

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol.* 54, 358–368. doi: 10.1037/h0044417

Liu, S., and Samuel, A. G. (2004). Perception of Mandarin lexical tones when F_0 information is neutralized. *Lang. Speech* 47, 109–138. doi: 10.1177/00238309040470020101

Maggu, A. R., Wong, P. C., Antoniou, M., Bones, O., Liu, H., and Wong, F. C. (2018). Effects of combination of linguistic and musical pitch experience on subcortical pitch encoding. *J. Neurolinguistics* 47, 145–155. doi: 10.1016/j.jneuroling.2018.05.003

Marie, C., Kujala, T., and Besson, M. (2012). Musical and linguistic expertise influence pre-attentive and attentive processing of non-speech sounds. *Cortex* 48, 447–457. doi: 10.1016/j.cortex.2010.11.006

Mok, P. P., and Zuo, D. (2012). The separation between music and speech: evidence from the perception of Cantonese tones. *J. Acoust. Soc. Am.* 132, 2711–2720. doi: 10.1121/1.4747010

Moulines, E., and Laroche, J. (1995). Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Commun.* 16, 175–205. doi: 10.1016/0167-6393(94)00054-E

Ong, J. H., Wong, P. C., and Liu, F. (2020). Musicians show enhanced perception, but not production, of native lexical tones. *J. Acoust. Soc. Am.* 148, 3443–3454. doi: 10.1121/10.0002776

Patel, A. D. (2011). Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. *Front. Psychol.* 2, 142. doi: 10.3389/fpsyg.2011.00142

Patel, A. D. (2012). "Language, music and the brain: A resource-sharing framework," in *Language and Music as Cognitive Systems*, eds P. Rebuschat (Oxford: Oxford University Press).

Patel, A. D. (2014). Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis. *Hear. Res.* 308, 98–108.

Peng, G., Zhang, C., Zheng, H. Y., Minett, J. W., and Wang, W. S. (2012). The effect of intertalker variations on acoustic-perceptual mapping in Cantonese and Mandarin tone systems. *J. Speech, Lang. Hear Res.* 55, 579–595.

Peng, G., Zheng, H. Y., Gong, T., Yang, R. X., Kong, J. P., and Wang, W. S. Y. (2010). The influence of language experience on categorical perception of pitch contours. *J. Phon.* 38, 616–624. doi: 10.1016/j.wocn.2010.09.003

Perrachione, T. K., Fedorenko, E. G., Vinke, L., Gibson, E., and Dilley, L. C. (2013). Evidence for shared cognitive processing of pitch in music and language. *PLoS ONE* 8, e73372. doi: 10.1371/journal.pone.0073372

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Percept. Psychophys.* 13, 253–260. doi: 10.3758/BF03214136

Pollack, I., and Pisoni, D. (1971). On the comparison between identification and discrimination tests in speech perception. *Psychon. Sci.* 24, 299–300. doi: 10.3758/BF03329012

Prom-on, S., Xu, Y., and Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *J. Acoust. Soc. Am.* 125, 405–424. doi: 10.1121/1.3037222

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. Available online at: https://www.r-project.org (accessed Sep 14, 2020).

Schneider, W., Eschman, A., and Zuccolotto, A. (2012). *E-Prime User's Guide*. Pittsburgh: Psychological Software Tools Inc.

Shen, G., and Froud, K. (2016). Categorical perception of lexical tones by English learners of Mandarin Chinese. *J. Acoust. Soc. Am.* 140, 4396–4403. doi: 10.1121/1.4971765

Stoll, G. (1984). Pitch of vowels: experimental and theoretical investigation of its dependence on vowel quality. *Speech Commun.* 3, 137–147. doi: 10.1016/0167-6393(84)90035-9

Tang, W., Xiong, W., Zhang, Y. X., Dong, Q., and Nan, Y. (2016). Musical experience facilitates lexical tone processing among Mandarin speakers: behavioral and neural evidence. *Neuropsychologia* 91, 247–253. doi: 10.1016/j.neuropsychologia.2016.08.003

Wang, W. S.-Y. (1976). "Language change," in *Origins and Evolution of Language and Speech,* eds S. R. Harnad, H. D. Steklis, and J. Lancaster(New York, NY: New York Academy of Sciences), 61–72. doi: 10.1111/j.1749-6632.1976.tb25472.x

Wang, W. S. -Y., Lehiste, I., Chuang, C., and Darnovsky, N. (1976). Perception of vowel duration. *J. Acoust. Soc. Am.* 60, S92–S92. doi: 10.1121/1.2003607

Wang, Y., Yang, X., and Liu, C. (2017). Categorical perception of mandarin chinese tones 1–2 and tones 1–4: effects of aging and signal duration. *J. Speech Lang. Hear. Res.* 60, 3667. doi: 10.1044/2017_JSLHR-H-17-0061

Whalen, D. H., and Levitt, A. G. (1995). The universality of intrinsic F_0 of vowels. *J. Phon.* 23, 349–366. doi: 10.1016/S0095-4470(95)80165-0

Wong, P. C., Skoe, E., Russo, N. M., Dees, T., and Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nat. Neurosci.* 10, 420–422. doi: 10.1038/nn1872

Wong, P. C. M., and Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Appl. Psycholinguist.* 28, 565–585. doi: 10.1017/S0142716407070312

Wu, H., Ma, X., Zhang, L., Liu, Y., Zhang, Y., and Shu, H. (2015). Musical experience modulates categorical perception of lexical tones in native Chinese speakers. *Front. Psychol.* 6, 436. doi: 10.3389/fpsyg.2015.00436

Xu, B. R., and Mok, P. (2012). "Cross-linguistic perception of intonation by Mandarin and Cantonese listeners," in Speech Prosody *2012* (Shanghai).

Xu, Y., Gandour, J. T., and Francis, A. L. (2006). Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *J. Acoust. Soc. Am.* 120, 1063–1074. doi: 10.1121/1.2213572

Yu, A. C. L., Lee, H., and Lee, J. (2014). "Variability in perceived duration : pitch dynamics and vowel quality," in *Proceedings of the 4th International Symposium on Tonal Aspects of Languages (TAL* 2014) (Nijmegen), 41–44.

Zatorre, R. J., and Halpern, A. R. (1979). Identification, discrimination, and selective adaptation of simultaneous musical intervals. *Percept. Psychophys.* 26, 384–395.

Zhao, T. C., and Kuhl, P. K. (2015a). Effect of musical experience on learning lexical tone categories. *J. Acoust. Soc. Am.* 137, 1452–1463. doi: 10.1121/1.4913457

Zhao, T. C., and Kuhl, P. K. (2015b). Higher-level linguistic categories dominate lower-level acoustics in lexical tone processing. *J. Acoust. Soc. Am.* 138, EL133–EL137. doi: 10.1121/1.4927632

Zheng, H. Y., Minett, J. W., Peng, G., and Wang, W. S. (2012). The impact of tone systems on the categorical perception of lexical tones: An event-related potentials study. *Lang. Cogn. Processes* 27, 184–209.

Zhu, J., Chen, X., and Yang, Y. (2021). Effects of amateur musical experience on categorical perception of lexical tones by native chinese adults: an ERP study

[Internet]. *Front. Psychol*. p. 690. Available online at: https://www.frontiersin.org/article/10.3389/fpsyg.2021.611189

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Psychology

ORIGINAL RESEARCH
published: 22 October 2021
doi: 10.3389/fpsyg.2021.717110

Music Does Not Facilitate Lexical Tone Normalization: A Speech-Specific Perceptual Process

Ran Tao[†], *Kaile Zhang*[†] *and Gang Peng*[*]

Research Centre for Language, Cognition, and Neuroscience, Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China

Listeners utilize the immediate contexts to efficiently normalize variable vocal streams into standard phonology units. However, researchers debated whether non-speech contexts can also serve as valid clues for speech normalization. Supporters of the two sides proposed a general-auditory hypothesis and a speech-specific hypothesis to explain the underlying mechanisms. A possible confounding factor of this inconsistency is the listeners' perceptual familiarity of the contexts, as the non-speech contexts were perceptually unfamiliar to listeners. In this study, we examined this confounding factor by recruiting a group of native Cantonese speakers with sufficient musical training experience and a control group with minimal musical training. Participants performed lexical tone judgment tasks in three contextual conditions, i.e., speech, non-speech, and music context conditions. Both groups were familiar with the speech context and not familiar with the non-speech context. The musician group was more familiar with the music context than the non-musician group. The results evidenced the lexical tone normalization process in speech context but not non-speech nor music contexts. More importantly, musicians did not outperform non-musicians on any contextual conditions even if the musicians were experienced at pitch perception, indicating that there is no noticeable transfer in pitch perception from the music domain to the linguistic domain for tonal language speakers. The findings showed that even high familiarity with a non-linguistic context cannot elicit an effective lexical tone normalization process, supporting the speech-specific basis of the perceptual normalization process.

Keywords: speech normalization, tone normalization, music, lexical tones, Cantonese

OPEN ACCESS

Edited by:
Chao-Yang Lee,
Ohio University, United States

Reviewed by:
Yue Wang,
Simon Fraser University, Canada
Keith K. W. Leung,
Simon Fraser University, Canada, in collaboration with reviewer YW
William Choi,
The University of Hong Kong,
Hong Kong SAR, China

Correspondence:
Gang Peng
gpeng@polyu.edu.hk

[†] *These authors have contributed equally to this work*

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

Received: *30 May 2021*
Accepted: *30 September 2021*
Published: *22 October 2021*

Citation:
Tao R, Zhang K and Peng G (2021) Music Does Not Facilitate Lexical Tone Normalization: A Speech-Specific Perceptual Process. Front. Psychol. 12:717110. doi: 10.3389/fpsyg.2021.717110

INTRODUCTION

Humans communicate in language and music. In both formats, the continuous acoustic signals are segmented and then categorized into abstract meaningful units (e.g., words and melodies). Musical performance and appreciation require deliberate practice and longitudinal exposure, but speech production and perception abilities are developed naturally. Albeit speech categorization is sometimes demanding since there is no one-to-one mapping between acoustic signals and linguistic units due to speaker variability. In speech production, speakers vary a lot in their vocal tract configurations, which results in a large individual difference in speech production (Peterson and Barney, 1952). Even speech production by the same speaker may change a lot in different situations (Newman et al., 2001). The inter- and intra-speaker variability blurs the boundary between two acoustically similar phonemes and makes them less distinguishable. For example, a male speaker's production of a high-level tone may have a similar pitch height as a female speaker's production of a mid-level tone (Peng et al., 2012). The word identification is slower and less accurate when the speech stimuli are presented in the mixed-speaker condition than in the blocked-talker condition (Nusbaum and Magnuson, 1997; Magnuson and Nusbaum, 2007), revealing an obstacle in speech perception introduced by high speech variability.

Language as a complex system also provided rich information for us to overcome the speech variability and achieve perceptual constancy. Ladefoged and Broadbent (1957) found that the ambiguous 'bVt' syllable was more likely perceived as 'bit' in a sentence with high first formant (F1) and as 'bet' in a sentence with low F1. This pioneering work demonstrates that the acoustic information embedded in context affects our interpretation of the target speech cues and thus to some extent reduces the ambiguity caused by the inter- and intra-talker variability, a process known as extrinsic normalization (Nearey, 1989). The extrinsic normalization has been widely observed in the perception of vowels, consonants, and lexical tones. The perception of Cantonese lexical tones relies heavily on the extrinsic context information. The primary acoustic correlate of lexical tones is the fundamental frequency (F0), and the height and the slope of F0 trajectory affect the tone differentiation (Gandour, 1983). However, Cantonese has three level tones, high-level tone, middle-level tone, and low-level tone, which can only be differentiated by pitch heights (Peng, 2006). Therefore, the contextual F0 which provides a good reference for listeners to estimate the relative pitch height of the target stimuli becomes important. Cantonese speakers' perception of three level tones was improved significantly (from 48.6% to above 90%) when the isolated tonal stimuli were embedded into speech contexts (Wong and Diehl, 2003). Wong and Diehl (2003) and Zhang et al. (2013,2017,2021) reported that an ambiguous Cantonese middle-level tone is more frequently perceived as the high-level tone after a context of low F0 and perceived as the low-level tone after a context of high F0, indicating a contrastive context effect in the lexical tone normalization process.

Spectrally Contrastive Encoding and General-Auditory Level Processing

Since the observation of context effect, different theories are proposed to explain the underlying mechanisms of the extrinsic normalization process. There is an ongoing dispute about whether the extrinsic normalization operates on the general-auditory level or the speech-specific level. The core issue of this debate mainly lies in the effectiveness of non-speech context on perceptual normalization. Huang and Holt (2009) reported that Mandarin lexical tone perception was contrastively affected by the non-speech context composed of the sine-wave harmonics or the pure tone. Specifically, listeners perceived the ambiguous lexical tones as high-level tone (Mandarin T1) if its preceding non-speech context had low frequency and as high-rising tone (Mandarin T2) if the preceding non-speech context had high frequency. More importantly, the contrastive context effect in non-speech contexts (albeit quantitatively smaller) was statistically comparable with that in the speech context. The effectiveness of non-speech contexts got strong support from a serial of studies on consonant normalization as well. The non-speech composed of sine-wave tones and white noise affected listeners' perception of /ga/-/da/continuum in a similar manner as the speech context (Lotto and Kluender, 1998; Holt, 2006), even if context and target were separated by more than one second or multiple intervening sounds (Holt, 2005). Japanese quails demonstrated the contrastive perception behavior after training, which further suggested that the normalization process was not constrained by the speech-specific processing (Lotto et al., 1997). Since non-speech is extralinguistic, Lotto and Kluender (1998) suggested that the normalization process required no specific linguistic knowledge and that it operated on the general perceptual level and depended on the spectral contrast between context and target stimuli.

The contrastive perceptual pattern is essentially consistent with forwarding energy masking, which shows that the target after the masker is perceived less accurately when the masker has acoustic energies in the same frequency region as the target (Moore, 1995; Viswanathan et al., 2013). Energy masking is partially caused by the inertia of the auditory nerve; that is, the basilar membrane takes time to recover after responding to the masker (Duifhuis, 1973). Aside from the physiological basis in the peripheral auditory system, the neural adaptation of the central auditory system also contributes to contrast encoding (e.g., Bertrand, 1997). As illustrated by the oddball paradigm, while being continuously exposed to the same auditory stimuli, neurons decrease their firing rates and become less active; when new stimuli are presented, neurons are activated and increase their firing rates, generating large event-related potential (ERP) amplitude (Polich, 2007). This working mechanism of neurons may also apply to the extrinsic normalization process. Neurons in the auditory cortex are responsive to different frequency regions (Sjerps et al., 2019). Neurons adapted by the preceding context are less responsive to the same frequencies in the following target, but neurons that do not fire in the context presentation are relatively more sensitive to the frequency ranges of the following sounds, resulting in spectrally contrastive perception (Stilp, 2020).

Context Tuning and Speech-Specific Level Processing

The normalization differences between speech and non-speech contexts were also reported. Francis et al. (2006) compared the normalization of Cantonese middle-level tones in speech and non-speech contexts. The speech context was meaningful Cantonese sentence /ŋɔ23 wui23 tɔk22 ji33 pɛi25 lɛi23 t'ɛŋ55/[I will read ji for you (to hear)], and the non-speech context was synthesized by applying the pitch contour of the speech context to the 'hummed' unintelligible neutral vocal tract /ə/ with Praat (Boersma and Weenink, 2016). They found that even though the non-speech context contained the crucial cues for the pitch range estimation (i.e., the same pitch contour as the speech context), native listeners showed almost no normalization effect in the non-speech context. However, the contrastive perceptual pattern was noticeable in the speech context. Their study revealed an unequal effect of speech and non-speech contexts in the lexical tone normalization process. Zhang et al. (2015) further tested the contribution of speech information at each level. They asked native Cantonese speakers to identify the ambiguous Cantonese middle-level tones in non-speech contexts (triangle waves), reversed speech contexts (normal speech reversed in time scale, sounding like foreign phonemes), meaningless speech contexts (Cantonese monosyllabic sequences), and meaningful speech contexts. The information in four contexts also decreased from meaningful speech (semantic, phonological, phonetic, and acoustic information), meaningless speech (phonological, phonetic, and acoustic information), reversed speech (phonetic and acoustic information), to non-speech (acoustic information). They found that meaningful speech exerted the largest normalization effect, which was followed by the meaningless speech contexts. The reversed speech context also showed some positive effects on the normalization process, but the normalization effect in non-speech was almost negligible. Francis et al. (2006) and Zhang et al. (2015) suggest that the normalization effect of non-speech context is not as prominent as speech context and that speech-specific information (i.e., semantic, phonological, and phonetic information) is necessary for the Cantonese level tone normalization.

Instead of a contrastive encoding of auditory signals, some researchers believe that the speech normalization process operates via the context tuning mechanism. According to the context tuning mechanism, listeners use extrinsic contextual information to compute a talker-specific mapping of acoustic patterns onto abstract linguistic units, and ambiguous target speech cues are identified by referring to that mapping. Joos (1948) described that a talker-specific vowel pattern can be quickly established even during their first greeting ‘How do you do?’. Three critical phonemes /a/, /j/, and /u/in the greeting can roughly outline the vowel space of that speaker since the /a/ is pronounced with the low central articulation gesture, the /u/ with the highest and strongest back articulation, and the /j/ with a higher and more forward articulation. The incoming acoustic signals can be categorized by referring to this vowel pattern. Joos’ description indicates that the mapping used in normalization is essentially an acoustic-phoneme mapping. To form such a mapping, linguistic knowledge is required, indicating a normalization process at the speech-specific level.

A Hybrid Model of Speech Perception: Co-existence of Exemplars and Abstract Linguistic Representations

Although general auditory contrastive encoding mechanism and the context tuning mechanism, to some extent, explain the context effect on ambiguous speech perception, they can hardly explain why the typical speech is almost not affected by context cues. For example, the perception of two endpoints of the speech continuum does not change in different context shifts (Johnson, 1990) and speakers whose pitch ranges are closer to the population mean are less affected by the context F0 as well (Zhang et al., 2013). These findings suggest that while perceiving speech in context, contextual cues only partially contribute to our final decision and that another ongoing perceptual mechanism that utilizes the specific characteristics of each token also affects our final phonemic categorization. The token-specific effect can be well explained by the exemplar-based theory which believes that the exemplars we encounter in our daily life form the mental representations of each phonological category, and that speech perception is a match between stored exemplars and the incoming signal (Johnson, 1997). Based on this account, the speaker-specific details are also kept with the exemplars and are helpful cues for speech perception.

The normalization approach emphasizes the computation of an abstract and speaker-independent mental representation, but the exemplar-based approach utilizes the speaker-specific details to match the stored exemplars. Considering that both abstract phonological categories and fine acoustic details were reported to affect speech perception, a hybrid model was proposed to accommodate these two different views (Tuller, 2003; Nguyen et al., 2009). In the hybrid model, the mental representation of each phonological category is a multi-layered construct. Listeners maintain multiple exemplars with speaker-specific details in the lower layer and these exemplars gradually decay into more abstract speaker-independent representations in the upper layer. Correspondingly, speech perception may be a multi-pathway process as well. The normalization process extracts invariant elements from speech tokens and matches them to the abstract phonological categories. Meanwhile, the details of speech tokens are kept in memory, which constrains speech categorization.

To investigate the mental representation of phonological units, Kimball et al. (2015) asked listeners to identify whether two sounds in a trial were the same or different. The different trial was either the phonological variation (presence or absence of a pitch accent) or the phonetic variation (different durations or F0 peaks). They found that listeners could accurately identify the trials differing either in the phonological level or the phonetic level. When the interval between two sounds increased, the accuracy dropped for the phonetic variation but not for the phonological variation. Their results suggested that both phonetic details and abstract phonological categories were kept in our memory, and phonological distinctions were more robust, supporting the hybrid model of speech perception.

Familiarity and the Speech Normalization Process

The hybrid model, especially the exemplar layer, challenges the account that the speech superiority is due to the speech-specific nature of the normalization process. Although previous research made spectral complexity and spectral contrast comparable in speech and non-speech contexts, speech and non-speech still differ in many aspects. In addition to the speech-specific information (i.e., phonetic, phonological, and semantic information) which may favor a speech-specific account, listeners have different familiarities with speech and non-speech stimuli. Compared with the non-speech contexts used in previous studies (e.g., pure tones, harmonic complex tones, triangular waves, hummed sound modeled on a neutral vocal tract, or iterated rippled noise), listeners are much more familiar with speech contexts. They store more exemplars of speech than non-speech in their daily exposure to sound, and robust phonemic representations are established during their long-term language acquisition and usage. The rich exemplars on the lower layer and the robust phonemic representation on the higher layer result in a stronger activation of speech context than non-speech context. Familiarity also affects the efficiency of speech perception, which is boosted due to the countless repetition in daily communication, but the decoding of non-speech is rare and thus less automatic. The spectral characteristics in the non-speech context are probably not utilized due to the weak activation and/or limited process.

Familiarity advantage has been widely reported in speech perception studies, for example, faster response to speech spoken in familiar languages (e.g., Hu et al., 2017) and better identification of familiar talkers' speech in noise (e.g., Nygaard and Pisoni, 1998). Familiarity advantage might exist in the perceptual normalization as well since previous studies found that listeners' speech perception is better with contexts spoken in their native language (e.g., Lee et al., 2009; Kang et al., 2016; Zhang, 2020). Lee et al. (2009) asked native Mandarin speakers and native English speakers to identify the digitally processed Mandarin tones either in isolation or with Mandarin context. These syllables were produced by either single talker or multiple talkers. Talker variability affected Mandarin and English listeners equally. However, Mandarin listeners made better use of context to compensate for the speech variability and improved the lexical tone identification, suggesting a language familiarity advantage in the extrinsic normalization of lexical tones. Similar results were also reported in the extrinsic normalization of segmental components. The context composed of vowel /y/ can facilitate French speakers but not English speakers to perceive ambiguous /s–ʃ/ sound probably because vowel /y/ exists in French but not in English (Kang et al., 2016). The context effect of the native context /ε i/ was more prominent than that of the non-native context /oe y/ when English speakers perceived ambiguous vowel /u–ɔ/ (Zhang, 2020).

However, different findings were also reported. Sjerps and Smiljanić (2013) tested how language familiarity affected vowel normalization. They asked native listeners of American English, Dutch, Spanish and Spanish-English bilinguals to perceive the ambiguous /sofo/-/sufu/ with contexts spoken in either Dutch, Spanish, or English. They found that the perceptual impact of precursor context was comparable in size across listeners' language backgrounds, indicating a weak effect of the language familiarity on the speech normalization process. Magnuson et al. (2021) tested how the talker familiarity affected the accommodation of speech variabilities. They asked native Japanese speakers to identify morae produced by either familiar talkers (family members) or strangers in either blocked- or mixed-talker conditions. Listeners always took a longer time to recognize morae in the mixed-talker condition even for the voice of their family members, indicating a constant cost for talker accommodation regardless of talker familiarity.

Albeit inconclusive, the familiarity difference between speech and non-speech material is a potential factor that contributes to the superiority of speech context in the normalization process. Probably, either familiarity or speech-specific information contributes to the normalization process, or they work together to facilitate speech normalization. By teasing familiarity and speech-specific information apart, we can test if the speech-specific information is the only or main contributor to the speech-superiority effect, which may partially clarify the long-standing dispute between the general-auditory and speech-specific basis of the extrinsic lexical tone normalization. If normalization effects are comparable between speech and non-speech contexts when the familiarity gap is controlled, the speech-specific information might not play a crucial role in extrinsic normalization and the normalization process is largely processed by a general-auditory mechanism. However, if speech context still shows a significantly better normalization effect than other contexts when their familiarities are comparable, this could be strong evidence for the speech-specific basis of extrinsic normalization.

To test the confounding factor familiarity, the present study compares the native Cantonese speakers' perception of ambiguous Cantonese level tones in the context of either speech, music, or synthesized non-speech. Music is one of the most meaningful and popular forms of non-verbal sound; like speech, it has been developed to take advantage of the efficiencies of the human auditory system (Baldwin, 2012). Music also follows syntax-like rules, makes use of pitch and rhythm, and has a ubiquitous presence across human civilizations (Zatorre et al., 2007; Patel, 2013). The prevalence and the use of pitch as an importation cue make music an ideal non-speech context for the present study to test the familiarity effect. For Cantonese speakers who never received professional musical training, their familiarities with the three contexts decrease from speech, music, to synthesized non-speech. Meanwhile, familiarity is further manipulated by including a group of Cantonese-speaking musicians who receive professional musical training and thus are much more familiar with musical materials than non-musicians. If familiarity is the main factor for the unequal effect of the speech and non-speech context, musicians are expected to perform better than non-musicians at least in the musical context, and meanwhile, both groups are expected to show a performance improvement when contexts change from synthesized non-speech, music to speech.

Music Experience and the Lexical Tone Normalization Process

To our knowledge, no study directly tested how music experience affects the lexical tone normalization. A few studies about the congenital amusia, a neurodevelopmental disorder of pitch processing (Ayotte et al., 2002) may shade light on this question from a different angle, that is how the music deficit affects the lexical tone normalization. People with congenital amusia (amusics) showed severe deficit in the perception of musical melody. Zhang C. et al. (2018) asked Cantonese-speaking amusics and controls to perceive the ambiguous Cantonese mid-level tone in contexts with different pitch heights. The control group showed noticeable normalization effect in speech context, but the normalization effect in speech context is much reduced for amusics. Shao and Zhang (2018) further tested the perception of six Cantonese tones with and without context for two groups. They found that controls performed better with context cues, but that amusics in most cases failed to benefit from the context cues. Similar result was also observed in Mandarin speakers. Liu et al. (2021) reported that Mandarin-speaking amusics cannot utilize the contextual information to perceive the ambiguous Mandarin T55–T35 continuum, but control group without amusia showed typical context effect in the Mandarin lexical tone perception. The studies about amusia suggest that the impaired pitch perception ability in music domain affects the lexical tone normalization. Based on the findings from amusics, it is natural to hypothesize that people with music experience probably perform better in the lexical tone normalization.

This hypothesis is somewhat supported by the studies which reported that music experience affects lexical tone perception (for a review, please see Ong et al., 2020). While detecting the subtle sentence-final pitch variation, French musicians performed better than non-musicians no matter in their native language (Schön et al., 2004) or the non-native language (Marques et al., 2007). French musicians could also detect the variation of Mandarin lexical tones better than non-musicians (Marie et al., 2011). English Musicians' identification and discrimination of Mandarin lexical tones were faster and more accurate than non-musicians (Alexander et al., 2005). Even musicians of tonal language speakers, for example, Mandarin musicians, showed increased sensitivity to the fine acoustic difference of mandarin tones (Wu et al., 2015). All these findings support a positive transfer from music experience to pitch perception in the language domain. Considering musicians' improved pitch perception ability, they are expected to extract the contextual pitch information more accurately and thus have a more precise pitch range reference to estimate the relative pitch height of the target tone. Therefore, these tone perception studies make it more reasonable to hypothesize that music experience boosts the extrinsic normalization of lexical tones.

The present study also explicitly tests this hypothesis by comparing musicians who receive intensive and professional music training with non-musicians who have rare music experience in a lexical tone normalization task. If the hypothesis holds, musicians are expected to show a stronger context effect than non-musicians at least in the speech context condition. Considering that musicians are reported to have better pitch perception ability in both linguistic and non-linguistic domains, they probably perform better than non-musicians in the non-speech contexts (music and synthesized non-speech) as well. If this is true, the extrinsic normalization of lexical tones is largely determined by the domain-general pitch processing ability but not the speech-specific processing, which to some extent is in line with the general auditory mechanism and the familiarity hypothesis (i.e., the frequent practice in pitch perception). On the contrary, if musicians fail to show any advantage over non-musicians in tone normalization in any kind of context conditions, the results will favor a speech-specific mechanism.

MATERIALS AND METHODS

Participants

Forty native Cantonese adults participated in this experiment, among whom 20 were categorized as non-musicians (10 female, Age = 21.9 ± 2.96) and 20 were categorized as musicians (10 female, Age = 23.6 ± 4.69). Participants were matched in their age [Welch's $t(32.1) = 1.325$, $p = 0.194$] and gender. Non-musicians were defined as individuals with less than 3 years of musical training except the mandatory courses in their primary or middle schools. Musicians were defined as individuals with at least 7 years of private musical training and still actively engaging in music (Wong et al., 2007; Wayland et al., 2010; Cooper and Wang, 2012), such as practicing music, studying in music major, or having a music-related occupation (e.g., band member, private music tutor, and music teacher in schools). The musicians had a diverse background of music learning experience and some of them reported having learned several kinds of instruments. Characteristics of musicians are summarized in the **Supplementary Table S1**. One non-musician was identified as ambidextrous using the Edinburgh handedness inventory (Oldfield, 1971) and the rest participants were right-handed. All participants reported no hearing loss, neuropsychiatric disorders, or brain injuries. Participants were compensated for their time and signed consent forms before the experiment. The experiment procedure was approved by the Human Subjects Ethics Sub-committee of The Hong Kong Polytechnic University.

Stimuli

Preparation of the stimuli and the experimental procedure followed previous work (Zhang et al., 2013, 2017; Tao and Peng, 2020). Stimuli consisted of contexts and targets in four different context conditions: a speech context condition, two non-linguistic contexts (e.g., synthesized non-speech and music), and a condition without context (coded as isolated hereafter). Speech contexts and all targets were produced by four native Cantonese talkers who were a female talker with a high pitch range, a female talker with a low pitch range, a male talker with a high pitch range, and a male talker with a low pitch range (coded as FH, FL, MH, and ML respectively). Speech context was a four-syllable meaningful sentence, i.e., 呢個字係 (/li55 ko33 tsi22 hɐi22/, "This word is meaning"). After recording the natural production of the sentence from the four talkers, the F0 trajectories of the sentences (see **Supplementary Figure S1**) were then lowered and raised three semitones to trigger the contrastive context effect (Wong and Diehl, 2003). Specifically, more high-level tone responses were expected in the lowered context condition and more low-level tone responses in the raised context condition. In sum, three sets of speech contexts were formed: a set of F0 lowered contexts, a set of F0 unshifted contexts, and a set of F0 raised contexts. All targets from four context conditions were the natural production of the Chinese character 意 (e.g., /ji33/mid-level tone, "meaning," also see **Supplementary Figure S1** for F0 trajectories).

The non-speech contexts were produced by applying the F0 trajectory and intensity profile from speech contexts to triangle

TABLE 1 | Mean fundamental frequency (Hz) of speech contexts, their counterparts, and targets.

		FH	FL	MH	ML
Speech/ non-speech	F0 raised	280.5	246.3	174.6	134.5
	F0 unshifted	236.8	208.1	148.4	113.8
	F0 lowered	198.2	173.9	124.5	96.8
Music	F0 raised	294.4	255.6	181.6	139.6
	F0 unshifted	247.7	215.4	153.8	117.6
	F0 lowered	207.6	181.2	128.7	98.7
Target		233.2	206.8	143.8	114.9

waves. The music contexts were piano notes that had the closest pitch height to each of the syllables in the speech context, which were generated using a Kurzweil K2000 synthesizer tuned to the standard A4 of 440 Hz (Peng et al., 2013). We chose the closest piano notes rather than synthesizing a piano sound with the mean F0 of each syllable to ensure that the musicians would feel as natural as possible when hearing these notes. The manipulation on speech F0 and selection of piano notes caused a slight discrepancy between conditions (see **Table 1** for a list of mean F0 of all contexts and Targets), however, the hierarchy of F0 between raised, unshifted, and lowered conditions were reliably reserved. **Figure 1** lower panel shows a schema of context stimuli preparation. All speech stimuli, including speech contexts and targets, were adjusted to 55 dB in intensity. The non-linguistic contexts, including non-speech and music contexts, were adjusted to 75 dB in intensity to match the hearing loudness of speech contexts. The duration of speech contexts was kept unchanged to reserve the natural production outcome (FH: 1005 ms, FL: 888 ms, MH: 811 ms, ML: 821 ms). The duration of non-speech contexts was the same as their corresponding speech contexts. The duration of music contexts was 1000 ms with each note lasting 250 ms.

Fillers were prepared with the same procedure. In the speech context condition, the filling context was two four-syllable sentences, i.e., 我而家讀 (/ŋo23 ji21 ka55 tuk2/, “Now I will read,” recorded from FL and MH) and 請留心聽 (/ts^{h}iŋ25 ləu21 sɐm55 t^{h}iŋ55/, “Please listen carefully to,” recorded from FH and ML). Target fillers were Chinese characters 意 (recorded from FL and MH) or 二 (e.g., /ji22/ low-level tone, “two,” recorded from FH and ML).

Experiment Procedure

All participants attended a word identification task in a sound-proof booth. Participants were asked to make a judgment on the target syllable following a preceding context. In each experiment trial, the target and context corresponded with each other, i.e., the target always followed the context produced by the same talker or its non-verbal counterparts. Participants were instructed to listen to both the context and the target attentively. Specifically, they first saw a 500 ms fixation in the middle of the screen followed by the context presented through earplugs, and then after a jittering silence (range: 300–500 ms), a target syllable was presented. In the isolated condition, participants heard the target without a context, i.e., the fixation was followed by the jittering silence immediately. Participants then made a judgment on the target syllable from three choices of 醫 (/ji55/ high-level tone, “doctor”), 意 (middle-level tone), or 二 (low-level tone) by pressing designated keys on the keyboard when they saw a cue on the screen. The cue was a question mark on the middle of the screen, delayed 800–1000 ms from the onset of the target (see **Figure 1**). In this kind of setting, reaction times were not meaningful indices of participants’ psycholinguistic processing and thus were not analyzed in this study. We focused on the participants’ judgments on the targets.

The four context conditions were grouped into four experimental blocks which were counterbalanced across participants to prevent order effect. The isolated condition block consisted of 16 repetitions of each target. The blocks of three context conditions each consisted of nine repetitions of three F0 shifts of four talkers.

Analysis

First, we evaluated the effect of a preceding context on the perception of middle-level tones by comparing the listeners’ response patterns on the targets following various contexts (and without a context, e.g., isolated condition) with a three-way ANOVA. Two within-subject factors were Context (isolated, music, non-speech, speech) and Choice (judging the target as high-, middle-, and low-level tones), and one between-subject factor was Group (musicians and non-musicians). In this analysis, we included the isolated condition and three context conditions in which the contexts’ F0 were kept unshifted (F0 unshifted context conditions), such that the targets’ F0 fell in the range of contexts’ F0. This analysis revealed whether the perception of Cantonese middle-level tone was regulated by preceding contexts. We were particularly interested in the comparison between the isolated condition and the other three context conditions as the isolated condition served as a benchmark indicating the response bias toward a middle-level tone when the target presented individually. According to Wong and Diehl (2003), the response rate of middle-level tone in the isolated condition could be around 50% across talkers. The context condition eliciting a different response rate than the isolated condition should inform us that the context provided useful information for listeners to adopt in normalizing the level tone. However, a lack of difference in F0 unshifted context conditions might not conclude that the context failed to support listeners’ lexical tone normalization, so a second analysis including the various F0 manipulations was performed, which focused on the contrastive context effect.

Following previous research on contrastive context effect (Wong and Diehl, 2003; Zhang et al., 2012, 2017), perceptual height (PH) and expected identification rate (IR) were analyzed to investigate participants’ lexical tone normalization performance. For the PH analysis, a response of high-level tone was coded as 6, middle-level tone as 3, and low-level tone as 1. This coding scheme reflected the acoustic difference among the three level tones and was straightforward when deciphering the results. The mean PH close to 6 indicated that participants generally perceived the targets as high-level tones. In an F0 lowered condition, this could serve as evidence of evoking participants’ tone normalization. The mean PH close to 1 indicated that

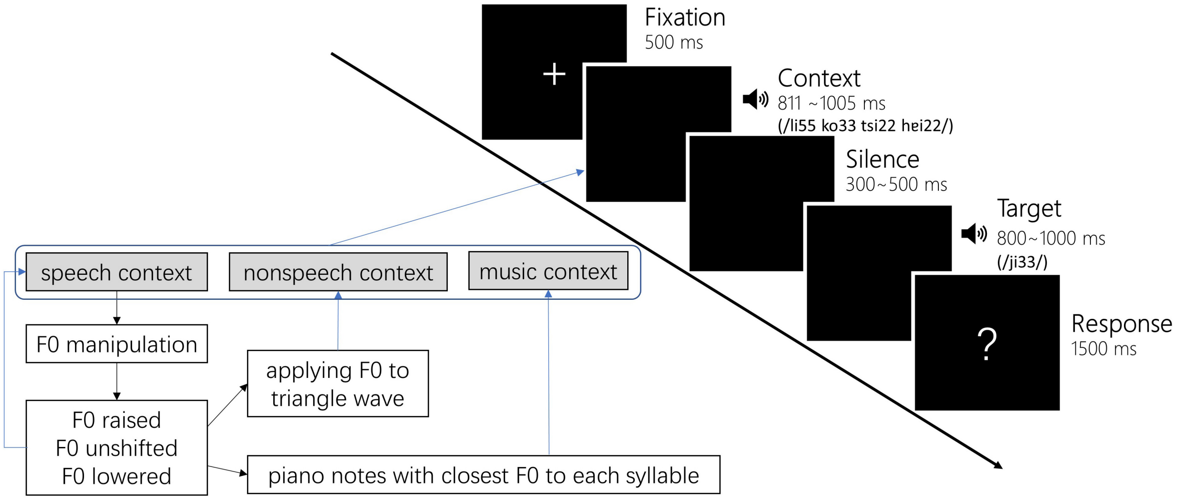


FIGURE 1 | Trial procedure of the word identification task and a schema depicting the context preparation.

participants generally perceived the targets as low-level tone. In an F0 raised condition, this could serve as evidence of evoking participants' tone normalization. The IR was the percentage of expected responses in each condition according to the contrastive context effect. The expected responses were the judgments that participants should make when the lexical tone normalization process was elicited, e.g., choosing low-level tone in the F0 raised condition, and choosing high-level tone in the F0 lowered condition. For targets following an F0 unshifted context, although there is no contrast between the context and target, it is expected that participants would perceive the target as a middle-level tone if the context provides sufficient information to reliably categorize the level tone.

We conducted three-way ANOVAs on PH and IR, where the isolated condition was excluded as it did not match the design matrix of other context conditions, e.g., there was no context and thus no F0 Shift manipulations. Two within-subject factors were Context (music, non-speech, speech) and Shift (F0 lowered, unshifted, raised), and one between-subject factor was Group (musicians, non-musicians). It is expected to see a contrastive context effect in speech context conditions and a lack of such an effect in non-speech context conditions. Following a speech-specific mechanism hypothesis, the music context conditions would not elicit a contrastive context effect, while the general-auditory mechanism hypothesis would expect music context to elicit a contrastive context effect, with higher magnitude seen in the musician group, e.g., an interaction between the three factors. The interaction among Context, Shift, and Group factors was most critical to the current study.

Previous research suggested that speech contexts produced by talkers with various F0 all elicited very high IR, while the specific pattern of responses was biased by the talkers' F0 (Zhang et al., 2013). For example, both female and male talkers with lower F0 elicited more low-level tone responses, and both female and male talkers with higher F0 elicited more high-level tone responses. The present study did not aim to follow up the discussion on the talker effect, nonetheless, the four talkers were included in the experiment to prevent response bias to a single talker and to increase the generalization of the results. Two talker-related factors, Gender (with two levels, female and male) and Pitch (with two levels, high pitch and low pitch) were included as control variables in all analyses for controlling their main effects and possible interactions with other factors (but see **Supplementary Figures S2, S3** for the illustrations of talker effects on PH and IR).

In all analyses, Greenhouse–Geisser correction was applied when the data violated the Sphericity hypothesis. Tukey method for comparing families of multiple estimates was applied for necessary *post hoc* analysis. The effect size of each significant main effect and interaction was reported in the form of general eta squared (η^2). The ANOVA procedure is robust for within-subject designs and used for analyzing data in this study to increase the comparability with previous research. However, we also performed a non-parametric version of ANOVA and found the results were highly similar (see **Supplementary Analysis**). All analysis was performed in R (version 4.0.5, R Core Team, 2021) with packages tidyverse (Wickham et al., 2019), rstatix (Kassambara, 2021), afex (Singmann et al., 2021), lsmeans (Lenth, 2016), and ggplot2 (Wickham, 2016) for data processing, statistics, and visualization.

RESULTS

Context Regulation on Targets' Response Rates

To evaluate how the context regulated participants' response rates on the three possible choices (high, middle, and low-level tones), ANOVA was performed on target response rates of F0 unshifted context conditions and isolated condition. Results revealed a main effect of Choice [$F(1.84,70.08) = 54.69$, $p < 0.001$, $\eta^2 = 0.226$]. The response rate of middle-level tone (mean ± SD = 52.5% ± 12.5%) was higher than other responses

(ps < 0.001), and response rate of low-level tone (30.1% ± 10.7%) was higher than high-level tone (17.7% ± 13.8%). There was an interaction between Context and Choice factors [$F(4.64, 176.47) = 26.54$, $p < 0.001$, $\eta^2 = 0.097$]. As in **Figure 2**, *post hoc* analysis revealed that the response rates of all three level tones following speech contexts were different from the isolated condition. The speech context yielded a higher response rate of middle-level tone than isolated condition (72.7% ± 21.2% vs. 47.6% ± 17.1%, $p < 0.001$), and lower responses rates of high- and low-level tones (11.6% ± 15.2% vs. 20.9% ± 19.4%, $p < 0.01$ and 15.7% ± 12.1% vs. 31.5% ± 14.4%, $p < 0.001$, respectively). The music and non-speech contexts, however, did not yield different response rates from isolated conditions (all ps > 0.1). The Group factor did not interact with other factors (ps > 0.4), suggesting that the above pattern was consistent across musicians and non-musicians. The results indicated that a speech context could regulate listeners' responses to targets: listeners had a higher rate of making the correct choice, i.e., choosing middle-level tone.

Extrinsic Normalization in Three Context Conditions

The lack of response rate differences from isolated conditions suggested that non-speech and music context could not facilitate tone categorization when the target F0 fell in the context's F0 range. However, the results could not conclude that non-speech and music context fail to facilitate tone categorization in a contrastive manner. Therefore, ANOVA was also performed on PH and IR, respectively. The isolated context condition was excluded from these analyses for not matching the other three context conditions on the Shift factor and not possible to elicit contrastive extrinsic normalization, e.g., there are no contexts and thus no F0 manipulations.

For the analysis on PH, a main effect of Context [$F(1.99,75.49) = 13.32$, $p < 0.001$, $\eta^2 = 0.040$] was found, with speech context yielded higher PH than non-speech and music contexts (ps < 0.01). There was no difference between PH yielded by non-speech and music contexts ($p = 0.341$). Additionally, a main effect of Shift was found [$F(1.21,46.13) = 240.46$, $p < 0.001$, $\eta^2 = 0.181$]. *Post hoc* analysis revealed that F0 lowered contexts yielded a higher PH than F0 unshifted contexts which was higher than F0 raised contexts (lower > unshifted > raised: 3.62 ± 0.61 > 2.89 ± 0.53 > 2.36 ± 0.58, all ps < 0.001). Not surprisingly, there was an interaction between the Context and Shift factors [$F(1.44, 54.80) = 216.33$, $p < 0.001$, $\eta^2 = 0.307$]. However, the decremental pattern of F0 manipulation (lowered > unshifted > raised) was only significant in speech contexts (5.28 ± 1.00 > 3.04 ± 0.47 > 1.48 ± 0.69, ps < 0.001), but not non-speech (2.90 ± 0.78, 2.88 ± 0.75, 2.87 ± 0.812) nor music (2.69 ± 0.76, 2.77 ± 0.74, 2.72 ± 0.76, all ps > 0.7) context conditions. The decrement of PH with F0 manipulation in speech context conditions is evident for the contrastive context effect, and thus indicate that listeners were elicited lexical tone normalization in the speech context conditions only but not in non-speech nor music context conditions.

It is also worth mentioning that the main effect of Group was not significant [$F(1,38) = 1.59$, $p = 0.214$, $\eta^2 = 0.009$]. Group factor did not interact with Context ($p = 0.928$) or Shift factors ($p = 0.879$), and there was not a three-way interaction among these factors ($p = 0.572$). Such a pattern indicated that musicians did not outperform non-musicians in any of the contexts with any kind of F0 manipulations (**Figure 3**).

The analysis on PH revealed that participants perceived targets as perceptually different tones only in speech contexts in both groups. To test whether participants behaved categorically following expectations of contrastive context effect, we performed ANOVA on their IR synthesized across each condition. The ANOVA on IR found a main effect of Context [$F(1.10,41.68) = 198.90$, $p < 0.001$, $\eta^2 = 0.365$], driven by that the speech context yielded a higher IR than non-speech and music contexts (ps < 0.001). The main effect of Shift was also significant [$F(1.67, 63.38) = 13.89$, $p < 0.001$, $\eta^2 = 0.058$]. The *post hoc* analysis revealed that both F0 unshifted (54.1% ± 12.4%) and raised (52.0% ± 12.8%) conditions yielded higher IR than F0 lower condition (38.7% ± 14.7%) (ps < 0.001), while the F0 unshifted and raised conditions yielded similar IR ($p = 0.794$). There was also an interaction between the Context and Shift factors [$F(3.47, 132.00) = 18.75$, $p < 0.001$, $\eta^2 = 0.054$], which was driven by the similarly high IRs yielded by F0 manipulations in the speech context conditions (lowered, unshifted, raised: 79.5% ± 27.1%, 72.7% ± 21.2%, 81.2% ± 22.2%, all ps > 0.1) and different but low IRs observed in non-speech and music contexts. In both non-speech and music contexts, the patterns of IRs yielded by F0 manipulations were similar: the F0 lowered (in non-speech and music: 20.4% ± 18.2% and 16.1% ± 15.6%) condition yielded a smaller IR than unshifted (44.6% ± 14.9% and 45.0% ± 13.3%) and raised (35.2% ± 16.3% and 39.7% ± 17.5%) conditions (all ps < 0.001), and the unshifted and raised conditions yielded similar IRs (ps > 0.05). The similarly high IRs observed in speech context with F0 manipulations indicated that listeners' performance followed the contrastive context effect's expectations: listeners are more likely to make a high-level tone judgment after hearing an F0 lowered speech context and a low-level tone judgment after hearing an F0 raised speech context. However, listeners, irrespective of non-musicians or musicians, did not show such a pattern in non-speech or music context conditions (**Figure 4**).

Regarding to IR, the Group factor did not show a main effect [$F(1,38) = 0.02$, $p = 0.890$, $\eta^2 < 0.001$] or any interactions with Context or Shift factors (ps > 0.2), and there was not a three-way interaction among the between and within-subject factors ($p = 0.211$). Echoing the results observed in the analysis of PH, such a pattern indicated that the musical training induced familiarity in music contexts did not facilitate listeners to take advantage of non-speech or music contexts for subsequent lexical tone normalization.

Explorative Analyses on the Effect of Piano Learning Experience on Extrinsic Normalization

A primary aim of this study was to compare musicians with non-musicians on their tone normalization performance following contexts with different familiarity. Surprisingly, the Group factor

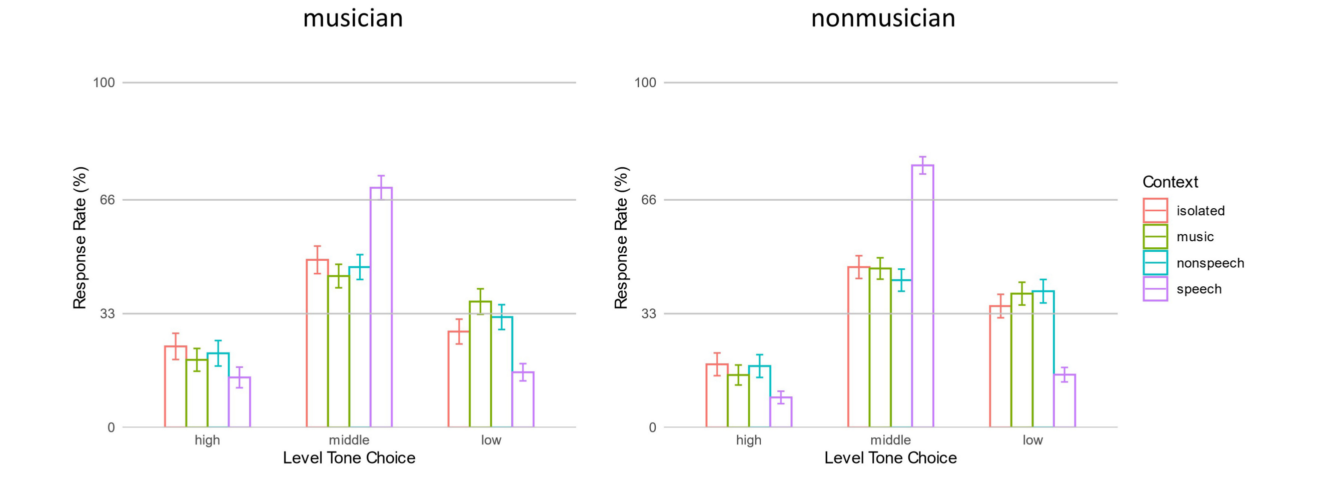


FIGURE 2 | The response rates of high, middle, and low-level tones in isolated and three context conditions. Error bars represent standard error.

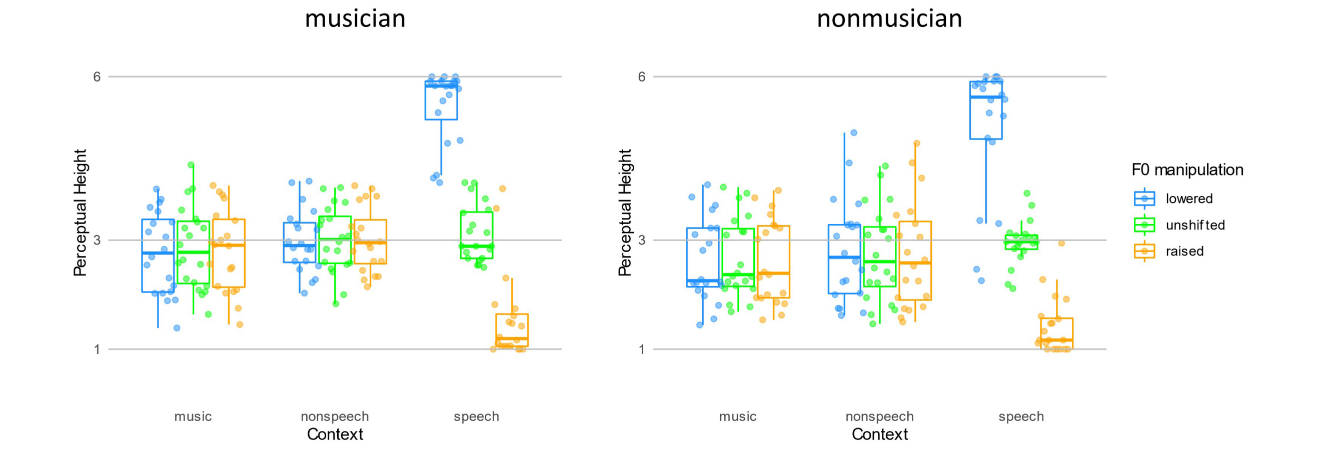


FIGURE 3 | The perceptual height of the targets following three contexts with F0 manipulations in two groups.

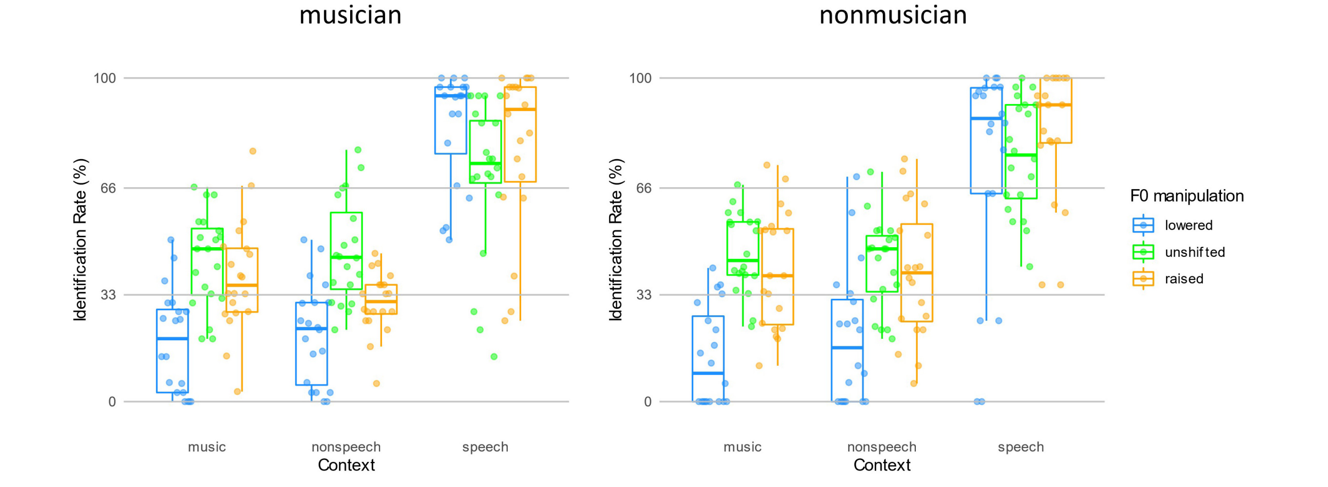


FIGURE 4 | The identification rate of the target following three contexts with F0 manipulations in two groups.

was not significant in the above analysis, and it did not show any interaction with other factors. One possible reason for such a result was that the musician group had a diverse musical background (**Supplementary Table S1**) and not all of them were equally familiar with piano notes used in the music context conditions. Thus, we sought to explore whether musicians with piano learning experiences could better facilitate them to adopt music contexts compared with others who did not learn piano before. As in **Supplementary Table S1**, 12 musicians reported piano learning experience and seven did not learn piano before the experiment. One participant failed to report her learned instrument and was excluded from the subsequent analysis. Within the 19 musicians, a repeated-measures ANOVA was performed to explore the effects of within-subject factors Context, Shift, and the between-subject factor piano learning experience (with two levels, yes and no, and coded as KnowPiano in the following text) on their PH and IR.

Critical to our interest, KnowPiano factor did not significantly influence musicians' PH [$F(1,17) = 2.73$, $p = 0.117$, $\eta^2 = 0.020$] or IR [$F(1,17) = 0.20$, $p = 0.661$, $\eta^2 < 0.001$]. KnowPiano factor did not interact with Context or Shift factors, either (all ps > 0.3), and there was not a three-way interaction ($p > 0.1$). The Context and Shift factors revealed similar main effects and interactions as the previous section, indicating that musicians, regardless of their piano learning experience, perceived targets following the expectation of contrastive context effect only in speech context conditions. This pattern was also consistent with the finding that musicians as a group did not outperform non-musicians in extrinsic tone normalization. The ANOVA tables are summarized in **Supplementary Table S2**.

DISCUSSION

In this study, we sought to clarify the possible interference of familiarity factor in listeners' speech normalization. The familiarity factor was manipulated on two dimensions with musical materials and musical training experience, respectively. Music is a commonly experienced non-verbal stimulus that has higher familiarity than rarely heard synthesized non-speech. In addition, a group of musicians with sufficient musical training was also included in the experiment, as their familiarity with music context was higher than the non-musician group. Previous studies conflict with each other on whether only speech context can provide valid information for listeners to adopt in mapping ambiguous acoustic signals to determinate linguistic units, possibly because the speech and non-speech contexts not only contrast in their linguistic features but also the listeners' familiarity with these contexts. According to the hybrid model (Tuller, 2003; Nguyen et al., 2009), the lack of non-speech exemplars in the lower layer due to unfamiliarity may account for the speech context superiority in speech normalization. Here we used a non-linguistic context, music, in addition to the conventionally used synthesized non-speech and speech contexts to probe listeners' lexical tone normalization in a word identification task. Our result showed that the music context did not trigger the lexical tone normalization process in either group. Musicians performed similarly to non-musicians in the music context. Additionally, musicians with piano training did not show a normalization advantage in the music context composed of piano notes compared with musicians without piano training. Overall, the results indicate that the familiarity factor does not interfere with listeners' lexical tone normalization.

Sjerps and Smiljanić (2013) and Magnuson et al. (2021) reported that the language familiarity and the talker familiarity did not facilitate the accommodation of speech variabilities (but see Lee et al., 2009 and Kang et al., 2016 for different results). By extending their works to the familiarity with different sounds (i.e., speech, music, triangle waves), the present study also failed to find the familiarity advantage in the lexical tone normalization process. Although familiarity advantage has been reported in several aspects of speech perception, for example, better speaker and word identification in noise (e.g., Johnsrude et al., 2013), familiarity might not directly affect the normalization process. Previous studies came up with different explanations for the absence of familiarity advantage in speech normalization. Sjerps and Smiljanić (2013) suggested that the normalization process occurred at the pre-phonemic level, and the acoustic cues were enough to normalize speech variability. Therefore, language familiarities that mainly differed at the phonological level showed no advantage in their studies. However, this explanation was not supported by the present study since we did not observe a reliable normalization effect in the non-speech and music contexts that provided acoustic cues. Magnuson et al. (2021) suggested that the talker normalization process which computes the talker-specific vocal tract characteristics probably overlaps with the talker recognition process. The talker familiarity advantage emerges only when the talker-specific speech characteristics were processed. That is, only when the listener recognized the identity of the talker, can they retrieve the stored exemplars or other mental representations of that talker. Further studies which explore the time course of taker identification and speech normalization should be conducted to test this hypothesis. Here we attempted to give another potential explanation for the absence of familiarity advantage in the lexical tone normalization process. Apart from more exemplars, the familiarity advantage could be aroused by the improved processing proficiency. The empirical studies supported that familiarity facilitates automatic face recognition and automatic speech prosody perception (Ylinen et al., 2010; Yan et al., 2017). Zhang et al. (2017) asked listeners to perform a Cantonese homophone judgment task while listening to speech or non-speech context in the normalization task. They found that the normalization results in both speech and non-speech contexts were not affected by the simultaneously ongoing secondary task. Although speech normalization probably is a cognitive-resource-dependent process (Nusbaum and Morin, 1992), Zhang et al. (2017) indicated that extracting information from both speech and non-speech context is automatic. The automatic extraction of the context pitch was almost not affected by familiarity, resulting in comparable results in music context and non-speech context, and between musicians and non-musicians.

The present study compared different contexts (speech, music notes, vs. triangle waves) and different groups of listeners (musicians vs. non-musicians). Neither dimension showed a familiarity advantage in Cantonese tone normalization, indicating that the speech superiority in Cantonese tone normalization is not due to familiarity but much more likely due to the speech-specific information in speech context. A previous study suggested that the richness of linguistic information influences the magnitude of tone normalization. The removal of semantic, phonological, and phonetic information gradually fails to elicit the contrastive context effect (Zhang et al., 2015). Other researchers hypothesized the speech-specific information enclosed in the talker-specific mapping of acoustic patterns onto linguistic units, and such a mapping is critical for tuning speech perception (Joos, 1948). Even the spectrally rotated non-speech that has more speechlike spectrotemporal dynamics could generate stronger normalization effects than the non-speech context without these speechlike properties (Sjerps et al., 2011). All these studies emphasized the importance of speech-specific (or at least speechlike) information in accommodating speech variability. As music notes and triangle waves in the present study contained no speech-specific information (even no speechlike spectrotemporal dynamics), the normalization process did not emerge. The necessity of the speech-specific information indicates that the successful lexical tone normalization process is largely operated via a speech-specific mechanism. It is worth noting that our findings could only conclude that the familiarity did not contribute to the final decision of the tone categorization. Future studies with a fine-grained temporal resolution (e.g., electrophysiological methods) may provide evidence on whether familiarity influences the early stages of normalization processing.

Aside from the speech-specific information, the coherence between context and target is another potential factor that leads to the speech-superiority in the normalization process. Speech context is more coherent with speech targets in many dimensions than music notes and triangle waves. This is partially supported by the congruency effect reported by Zhang et al. (2017). They found that the pitch height estimation of the non-speech target was better with the non-speech than speech context, and the lexical tone perception of the speech target was better with the speech than the non-speech context. Although the experimental design of the present study cannot tease apart the context-target coherence and the speech-specific information, the coherence hypothesis to some extent is in line with the domain-specific sound process, which in turn supports the speech-specific normalization process. Further studies which include a music context-music target and triangle wave context - triangle wave target could be ideal to test the context-target coherence hypothesis.

Although the studies from our group (e.g., Zhang et al., 2013, 2017) and other research groups (e.g., Francis et al., 2006) consistently revealed the necessity of the speech-specific information in normalizing Cantonese level tones, the normalization of other linguistic units showed mixed results. Huang and Holt (2009) compared the normalization of Mandarin T1 and T2 in the speech context and the non-speech contexts composed of either sine-wave harmonics or pure tone. They found the statistically comparable normalization effect between speech and two non-speech contexts. However, with an almost similar paradigm, Chen and Peng (2016) found the normalization effect in the speech context but not in the non-speech context (triangle waves). The results for the segmental normalization, for example, vowels, are also complex. Sjerps et al. (2011) found that the non-speech context could hardly help the normalization of vowels but once it shared some speech-like spectrotemporal features, the normalization effect emerged. Zhang K. et al. (2018) reported that although at the group level there was no normalization effect for the non-speech context, around half of the participants did show a contrastive context effect in the non-speech condition. These results suggest that non-speech contexts to some extent affect vowel normalization. The discrepancy between level tones and other speech units might come from the acoustic cues that contribute to their identification (Sjerps et al., 2018). The differentiation of level tones mainly depends on the pitch height and the contextual information is important to tell the relative pitch height. This special feature makes the Cantonese level heavily rely on contextual information. However, more than one cue affects the perception of vowels. The pitch and the formant pattern within the target syllable also contribute to the vowel identification (Johnson, 2005). Consequently, the vowel normalization probably relies less on the context information. Although the information in non-speech context is not as rich as that in speech context, it is enough to trigger successful vowel normalization. It is worth noting that no matter for segments or suprasegments, the normalization effect is salient in speech context and does not always appear in non-speech context. Therefore, although the speech-specific context information might not be indispensable for the normalization of phonemes containing rich acoustic cues (e.g., vowels), the superiority of speech context in the extrinsic normalization largely holds in speech perception.

To our knowledge, there was no study directly testing whether music experience facilitates lexical tone normalization. The present study uniquely and empirically probed into this question by comparing musicians and non-musicians. Musicians who are trained intensively in perceiving the fine pitch differences are expected to form a more precise pitch range reference to estimate the relative pitch height of the target lexical tone, and consequently, show a stronger contrastive context effect than non-musicians. However, musicians in the present study showed no advantage in the lexical tone normalization in the speech context, suggesting that there is almost no positive transfer in the pitch encoding from the music domain to the linguistic domain. This finding is somewhat in line with previous studies about musicians of tonal language speakers. Although non-tonal language speakers showed the music experience benefit in the lexical tone discrimination and identification (e.g., French speakers in Marie et al., 2011 and English speakers in Alexander et al., 2005), this benefit reduced a lot for tonal language speakers. Wu et al. (2015) and Zhu et al. (2021) reported that Mandarin musicians only showed the increased sensitivity to the within-category differences which was not important for the lexical tone categorization. Cooper and Wang (2012) found that either tone-language or music experience facilitated the lexical tone identification, but that the combination of two did not lead to better results than either experience alone. By investigating the extrinsic normalization of lexical tones, the present study extended these findings and showed that musicians of tonal language speakers had almost no observable advantage in identifying the relative pitch height in the linguistic domain. Zhang (2020) and Zhang and Peng (2021) found that the normalization process is largely implemented at the phonetic and phonological processing stages. Wu et al. (2015) and Zhu et al. (2021) suggested that acoustic processing was reliably enhanced even with limited musical training (4–5 years of amateur learning), but the musical training did not benefit the phonological processing. This might account for why musicians did not outperform non-musicians in the normalization task. Besides, musicians showed no normalization advantage in the non-verbal context (i.e., piano notes and triangle waves) as well. Although musicians have a more precise encoding of pitch information, the pitch extracted from non-speech contexts is still not enough for them to establish an effective talker-specific reference, indicating that additional speech-specific information is necessary for the successful normalization process (Zhang et al., 2015). Shao and Zhang (2018), Zhang C. et al. (2018), and Liu et al. (2021) consistently reported that amusics of tonal language speakers who are impaired in music pitch perception also show impaired lexical tone normalization, indicating a negative transfer from music pitch processing to linguistic pitch processing. By investigating the musicians' lexical tone normalization, the present study, however, failed to find a positive transfer from the music domain to the linguistic domain. It is reported that music and language share the similar acoustic parameters and the similar process at the lower level (i.e., the acoustic level), which leads to the observed positive/negative transfer across two domains (Patel, 2013). The accurate perception of the acoustic differences is the basis for the successful processing at the higher level (i.e., the phonological identification). Amusics who are impaired at perceiving the fine acoustic differences are less likely successful at the lexical tone normalization which is largely implemented at the phonetic and phonological level, especially when the demand for pitch sensitivity is high (Wang and Peng, 2014). Meanwhile, successful phonological identification requires acoustic differentiation ability, but the basic acuity shared by normal tonal language speakers (non-amusics) is enough (Cooper and Wang, 2012). This might be the reason why Cantonese musicians who have higher ability at telling fine acoustic differences performed equally well as Cantonese non-musicians in the lexical tone normalization.

CONCLUSION

In this study, to evaluate whether the familiarity factor mediates the speech superiority in lexical tone normalization, we compared musicians' and non-musicians' perception of Cantonese level tones in speech, music, and non-speech contexts. We found that despite two groups of participants showed clear contrastive context effect in speech context conditions, neither group showed such an effect in non-speech or music context conditions. The familiarity of music could not increase its usefulness in listeners' speech normalization, even it was longitudinally learned and practiced by listeners. Thus, our findings add more evidence to support the speech-specific mechanism in explaining the speech normalization process. The present study also found that even though musicians have the sophisticated pitch perception ability in music, their music experience does not boost the pitch height estimation in the linguistic domain as revealed by the comparable extrinsic normalization results of musicians and non-musicians.

DATA AVAILABILITY STATEMENT

The data that support the findings of the current study are available upon reasonable request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Human Subjects Ethics Sub-committee of The Hong Kong Polytechnic University. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

RT and GP conceived and designed the experiment. RT and KZ implemented the experiment, and collected and analyzed the data. All the authors interpreted the data, wrote the manuscript, and approved the submitted version.

FUNDING

This work was supported in part by the General Research Fund (No. 15607518) from the Research Grants Council (RGC) of Hong Kong.

ACKNOWLEDGMENTS

We thank Xue Chen, Chi Tat Li, Yaching Tsai, and Man Lok Wong for their great help in recruiting musicians.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.717110/full#supplementary-material

Table 1.DOCX | Supplementary analysis, supplementary figures, and supplementary tables.

Audio 1–8.WAV | Sound files of speech contexts and targets produced by four talkers (MH/ML/FH/FL).

REFERENCES

Alexander, J. A., Wang, P. C. M., and Bradlow, A. R. (2005). Lexical tone perception in musicians and non-musicians. *Paper Presented at the 9th European Conference on Speech Communication and Technology*, Lisbon, 397–400.

Ayotte, J., Peretz, I., and Hyde, K. (2002). Congenital amusia. A group study of adults afflicted with a music-specific disorder. *Brain* 125, 238–251. doi: 10.1093/brain/awf028

Baldwin, C. L. (2012). *Auditory Cognition and Human Performance Research and Applications*, ed. I. Ebrary Boca Raton, Fla: Taylor & Francis.

Bertrand, D. (1997). "Auditory neural processing of speech," in *The Handbook of Phonetic Sciences*, eds W. J. Hardcastle, and J. Laver (Oxford: Blackwell), 507–538. doi: 10.1177/002383098002300107

Boersma, P., and Weenink, D. (2016). *Praat: Doing Phonetics by Computer [Computer program]. Version 6.0.16*. Available online at: http://www.praat.org/ (accessed August 10, 2016).

Chen, F., and Peng, G. (2016). Context effect in the categorical perception of mandarin tones. *J. Signal Process. Syst.* 82, 253–261. doi: 10.1007/s11265-015-1008-2

Cooper, A., and Wang, Y. (2012). The influence of linguistic and musical experience on Cantonese word learning. *J. Acoust. Soc. Am.* 131, 4756–4769. doi: 10.1121/1.4714355

Duifhuis, H. (1973). Consequences of peripheral frequency selectivity for nonsimultaneous masking. *J. Acoust. Soc. Am.* 54, 1471–1488. doi: 10.1121/1.1914446

Francis, A. L., Ciocca, V., Wong, N. K. Y., Leung, W. H. Y., and Chu, P. C. Y. (2006). Extrinsic context affects perceptual normalization of lexical tone. *J. Acoust. Soc. Am.* 119, 1712–1726. doi: 10.1121/1.2149768

Gandour, J. (1983). Tone perception in Far Eastern languages. *J. Phon.* 11, 149–175. doi: 10.1016/s0095-4470(19)30813-7

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychol. Sci.* 16, 305–312. doi: 10.1111/j.0956-7976.2005.01532.x

Holt, L. L. (2006). The mean matters: effects of statistically defined nonspeech spectral distributions on speech categorization. *J. Acoust. Soc. Am.* 120, 2801–2817. doi: 10.1121/1.2354071

Hu, X., Wang, X., Gu, Y., Luo, P., Yin, S., Wang, L., et al. (2017). Phonological experience modulates voice discrimination: evidence from functional brain networks analysis. *Brain Lang.* 173, 67–75.

Huang, J., and Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *J. Acoust. Soc. Am.* 125, 3983–3994. doi: 10.1121/1.3125342

Johnson, K. (1990). Contrast and normalization in vowel perception. *J. Phon.* 18, 229–254. doi: 10.1016/s0095-4470(19)30391-2

Johnson, K. (1997). "Speech perception without speaker normalization: an exemplar model," in *Talker Variability in Speech Processing*, eds K. Johnson and J. W. Mullennix (Cambridge, MA: Academic Press), 145–166.

Johnson, K. (2005). "Speaker normalization in speech perception," in *The Handbook of Speech Perception Blackwell handbooks in linguistics*, eds D. B. Pisoni and R. E. Remez (Hoboken, NY: Blackwell Publishing), 363–389. doi: 10.1002/9780470757024.ch15

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., and Carlyon, R. P. (2013). Swinging at a cocktail party: voice familiarity aids speech perception in the presence of a competing voice. *Psychol. Sci.* 24, 1995–2004. doi: 10.1177/0956797613482467

Joos, M. (1948). Acoustic phonetics. *Language* 24, 1–136.

Kang, S., Johnson, K., and Finley, G. (2016). Effects of native language on compensation for coarticulation. *Speech Commun.* 77, 84–100. doi: 10.1016/j.specom.2015.12.005

Kassambara, A. (2021). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. R package version 0.7.0. Available online at: https://CRAN.R-project.org/package=rstatix (accessed February 13, 2021).

Kimball, A. E., Cole, J., Dell, G., and Shattuck-Hufnagel, S. (2015). "Categorical vs. episodic memory for pitch accents in english," in *Proceedings of the Intnational Congress Phonetic Science*, (Glasgow: University of Glasgow).

Ladefoged, P., and Broadbent, D. E. (1957). Information conveyed by vowels. *J. Acoust. Soc. Am.* 29, 98–104. doi: 10.1121/1.1908694

Lee, C. Y., Tao, L., and Bond, Z. S. (2009). Speaker variability and context in the identification of fragmented Mandarin tones by native and non-native listeners. *J. Phon.* 37, 1–15. doi: 10.1016/j.wocn.2008.08.001

Lenth, R. V. (2016). Least-squares means: the R package lsmeans. *J. Stat. Softw.* 69, 1–33. doi: 10.18637/jss.v069.i01

Liu, F., Yin, Y., Chan, A. H. D., Yip, V., and Wong, P. C. M. (2021). Individuals with congenital amusia do not show context-dependent perception of tonal categories. *Brain Lang.* 215:104908. doi: 10.1016/j.bandl.2021.104908

Lotto, A. J., and Kluender, K. R. (1998). General contrast effects in speech perception: effect of preceding liquid on stop consonant identification. *Percept. Psychophys.* 60, 602–619. doi: 10.3758/BF03206049

Lotto, A. J., Kluender, K. R., and Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *J. Acoust. Soc. Am.* 102, 1134–1140. doi: 10.1121/1.419865

Magnuson, J. S., and Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *J. Exp. Psychol.* 33, 391–409. doi: 10.1037/0096-1523.33.2.391

Magnuson, J. S., Nusbaum, H. C., Akahane-Yamada, R., and Saltzman, D. (2021). Talker familiarity and the accommodation of talker variability. *Atten. Percept. Psychophys.* 83, 1842–1860. doi: 10.3758/s13414-020-02203-y

Marie, C., Delogu, F., Lampis, G., Belardinelli, M. O., and Besson, M. (2011). Influence of musical expertise on segmental and tonal processing in Mandarin Chinese. *J. Cogn. Neurosci.* 23, 2701–2715. doi: 10.1162/jocn.2010.21585

Marques, C., Moreno, S., Castro, S. L., and Besson, M. (2007). Musicians detect pitch violation in a foreign language better than nonmusicians: behavioral and electrophysiological evidence. *J. Cogn. Neurosci.* 19, 1453–1463. doi: 10.1162/jocn.2007.19.9.1453

Moore, B. C. J. (1995). *Hearing: Handbook of Perception and Cognition*, 2nd Edn, ed. B. C. J. Moore Cambridge, MA: Academic Press.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *J. Acoust. Soc. Am.* 85, 2088–2113. doi: 10.1121/1.397861

Newman, R. S., Clouse, S. A., and Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *J. Acoust. Soc. Am.* 109, 1181–1196. doi: 10.1121/1.1348009

Nguyen, N., Wauquier, S., and Tuller, B. (2009). "The dynamical approach to speech perception: from fine phonetic detail to abstract phonological categories," in *Approaches to Phonological Complexity*, eds F. Pellegrino, E. Marsico, I. Chitoran, and C. Coupé (Berlin: Walter de Gruyter), 193–217. doi: 10.1515/9783110223958.191

Nusbaum, H. C., and Magnuson, J. S. (1997). "Talker normalization: phonetic constancy as a cognitive process," in *Talker Variability and Speech Processing, June* 2016, eds K.A. Johnson, and J.W. Mullennix (New York, NY: Academic Press), 109–132. doi: 10.1121/1.2028337

Nusbaum, H. C., and Morin, T. M. (1992). "Paying attention to difference among talkers," in *Speech Perception, Speech Production, and Linguistic Structure*, eds Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (Amsterdam: IOS Press), 113–134.

Nygaard, L. C., and Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Percept. Psychophys.* 60, 355–376. doi: 10.3758/BF03206860

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* 9, 97–113. doi: 10.1016/0028-3932(71)90067-4

Ong, J. H., Tan, S. H., Chan, A. H. D., and Wong, F. C. K. (2020). "The effect of musical experience and congenital amusia on lexical tone perception, production, and learning: a review," in *Speech Perception, Production and Acquisition. Chinese Language Learning Sciences*, eds H. Liu, F. Tsao, and P. Li (Singapore: Springer) doi: 10.1007/978-981-15-7606-5_8

Patel, A. D. (2013). "Sharing and nonsharing of brain resources for language and music," in *Language, Music, and the Brain*, ed. M. A. Arbib (Cambridge, MA: The MIT Press), 329–356. doi: 10.7551/mitpress/9780262018104.003.0014

Peng, G. (2006). Temporal and tonal aspects of Chinese syllables: a corpus-based comparative study of Mandarin and Cantonese. *J. Chin. Linguist.* 34, 134–154.

Peng, G., Deutsch, D., Henthorn, T., Su, D., and Wang, W. S.-Y. (2013). Language experience influences non-linguistic pitch perception. *J. Chin. Linguist.* 41, 447–467.

Peng, G., Zhang, C., Zheng, H., Minett, J. W., Territories, N., and Mandarin, C. (2012). The effect of intertalker variations on acoustic–perceptual mapping in cantonese and Mandarin tone systems. *J. Speech Lang. Hear. Res.* 55, 579–596. doi: 10.1044/1092-4388

Peterson, G. E., and Barney, H. L. (1952). Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* 24, 175–184. doi: 10.1121/1.1906875

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* 118, 2128–2148. doi: 10.1016/j.clinph.2007.04.019

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria. Available online at: https://www.R-project.org/ (accessed May 30, 2021).

Schön, D., Magne, C., and Besson, M. (2004). The music of speech: music training facilitates pitch processing in both music and language. *Psychophysiology* 41, 341–349. doi: 10.1111/1469-8986.00172.x

Shao, J., and Zhang, C. (2018). Context integration deficit in tone perception in Cantonese speakers with congenital amusia. *J. Acoust. Soc. Am.* 144, EL333–EL339. doi: 10.1121/1.5063899

Singmann, H., Bolker, B., Westfall, J., Aust, F., and Ben-Shachar, M. S. (2021). *afex: Analysis of Factorial Experiments.* R Package Version 0.28-1. Available online at: https://CRAN.R-project.org/package=afex (accessed July 22, 2021).

Sjerps, M. J., and Smiljanić, R. (2013). Compensation for vocal tract characteristics across native and non-native languages. *J. Phon.* 41, 145–155. doi: 10.1016/j.wocn.2013.01.005

Sjerps, M. J., Fox, N. P., Johnson, K., and Chang, E. F. (2019). Speaker-normalized sound representations in the human auditory cortex. *Nat. Commun.* 10:2465. doi: 10.1038/s41467-019-10365-z

Sjerps, M. J., Mitterer, H., and McQueen, J. M. (2011). Listening to different speakers: on the time-course of perceptual compensation for vocal-tract characteristics. *Neuropsychologia* 49, 3831–3846. doi: 10.1016/j.neuropsychologia.2011.09.044

Sjerps, M. J., Zhang, C., and Peng, G. (2018). Lexical tone is perceived relative to locally surrounding context, vowel quality to preceding context. *J. Exp. Psychol. Hum. Percept. Perform.* 44, 914–924. doi: 10.1037/xhp0000504

Stilp, C. (2020). Acoustic context effects in speech perception. *WIREs Cogn. Sci.* 11:e1517. doi: 10.1002/wcs.1517

Tao, R., and Peng, G. (2020). "Music and speech are distinct in lexical tone normalization processing," in *Proceeding of the 34th Pacific Asia Conference on Language, Information and Computation*, Hanoi.

Tuller, B. (2003). Computational models in speech perception. *J. Phon.* 31, 503–507. doi: 10.1016/S0095-4470(03)00018-4

Viswanathan, N., Magnuson, J. S., and Fowler, C. A. (2013). Similar response patterns do not imply identical origins: an energetic masking account of nonspeech effects in compensation for coarticulation. *J. Exp. Psychol.* 39, 1181–1192. doi: 10.1037/a0030735

Wayland, R., Herrera, E., and Kaan, E. (2010). Effects of musical experience and training on pitch contour perception. *J. Phon.* 38, 654–662. doi: 10.1016/j.wocn.2010.10.001

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer International Publishing.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., et al. (2019). Welcome to the tidyverse. *J. Open Source Softw.* 4:1686. doi: 10.21105/joss.01686

Wong, P. C. M., and Diehl, R. L. (2003). Perceptual normalization for inter- and intratalker variation in Cantonese level tones. *J. Speech Lang. Hear. Res.* 46, 413–421. doi: 10.1044/1092-4388(2003/034)

Wang, X., and Peng, G. (2014). Phonological processing in Mandarin speakers with congenital amusia. *J. Acoust. Soc. Am.* 136, 3360–3370. doi: 10.1121/1.4900559

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., and Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nat. Neurosci.* 10, 420–422. doi: 10.1038/nn1872

Wu, H., Ma, X., Zhang, L., Liu, Y., Zhang, Y., and Shu, H. (2015). Musical experience modulates categorical perception of lexical tones in native Chinese speakers. *Front. Psychol.* 6:436. doi: 10.3389/fpsyg.2015.00436

Yan, X., Young, A. W., and Andrews, T. J. (2017). The automaticity of face perception is influenced by familiarity. *Atten. Percept. Psychophys.* 79, 2202–2211. doi: 10.3758/s13414-017-1362-1

Ylinen, S., Uther, M., Latvala, A., Vepsäläinen, S., Iverson, P., Akahane-Yamada, R., et al. (2010). Training the brain to weight speech cues differently: a study of finnish second-language users of english. *J. Cogn. Neurosci.* 22, 1319–1332. doi: 10.1162/jocn.2009.21272

Zatorre, R. J., Chen, J. L., and Penhune, V. B. (2007). When the brain plays music: auditory–motor interactions in music perception and production. *Nat. Rev. Neurosci.* 8, 547–558. doi: 10.1038/nrn2152

Zhang, C., Peng, G., and Wang, W. S. Y. (2013). Achieving constancy in spoken word identification: time course of talker normalization. *Brain Lang.* 126, 193–202. doi: 10.1016/j.bandl.2013.05.010

Zhang, C., Peng, G., Wang, X., and Wang, W. S. (2015). "Cumulative effects of phonetic context on speech perception," in *Proceedings of the 18th International Congress of Phonetic Sciences*, (Glasgow: University of Glasgow).

Zhang, C., Shao, J., and Chen, S. (2018). Impaired perceptual normalization of lexical tones in Cantonese-speaking congenital amusics. *J. Acoust. Soc. Am.* 144, 634–647. doi: 10.1121/1.5049147

Zhang, K. (2020). *The Cognitive Mechanisms Underlying the Extrinsic Perceptual Normalization of Vowels*. Ph.D. thesis. Hung Hom: The Hong Kong Polytechnic University.

Zhang, K., and Peng, G. (2021). The time course of normalizing speech variability in vowels. *Brain Lang.* 222:105028. doi: 10.1016/j.bandl.2021.105028

Zhang, C., Peng, G., and Wang, W. S. (2012). Unequal effects of speech and nonspeech contexts on the perceptual normalization of Cantonese level tones. *J. Acoust. Soc. Am.* 132, 1088–1099. doi: 10.1121/1.4731470

Zhang, K., Sjerps, M. J., and Peng, G. (2021). Integral perception, but separate processing: the perceptual normalization of lexical tones and vowels. *Neuropsychologia* 156:107839. doi: 10.1016/j.neuropsychologia.2021.107839

Zhang, K., Sjerps, M. J., Zhang, C., and Peng, G. (2018). "Extrinsic normalization of lexical tones and vowels: beyond a simple contrastive general auditory mechanism," in *Proceedings of the TAL2018, Sixth International Symposium on Tonal Aspects of Language*, Berlin, 227–231, doi: 10.21437/tal.2018-46

Zhang, K., Wang, X., and Peng, G. (2017). Normalization of lexical tones and nonlinguistic pitch contours: implications for speech-specific processing mechanism. *J. Acoust. Soc. Am.* 141, 38–49. doi: 10.1121/1.4973414

Zhu, J., Chen, X., and Yang, Y. (2021). Effects of amateur musical experience on categorical perception of lexical tones by native chinese adults: an ERP study. *Front. Psychol.* 12:611189. doi: 10.3389/fpsyg.2021.611189

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

ORIGINAL RESEARCH
published: 26 January 2022
doi: 10.3389/fpsyg.2021.770425

Effects of Music Training on the Auditory Working Memory of Chinese-Speaking School-Aged Children: A Longitudinal Intervention Study

Peixin Nie[1,2,3], *Cuicui Wang*[3], *Guang Rong*[4], *Bin Du*[3], *Jing Lu*[3], *Shuting Li*[3], *Vesa Putkinen*[2,5,6], *Sha Tao*[3] *and Mari Tervaniemi*[1,2,7]*

[1] Cicero Learning, Faculty of Educational Sciences, University of Helsinki, Helsinki, Finland, [2] Cognitive Brain Research Unit, Faculty of Medicine, University of Helsinki, Helsinki, Finland, [3] State Key Laboratory of Cognitive Neuroscience and Learning and IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China, [4] HiperCog Group, Department of Education, Faculty of Educational Sciences, University of Helsinki, Helsinki, Finland, [5] Turku PET Centre, University of Turku, Turku, Finland, [6] Turku University Hospital, Turku, Finland, [7] Advanced Innovation Center for Future Education, Beijing Normal University, Beijing, China

Edited by:
Caicai Zhang,
The Hong Kong Polytechnic University, Hong Kong SAR, China

Reviewed by:
E. Glenn Schellenberg,
University of Toronto, Canada
Gavin M. Bidelman,
University of Memphis, United States

***Correspondence:**
Mari Tervaniemi
mari.tervaniemi@helsinki.fi

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

Received: *03 September 2021*
Accepted: *07 December 2021*
Published: *26 January 2022*

Citation:
Nie P, Wang C, Rong G, Du B, Lu J, Li S, Putkinen V, Tao S and Tervaniemi M (2022) Effects of Music Training on the Auditory Working Memory of Chinese-Speaking School-Aged Children: A Longitudinal Intervention Study. Front. Psychol. 12:770425. doi: 10.3389/fpsyg.2021.770425

Music expertise is known to be beneficial for cognitive function and development. In this study, we conducted 1-year music training for school children (*n* = 123; 7–11 years of age before training) in China. The children were assigned to music or second-language after-class training groups. A passive control group was included. We aimed to investigate whether music training could facilitate working memory (WM) development compared to second-language training and no training. Before and after the training, auditory WM was measured via a digit span (DS) task, together with the vocabulary and block tests of the Wechsler Intelligence Scale for Child IV (WISC-IV). The results of the DS task revealed superior development in the music group compared to the other groups. However, further analysis of DS forward and backward tasks indicated that the performance of the three training/non-training groups only differed significantly in DS backward scores, but not in the DS forward scores. We conclude that music training may benefit the central executive system of WM, as reflected by the DS backward task.

Keywords: second-language training, music, training effect, transfer, randomized controlled trial, propensity score method

INTRODUCTION

The effects of music expertise beyond music/sound-related skills have been increasingly investigated since the 1990s. Studies suggest that individuals with music exposure perform better in tasks measuring language abilities, such as foreign language pronunciation skills (Milovanov et al., 2010), phonological awareness (Linnavalli et al., 2018), and verbal intelligence (Moreno et al., 2012) than those who without music exposure. In addition to these transfer effects on linguistic function, associations between music exposure and higher-level cognitive abilities, which may indicate far-transfer effects, have also been reported, for example, in non-verbal intelligence (Schellenberg, 2006) and academic skills (dos Santos-Luiz et al., 2016).

Despite these findings, the existence and interpretation of these far-transfer effects remain unclear. One perspective is that music lessons may enhance general cognitive abilities (e.g., WM and executive functions), and these abilities may mediate the amount of benefit received from music lessons to music-unrelated performance (Hannon and Trainor, 2007; Moreno and Bidelman, 2014). In other words, the high demands on listening, attention, and controlling behavior during the music learning process may facilitate domain-general executive functions. Prior studies have provided numerous findings regarding the effects of music training on cognitive functions (working memory: Roden et al., 2012; attention: Strait et al., 2015; executive functions: Degé et al., 2011; Slevc et al., 2016; for review: Talamini et al., 2017). However, the findings are inconsistent to some extent across different studies and measurements. Talamini et al. (2017) summarized thirty-seven studies in their meta-analysis and revealed a small effect size (g = 0.29) for long-term memory, a medium effect size (g = 0.57) for short-term memory, and a medium effect size (g = 0.56) for WM. Sala and Gobet (2017), in their meta-analysis, also reported a small effect size (d = 0.34) of music training on memory-related abilities, and the effect size was even smaller if random allocation of participants was conducted (Sala and Gobet, 2020). Recently, Bigand and Tillmann (2021) repeated Sala and Gobet's (2020) analysis using the same data file, which resulted in stronger and more significant results. Here, we focus on music training's effects on auditory WM, which has been viewed as predictive of other cognitive functions, such as general fluid intelligence and cognitive flexibility (Kane et al., 2004; Blackwell et al., 2009). Some researchers have proposed that WM may play an essential role in mediating music training effects (Moreno and Bidelman, 2014). Unlike short-term memory, WM requires not only temporary storage, but also the processing and manipulation of information (Baddeley, 1992). According to Baddeley and Hitch's (1974) model, WM consists of two slave systems and a central executive system. The two slave systems—visual-spatial sketchpad and phonological loop—provide the fundamental basis for storing and maintaining visual-spatial and verbal-linguistic information, respectively. The central executive system reflects on domain-general processing and provides a certain workspace for ongoing information manipulation and other cognitive activities. WM refers to a wide range of information processing, including visual-spatial, verbal, and auditory WM. Consequently, many types of WM tests were developed based on the different types of WM and the three processing systems in Baddeley's (1992) model, which include, for example, the DS forward task, the DS backward task, the matrix span test, the Corsi block span test, the complex span test, and so on (Talamini et al., 2017).

The DS task (Wechsler, 2003) is a valid and commonly used test for measuring verbal WM from both storage and executive perspectives. The test consists of two parts: DS forward and DS backward. The former requires accurate repetition of a presented number sequence, which may represent the component phonological loop in the model. In contrast, the latter requires participants to repeat the numbers in reverse order and, therefore, requires further manipulation of the numbers and executive processing while storing them. Previous research has found evidence of the enhancement of both aspects in adult musicians and musically trained children when compared to untrained individuals. However, the literature is inconsistent regarding which component of the WM is enhanced in musically active individuals. For example, in Suárez et al. (2016), enhanced memory performance in musicians was found in the DS back task, reflecting central executive functions, but not in the DS forward task, reflecting the phonological loop. Similarly, Guo et al.'s (2018) training study of 6–8-year-old children found greater improvement in DS backward scores in the music group than in the control group. In contrast, Saarikivi et al. (2019) investigated the development of WM in children and adolescents aged nine through twenty and reported that musically trained participants outperformed their non-trained peers only on the DS forward test. Similar evidence has been found in other studies (Lee et al., 2007; Hansen et al., 2013). Schellenberg (2011) found that 9–12-year-old children who had music training obtained significantly higher DS total scores than children in the control group, while Virtala et al. (2014) reported that there were only marginally significant differences in DS total scores between adult musicians and non-musicians.

However, the limitations and inconsistencies of implementation were unavoidable in the reported studies. Some studies implemented interventions that may have been too short and, therefore, unable to observe the enhancement. For example, in Guo et al.'s (2018) study, the training sessions lasted for only 6 weeks; in Shen et al.'s (2019) study, the training program duration was 12 weeks. Some studies (Lee et al., 2007; Schellenberg, 2011; Suárez et al., 2016) were cross-sectional and directly compared musically trained and untrained children or adults, which may make it difficult to draw conclusions regarding causation. Furthermore, in some studies, the sample size was relatively small, so the results may not be generalized to a larger population (Fujioka et al., 2006; Virtala et al., 2014; Kumar and Krishna, 2019) or may lead to false-positive results (Button et al., 2013).

Language and music share similar cognitive demands, including auditory, somatosensory, visual, and cross-modal processing. Previous research has suggested that bilingualism also benefits one's executive functions, especially inhibitory control (Bialystok et al., 2004; Carlson and Meltzoff, 2008; Bialystok and DePape, 2009), as well as WM (Grundy and Timmer, 2017; Antón et al., 2019), although Alain et al. (2018) found that the effect of bilingualism on WM may be supported by different neural activities from those of music expertise. Antón et al. (2019) found that bilingual children outperformed monolinguals on the DS backward but not on the DS forward. The effects of bilingualism reflect a possible role of language processing on more general cognitive functions.

Addressing the limitations of the previous studies above, in our study, we used an randomized controlled trial (RCT) design and investigated the effects of music training on WM performance during a 1-year longitudinal training program in Beijing, China. Over 100 elementary school children were recruited and randomly allocated to music and second-language training groups. Language training was chosen as an active control to investigate the possible unique effect of music training on WM, apart from language learning. A passive control group was also included. To further balance the possible bias between the training groups that may result from dropouts, a propensity score method that is commonly used in medical experiments was applied for the data analysis. We aimed to investigate if and how 1 year of extracurricular group-based music training can benefit school-aged children's WM compared with language training and no training. In addition, we aimed to determine whether there were other more general effects of music training on children's cognitive development in terms of verbal and spatial skills.

MATERIALS AND METHODS

Participants

One hundred and nineteen children from 6 to 10 years of age were recruited at the first stage of the study and randomly assigned to the language (n = 60) or music (n = 59) groups. Nineteen children (fourteen boys) in the music group and seven (four boys) in the language group were unable to attend the courses due to scheduling conflicts. These twenty-six children, along with eleven newly recruited children from the same school, formed the passive control group (n = 37). Three children in the music group, three in the language group, and one in the passive control group voluntarily withdrew from the study. This resulted in 123 participants at the baseline stage: fifty in the language group, 37 in the music group, and 36 in the control group. All participants were native Chinese speakers. Twelve children (three in the music group, three in the language group, and six in the control group) failed to attend the training classes and take the post-training tests. Thus, there were 111 participants in the post-test: language group (n = 47), music group (n = 34), and control group (n = 30). In the analysis, outliers were defined as those with baseline scores for forward or backward DS tasks that were more than three standard deviations from the mean. Thus, three participants were identified: two had exceptional scores on the DS forward pre-test, and one had an exceptional score on the DS backward pre-test. Since only one had attended the post-test, 110 participants were included in the analysis: 46 in the language group (23 boys), 44 in the music group (eight boys), and thirty in the control group (twenty-two boys). **Table 1** shows additional descriptive statistics.

Parents provided written informed consent and were compensated for local transportation and time. - The study was approved by the Institutional Review Board at the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, and conducted in accordance with the norms of the Declaration of Helsinki.

Training Procedure

The training program was based on a large longitudinal study conducted in Beijing (Tervaniemi et al., 2021). The training sessions lasted for two semesters, during which the children received 50 1-h sessions of music/language training after their normal school curriculum. The curriculum of music training combined the Kodaly method with a well-established curriculum for basic knowledge of music, music theory, and solfeggio (Zhao, 2008), which includes fundamental rhythm and pitch skills, sight reading, and singing. Language training taught English as a second language, focusing on English word decoding, phonics, and vocabulary. Teaching materials included relevant textbooks: *Letter Land* (Wendon, 2009; Holt, 2011), *Root Phonics English* (Sun and Lytton, 2010), and *Pandeng English* (Pandeng English Project Team of State Key Laboratory of Cognitive Neuroscience and Learning at Beijing Normal University, 2012). This training protocol has been used in previous studies (Li et al., 2018; Xu et al., 2021). Teachers who were professionally trained in music and English language instruction at the master's level were hired for this project. A lead teacher always conducted the lesson in each class with an assistant teacher to help children with difficulties and assist with classroom management. During the last session of each semester, a Harvest Festival was held in each class to motivate children's learning in the classes; those who had studied diligently or performed well received a prize at the festival. **Supplementary Table 1** presents information about the fidelity check. The results of the fidelity check, in both the English and music training programs, showed good adherence to the teaching curriculum and plans (ratings above 4.8 on a 1–5 Likert scale). No group differences were observed in the most (first three) categories of fidelity ratings. However, the students' involvement in music classes was significantly better than in English classes.

Students' attendance was recorded; the average attendance was higher than 80% (**Table 1**). During the programs, the children were asked: "Did you generally like the sessions?" They answered using a 5-point Likert scale (1 = *I hate it*; 2 = *I don't like it*; 3 = *I don't know*; 4 = *I like it a bit*; 5 = *I like it very much*). In the music group, the mean score for this question was 4.3, and in the language group, it was 4.7, without significant group differences [$t(56.24) = 1.83$; $p = 0.072$; Cohen's $d = 0.488$].

Behavioral Measurements

Background Questionnaire

Demographic questions included the children's gender, age, parents' ages, and the family's socioeconomic status (SES). Family SES was based on the educational level of both parents, from none to doctoral level. The family's annual income was also reported by the parents. The data were further divided into two categories according to the respective median: higher family income (above CNY 100,000 annually; approximately USD 15,469) vs. lower family income (less than CNY 100,000) and a higher level of education (above high school) vs. lower education (up to high school). **Table 1** shows additional descriptive statistics.

Wechsler Intelligence Scale for Child IV

Three subtests were chosen from the Chinese version of the WISC-IV test (Zhang, 2009): DS, block design, and vocabulary. They were conducted before and after the training programs.

Digit span measures short-term auditory memory and WM. The test consisted of the forward DS and backward DS subtests. In the forward span task, the children were presented with a series of numbers and asked to repeat all the numbers in the same order. In the backward span task, the children also heard a series of numbers, but were asked to recall them in reverse order.

TABLE 1 | Descriptive statistics of the background variables in three groups at baseline.

	Music group **(*n* = 34)**	**Language group** **(*n* = 46)***	**Control group** **(*n* = 30)**	***df***	***F* or *t***	***p***
Age	8.75 ± 0.78	8.48 ± 0.82	8.57 ± 0.80	107	1.102	0.336
Attendance rate (%)	84.00 ± 20.45	86.65 ± 19.25	–	78	0.593	0.555
WISC-block design	10.26 ± 4.60	9.26 ± 4.31	9.97 ± 4.58	107	0.534	0.588
WISC-vocabulary	12.35 ± 3.32	11.61 ± 3.04	11.83 ± 2.94	107	0.571	0.566
				N	**X^2**	***p***
Father's education						
Higher level (*n*)	17 (50%)	19 (42.2%)	17 (56.7%)	109	1.541	0.463
Lower level (*n*)	17 (50%)	26 (56.5%)	13 (43.3%)			
Mother's education						
Higher level (*n*)	15 (44.1%)	23 (50%)	13 (43.3%)	109	0.579	0.749
Lower level (*n*)	19 (55.9%)	22 (47.8%)	17 (56.7%)			
Family income						
Higher income (*n*)	18 (52.9%)	19 (41.3%)	16 (53.3%)	109	1.258	0.533
Lower income (*n*)	16 (47.1%)	26 (57.8%)	14 (46.7%)			
Gender						
Boys (*n*)	8 (23.5%)	23 (50%)	22 (73.3%)	110	15.938	<0.0001
Girls (*n*)	26 (76.5%)	23 (50%)	8 (26.7%)			

**The SES scores of one participant in English group was missing, n = 45 when the variable is "family income", "father education" and "mother education".*

The number of correctly remembered trials was recorded as the original score for each forward span task and backward span task. The standardized total scores were calculated using Chinese norms (Zhang, 2009).

Block design measures children's spatial ability. Within a limited time, the participants were asked to assemble blocks to reproduce the given designs, matching the white-and-red design pictured. Each block has two red sides, two white sides, and two sides that are half white and half red. The original scores were the total number of trails in which the children successfully placed all the blocks within the limited time. The standardized total scores were calculated using Chinese norms (Zhang, 2009). The designs were arranged in order of increasing difficulty.

Vocabulary is an untimed verbal core subtest. The test measures verbal fluency, concept formation, and word knowledge and is comprised of twenty-five vocabulary words presented in order of increasing difficulty. The children were asked to explain the meaning of each word. The tasks stopped when the children failed to correctly explain the words. The original scores were obtained from the total number of correctly explained words. The standardized total scores were calculated using Chinese norms (Zhang, 2009).

Data Analysis Procedure

In a longitudinal research project with an intervention design, ideally, the groups of participants have balanced background variables to achieve the validity of the between-group comparison. This balance is usually a spontaneous subsequence when randomization is carried out on a sufficiently large sample. However, the small sample of 110 participants in this study (divided into two intervention groups and one control group), as well as the possibility of dropouts, may have caused the risk of imbalanced covariates. Thus, we adopted the propensity score (PS) method (Ho et al., 2007; Hansen and Bowers, 2008) to control for the participants' baseline characteristics. By using PS, this study was able to create balanced groups in which pairs of participants were similar, except for their experimental statuses, so that the main effect of the intervention could be unbiasedly estimated.

To calculate PS, logistic regression is usually used in this kind of analysis to predict the probability of being in the case group. Then, participants from one experimental group were matched with participants from the other groups on the magnitude of their scores to create groups with balanced covariates. However, research has shown that this procedure can be problematic in studies with small sample sizes. Holmes and Olsen (2010) examined three strategies for smaller sizes ($n = 112$) and recommended that using the PS score as a covariate be the optimal method for analyzing small sample data. Thus, we followed the recommended steps (Leite, 2016) as follows: (1) Estimate the scores—a multinomial logistic regression method was used for estimating the propensity scores. Age, gender as a dummy variable, WISC-block score, and WISC-vocabulary score were included as covariates in the model. The baseline measures were also entered in the model, including either the DS total score, DS forward score, or DS backward score. (2) Calculate the propensity weights—following Leite's (2016) approach, a propensity weight (PW) for each participant was further obtained by calculating the inverse of each PS. PWs were then assigned to the entire dataset. (3) Evaluation of covariate balance—pre-and post-experiment balancing of confounders between treatment groups, namely, the music group, language group, and passive control group, needs to be checked and reported in PS studies. We made a between-group pairwise comparison for each covariate and calculated the absolute standardized effect size and *p*-value. Effect sizes above 0.25 (Hansen, 2004) or *p*-values below 0.05 (Rubin, 2001) are considered a large imbalance of the covariate.

Finally, an analysis of covariate (ANCOVA) was conducted to estimate the group differences on post-test DS measures, with PW as the covariate (Holmes and Olsen, 2010). Based on previous research, we hypothesized that the effects of music training may differ on forward and backward tests (Saarikivi et al., 2019), so we analyzed the DS data separately for the forward and backward scores. Further multiple comparisons between groups were performed using the Bonferroni adjustment.

The data analysis was conducted in R (R Core Team, 2020). The *vglm* function in the "VGAM" package (Yee, 2020) and the *bal.stat* function in the "twang" package (Cefalu et al., 2021) were used to conduct multinomial logistic regressions and to assess the imbalance of the confounding variables, respectively. The function *emmeans* in the package "emmeans" (Lenth, 2020) was used for the *post hoc* test in the ANCOVA.

RESULTS

Table 1 shows the demographic variables for the three groups in terms of age at baseline, SES, baseline IQ, and attendance rate. No other significant differences were found in these variables among the three groups except for gender distribution—there were more boys in the control group than in the music group. This imbalance was caused by the selective participation of boys in the activities: despite random allocation, of the twenty-six participants who did not want to join the experimental (music, language) groups but who went to the passive control group, nineteen were boys.

We applied PS analysis to balance the bias from gender and other baseline measures. **Table 2** summarizes the pairwise covariate balance before and after adjusting with the PW from multinomial logistic regression, as well as the unadjusted balance in the baseline for DS total, DS forward, and DS backward, respectively. An effect size of 0.25 or greater is considered large (Hansen, 2004). As demonstrated in the table, the weights obtained with the multinomial logistic regression models provided a good covariate balance and were used in the final analysis.

Figure 1 shows a comparison of the DS scores among the three groups. The one-way ANCOVA was conducted to determine the group differences in the DS score in the post-test, after controlling for the propensity weights in the pre-test. Thus, the effects of the intervention can be estimated after controlling for prior differences. For the standardized DS total score, the results showed there was a significant group effect after controlling for the propensity weights in the pre-test [$F(2, 106) = 3.598$, $p = 0.031$]. However, the *post hoc* test showed that there was a significant difference between the music and passive control groups ($p = 0.029$) but no significant differences between the music and language groups or between the language and passive control groups ($p > 0.05$).

Next, we analyzed the outcome of the forward and backward subtests separately to reveal whether music training affected the phonological loop reflected by the forward subtest or the central executive system reflected by the backward subtest. As our main finding, we identified a discrepancy between the DS forward and backward raw scores as follows. For DS forward raw scores, the one-way ANCOVA showed no difference between the groups after controlling for the propensity weights in the pre-test [$F(2, 106) = 0.583$, $p = 0.560$]. In contrast, for DS backward raw scores, the one-way ANCOVA showed a significant group effect after controlling for propensity weights in the pre-test [$F(2, 106) = 5.038$, $p = 0.008$]. The music group outperformed the passive control group ($p = 0.013$) and the language group ($p = 0.039$); there were no differences between the language group and the passive control group ($p = 0.14$).

TABLE 2 | Summary of covariate balance.

	Group 1	Group 2	Method	Maximum standardized effect size	*P*-value
DS total	Music	English	Unadjusted	**0.58**	0.12
	Music	Control	Unadjusted	**1.36**	**0.01**
	English	Control	Unadjusted	**0.98**	0.18
	Music	English	MLR	0.12	0.63
	Music	Control	MLR	0.24	0.33
	English	Control	MLR	0.12	0.33
DS forward	Music	English	Unadjusted	**0.58**	**0.01**
	Music	Control	Unadjusted	**1.45**	0.25
	English	Control	Unadjusted	**1.3**	0.63
	Music	English	MLR	**0.26**	0.33
	Music	Control	MLR	**0.41**	0.15
	English	Control	MLR	0.16	0.15
DS backward	Music	English	Unadjusted	**0.7**	**0.02**
	Music	Control	Unadjusted	**1.23**	**<0.001**
	English	Control	Unadjusted	**0.67**	0.46
	Music	English	MLR	0.21	0.46
	Music	Control	MLR	0.18	0.45
	English	Control	MLR	0.1	0.45

1. GBM, general boosted model; MLR, multinomial logistic regression; 2. Numbers in ***bold*** *indicates the standardized effect size over 0.25 or p-value below 0.05 indicating imbalance on those covariates. Background variables include age, attendance rate, SES, gender, and general baseline cognitive abilities. For SES, "higher level" is defined as the reported education level or family income was higher than the median; "lower level" is defined as the reported education level or family income was lower than the median.*

The other two measures—block subtest scores and vocabulary subtest scores—were analyzed with the same procedure as PS analysis. The one-way ANCOVA with the propensity weights as covariates showed that there were no significant group differences on either the block subtest score [$F(2, 106) = 0.464$, $p = 0.630$], or the vocabulary subtest score [$F(2, 106) = 0.593$, $p = 0.554$].

DISCUSSION

The aim of our study was to investigate the effects of music training on auditory WM in school-aged children. The results revealed different effects of interventions, namely music training, language training, and no training, on the performance of DS tasks. On the general performance of the DS task, the musically trained group showed significant superiority compared to the control group after controlling for prior bias before the training

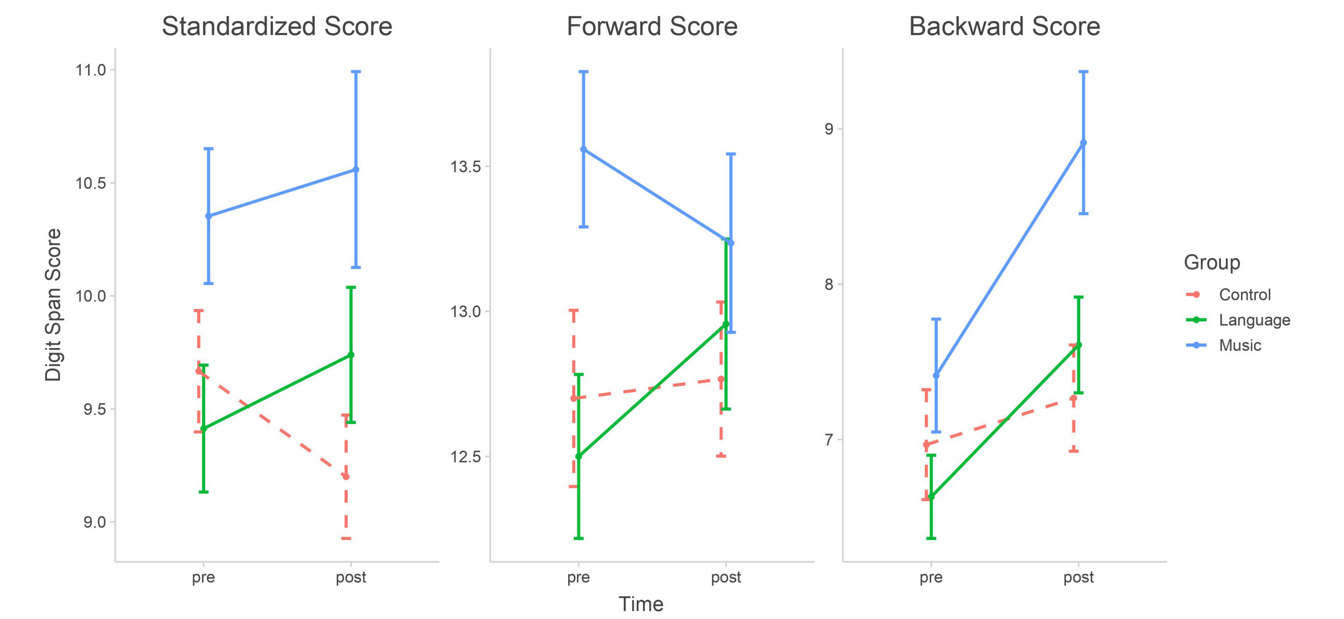


FIGURE 1 | Comparisons of digit span scores between groups (Mean and SE). Music group gained significant improvement compared with Language and Control group in digit span backward scores. However, no significant interaction between Group and Time was found in the digit span forward scores or digit span standardized score.

and the baseline level of the DS performance. However, this superiority was observed only in the DS backward performance. Regarding the DS forward performance, no such difference was found between the groups.

This result is in line with previous research indicating that DS forward and DS backward reflect different cognitive functions. Reynolds (1997), using factor analysis, found that forward and backward tasks indicate two distinct memory processes. Furthermore, in a study investigating attention deficits and DS performance, only DS backward scores predicted children with attention deficit hyperactivity disorder, while the DS forward task did not (Rosenthal et al., 2006). Our results support the view that DS forward and DS backward are distinct, measuring different cognitive processes—DS forward involves short-term auditory memory processes, whereas DS backward involves additional components of attention and executive functions.

Our results show that music training may be more beneficial for attention and executive memory processes, which is indicated by enhanced DS backward scores. This supports previous findings of positive associations of music expertise with the DS backward task (Guo et al., 2018) and higher cognitive functions, such as WM (Roden et al., 2014; D'Souza et al., 2018) and other executive functions (Degé et al., 2011; Saarikivi et al., 2016; Jaschke et al., 2018; Shen et al., 2019).

Notably, the negative results of the DS forward test were discrepant with previous findings. George and Coch (2011) found that DS forward scores were positively correlated with years of music training. Accordingly, Saarikivi et al. (2019) found that musically trained children and adolescents outperformed their untrained peers in DS forward but not DS backward tasks. They argued that music training may benefit WM, specifically in retaining and reproducing auditory sequences rather than in updating information in the mind. However, in the current study, music training did not produce a significant improvement in maintaining information indexed by the DS forward tasks.

One possible reason may be that this results from having a language background than in the majority of the literature—the participants in previous studies were speakers of non-tonal languages, whereas in the present study the spoken language is Chinese mandarin regarded as tonal language.[1] Bidelman et al. (2013) found that speakers of Cantonese, a tonal language, outperformed speakers of non-tonal languages on the tonal memory task, in which participants were asked to judge whether the probe tone was present in a four-tone sequence they had heard before.

In this study, the digit sequences in Mandarin, which is a tonal language, always have the same tones, and these tones may sound like melodies to children. The daily experience of listening and speaking melodic sentences may equip children with better auditory memory than the non-tonal language speakers, even without music training. While the performance of the DS forward task consequently benefited from the tonal melodies created by the digit sequences, the children might have already possessed a good level of memory for the DS forward, and music training may not be beneficial comparatively. The DS forward score in the music group might have dropped slightly because of the random fluctuation in the children's performances. However,

[1] Here, tonal language (Chinese, Vietnamese, Thai, etc.) is one in which the same series of sounds can have different semantic meanings depending on the tones (pitch) of the word. In contrast, in non-tonal languages (English, Spanish, etc.), the word's meaning is not influenced by pitch.

when the task was to list numbers in reverse order for the DS backward task, this melodic cue of the digit sequences was no longer helpful.

Another difference between our earlier findings and the literature can be found in the type of music training. While the training in this study was group-based and given as extracurricular lessons to schoolchildren, in previous studies, the musically trained participants were involved in instrumental training programs. Consequently, the discrepancy in the results may be explained by different demands of the given training; individual lessons emphasized fine-grained auditory functions, while group-based lessons in our study focused on acquiring music knowledge—for example, the recognition and classification of rhythm patterns and melodies, as well as interactions with teachers and peers. Thus, attention and executive functions might be practiced more than in other programs.

Next, we discuss the limitations of our study. Our initial purpose was to randomly assign the children to groups. However, there was a high dropout rate before the onset of the training program—several children dropped out of classes because of "scheduling conflicts." This might have led to an initial group difference before the training in the DS task but interestingly not in the block design and vocabulary task. It turns out that motivation and other environmental "hidden factors," such as school achievements and parents' personalities and parenting styles, may become critical barriers during random assignments (Schellenberg, 2020). When there was a weak commitment from the participants, those less motivated tended to choose other activities instead of staying in the classes.

However, if this issue were considered, what would happen if the less-motivated children were forced to stay and participate in the music group lessons? In addition to being unethical, it might still lead to an imbalance in motivation across groups, which could also impact the training effect. Moreover, some researchers have argued that randomization and the inference of causality are complicated. The group difference might still be present because of either gene-influenced individual differences or environmental factors, even if they were absent before the training (Schellenberg, 2020). Therefore, while solving the practical challenges of random assignment in a study, more factors, such as individual and familial background, should also be considered during the design, observation, and analysis processes of a training study in children.

In sum, we found that group-based music training enhanced children's auditory WM in terms of the executive system, as indexed by the DS backward test. In contrast, there was no evidence of the enhancement of simple storage of the digit WM, as indexed by the DS forward, resulting from music training. This could be due to the native tonal language background of the children, which may help their phonological storage with or without music training. To conclude, our results indicate that music training may enhance children's ability to manipulate information as a higher-order cognitive process, but not their simple storage capacity of auditory information.

DATA AVAILABILITY STATEMENT

The statistical data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board at the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

MT and ST designed the research plan. PN, CW, BD, SL, and JL monitored the training programs and conducted the research under the mentorship of ST. PN analyzed the data, wrote the initial draft of the manuscript, and prepared the data figures. GR helped with the data analysis and results reporting. All authors contributed to the revision of the manuscript and accepted the final version of it.

FUNDING

The Open Fund Project of the State Key Laboratory for Cognitive and Neuroscience, the Beijing Brain Initiative of Beijing Municipal Science and Technology Commission (Z181100001518003), the National Science Foundation of China (71834002), the 111 Project (BP0719032), the Beijing Normal University, and the University of Helsinki supported the study.

ACKNOWLEDGMENTS

We wish to thank Ms. Fei Xie, Huimin Yu, Yanmei Pan, and Yun Zhai for their help with data collection; Ms. Yuhua Chen, Weijuan Wang, Liyang Fan, Yiling You, Ge Li, Xizi Yang, Lin Zhang, Lingkai Zhang, Yichen Fan, Wei Wei, Yu Li, Jiaming Zhou, and Yiyang Wu for their help with training and material preparation; Ms. Hui Li, Mr. Jianchun Lv, and Ms. Runli Li for their help with participant recruitment and coordination.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.770425/full#supplementary-material

REFERENCES

Alain, C., Khatamian, Y., He, Y., Lee, Y., Moreno, S., Leung, A. W., et al. (2018). Different neural activities support auditory working memory in musicians and bilinguals. *Ann. N. Y. Acad. Sci.* 1423, 435–446. doi: 10.1111/nyas.13717

Antón, E., Carreiras, M., and Duñabeitia, J. A. (2019). The impact of bilingualism on executive functions and working memory in young adults. *PLoS One* 14:e0206770. doi: 10.1371/journal.pone.0206770

Baddeley, A. (1992). Working memory. *Science* 255, 556–559.

Baddeley, A. D., and Hitch, G. (1974). Working memory. *Psychol. Learn. Motiv.* 8, 47–89.

Bialystok, E., and DePape, A. M. (2009). Musical expertise, bilingualism, and executive functioning. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 565–574. doi: 10.1037/a0012735

Bialystok, E., Craik, F. I., Klein, R., and Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: evidence from the Simon task. *Psychol. Aging* 19:290. doi: 10.1037/0882-7974.19.2.290

Bidelman, G. M., Hutka, S., and Moreno, S. (2013). Tone language speakers and musicians share enhanced perceptual and cognitive abilities for musical pitch: evidence for bidirectionality between the domains of language and music. *PLoS One* 8:e60676. doi: 10.1371/journal.pone.0060676

Bigand, E., and Tillmann, B. (2021). Near and far transfer: is music special? *Mem. Congit.* Online ahead of print. doi: 10.3758/s13421-021-01226-6

Blackwell, K. A., Cepeda, N. J., and Munakata, Y. (2009). When simple things are meaningful: working memory strength predicts children's cognitive flexibility. *J. Exp. Child Psychol.* 103, 241–249. doi: 10.1016/j.jecp.2009.01.002

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376. doi: 10.1038/nrn3475

Carlson, S. M., and Meltzoff, A. N. (2008). Bilingual experience and executive functioning in young children. *Dev. Sci.* 11, 282–298. doi: 10.1111/j.1467-7687.2008.00675.x

Cefalu, M., Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., and Burgette, L. (2021). *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups. R Package Version 2.4.*

D'Souza, A. A., Moradzadeh, L., and Wiseheart, M. (2018). Musical training, bilingualism, and executive function: working memory and inhibitory control. *Cogn. Res. Principles Implications* 3:11. doi: 10.1186/s41235-018-0095-6

Degé, F., Kubicek, C., and Schwarzer, G. (2011). Music lessons and intelligence: a relation mediated by executive functions. *Music Percept.* 29, 195–201. doi: 10.1111/j.2044-8295.2011.02008.x

dos Santos-Luiz, C., Mónico, L. S. M., Almeida, L. S., and Coimbra, D. (2016). Exploring the long-term associations between adolescents' music training and academic achievement. *Musicae Sci.* 20, 512–527.

Fujioka, T., Ross, B., Kakigi, R., Pantev, C., and Trainor, L. J. (2006). One year of musical training affects development of auditory cortical-evoked fields in young children. *Brain* 129, 2593–2608. doi: 10.1093/brain/awl247

George, E. M., and Coch, D. (2011). Music training and working memory: an ERP study. *Neuropsychologia* 49, 1083–1094. doi: 10.1016/j.neuropsychologia.2011.02.001

Grundy, J. G., and Timmer, K. (2017). Bilingualism and working memory capacity: a comprehensive meta-analysis. *Second Lang. Res.* 33, 325–340.

Guo, X., Ohsawa, C., Suzuki, A., and Sekiyama, K. (2018). Improved digit span in children after a 6-week intervention of playing a musical instrument: an exploratory randomized controlled trial. *Front. Psychol.* 8:2303. doi: 10.3389/fpsyg.2017.02303

Hannon, E. E., and Trainor, L. J. (2007). Music acquisition: effects of enculturation and formal training on development. *Trends Cogn. Sci.* 11, 466–472. doi: 10.1016/j.tics.2007.08.008

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *J. Am. Stat. Assoc.* 99, 609–618. doi: 10.1198/016214504000000647

Hansen, B. B., and Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Stat. Sci.* 23, 219–236.

Hansen, M., Wallentin, M., and Vuust, P. (2013). Working memory and musical competence of musicians and non-musicians. *Psychol. Music* 41, 779–793.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* 15, 199–236. doi: 10.1093/pan/mpl013

Holmes, W., and Olsen, L. (2010). "Using propensity scores with small samples," in *Proceedings of the Annual Meetings of the American Evaluation Association* (San Antonio, TX).

Holt, L. (2011). *Letterland Beyond ABC*. Cambridge: Letterland International Press.

Jaschke, A. C., Honing, H., and Scherder, E. J. (2018). Longitudinal analysis of music education on executive functions in primary school children. *Front. Neurosci.* 12:103. doi: 10.3389/fnins.2018.00103

Kane, M. J., Tuholski, S. W., Hambrick, D. Z., Wilhelm, O., Payne, T. W., and Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *J. Exp. Psychol. Gen.* 133, 189–217. doi: 10.1037/0096-3445.133.2.189

Kumar, P. V., and Krishna, R. (2019). Exploring music induced auditory processing differences among vocalists, violinists and non-musicians. *Int. J. Health Sci. Res.* 9, 13–21.

Lee, Y. S., Lu, M. J., and Ko, H. P. (2007). Effects of skill training on working memory capacity. *Learn. Instruction* 17, 336–344.

Leite, W. (2016). "Propensity score methods for multiple treatments," in *Practical Propensity Score Methods Using R* (Thousand Oaks, CA: Sage Publications).

Lenth, R. (2020). *emmeans: Estimated Marginal Means, aka Least-Squares Means. R package Version 1.4.7.*

Li, S., Tao, S., Joshi, R. M., and Xu, Q. (2018). Second-language reading difficulties among native Chinese-speaking students learning to read English: the roles of native-and second-language skills. *Read. Res. Q.* 53, 423–441.

Linnavalli, T., Putkinen, V., Lipsanen, J., Huotilainen, M., and Tervaniemi, M. (2018). Music playschool enhances children's linguistic skills. *Sci. Rep.* 8:8767. doi: 10.1038/s41598-018-27126-5

Milovanov, R., Pietilä, P., Tervaniemi, M., and Esquef, P. A. (2010). Foreign language pronunciation skills and musical aptitude: a study of Finnish adults with higher education. *Learn. Ind. Differ.* 20, 56–60.

Moreno, S., and Bidelman, G. M. (2014). Examining neural plasticity and cognitive benefit through the unique lens of musical training. *Hear. Res.* 308, 84–97. doi: 10.1016/j.heares.2013.09.012

Moreno, S., Ellen Bialystok, R. B., Schellenberg, E. G., Cepeda, J. N., and Chau, T. (2012). Short-term music training enhances verbal intelligence and executive function. *Psychol. Sci.* 22, 1425–1433. doi: 10.1177/0956797611416999

Pandeng English Project Team of State Key Laboratory of Cognitive Neuroscience, and Learning at Beijing Normal University (2012). *Pandeng English Reading Series*. Beijing: Beijing Normal University Publishing House.

R Core Team (2020). *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reynolds, C. R. (1997). Forward and backward memory span should not be combined for clinical analysis. *Arch. Clin. Neuropsychol.* 12, 29–40.

Roden, I., Grube, D., Bongard, S., and Kreutz, G. (2014). Does music training enhance working memory performance? findings from a quasi-experimental longitudinal study. *Psychol. Music* 42, 284–298.

Roden, I., Kreutz, G., and Bongard, S. (2012). Effects of a school-based instrumental music program on verbal and visual memory in primary school children: a longitudinal study. *Front. Psychol.* 3:572. doi: 10.3389/fpsyg.2012.00572

Rosenthal, E. N., Riccio, C. A., Gsanger, K. M., and Jarratt, K. P. (2006). Digit Span components as predictors of attention problems and executive functioning in children. *Arch. Clin. Neuropsychol.* 21, 131–139. doi: 10.1016/j.acn.2005.08.004

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv. Outcomes Res. Methodol.* 2, 169–188. doi: 10.1023/A:1020363010465

Saarikivi, K. A., Huotilainen, M., Tervaniemi, M., and Putkinen, V. (2019). Selectively enhanced development of working memory in musically trained children and adolescents. *Front. Integr. Neurosci.* 13:62. doi: 10.3389/fnint.2019.00062

Saarikivi, K., Putkinen, V., Tervaniemi, M., and Huotilainen, M. (2016). Cognitive flexibility modulates maturation and music-training-related changes in neural sound discrimination. *Eur. J. Neurosci.* 44, 1815–1825. doi: 10.1111/ejn.13176

Sala, G., and Gobet, F. (2017). Does far transfer exist? negative evidence from chess, music, and working memory training. *Curr. Dir. Psychol. Sci.* 26, 515–520. doi: 10.1177/0963721417712760

Sala, G., and Gobet, F. (2020). Cognitive and academic benefits of music training with children: a multilevel meta-analysis. *Mem. Cogn.* 48, 1429–1441. doi: 10.3758/s13421-020-01060-2

Schellenberg, E. G. (2006). Long-term positive associations between music lessons and IQ. *J. Educ. Psychol.* 98, 457–468.

Schellenberg, E. G. (2011). Examining the association between music lessons and intelligence. *Br. J. Psychol.* 102, 283–302.

Schellenberg, E. G. (2020). "Music training, individual differences, and plasticity," in *Educational Neuroscience: Development Across the Lifespan*, eds M. S. C. Thomas, D. Mareschal, and I. Dumontheil (Milton Park: Routledge).

Shen, Y., Lin, Y., Liu, S., Fang, L., and Liu, G. (2019). Sustained effect of music training on the enhancement of executive function in preschool children. *Front. Psychol.* 10:1910. doi: 10.3389/fpsyg.2019.01910

Slevc, L. R., Davey, N. S., Buschkuehl, M., and Jaeggi, S. M. (2016). Tuning the mind: exploring the connections between musical ability and executive functions. *Cognition* 152, 199–211. doi: 10.1016/j.cognition.2016.03.017

Strait, D. L., Slater, J., O'Connell, S., and Kraus, N. (2015). Music training relates to the development of neural mechanisms of selective auditory attention. *Dev. Cogn. Neurosci.* 12, 94–104. doi: 10.1016/j.dcn.2015.01.001

Suárez, L., Elangovan, S., and Au, A. (2016). Cross-sectional study on the relationship between music training and working memory in adults. *Aust. J. Psychol.* 68, 38–46.

Sun, K. R., and Lytton, K. (2010). *Root Phonics English, Nanchang, P. R.* China: Jiangxi People's Press.

Talamini, F., Altoè, G., Carretti, B., and Grassi, M. (2017). Musicians have better memory than nonmusicians: a meta-analysis. *PLoS One* 12:e0186773.

Tervaniemi, M., Putkinen, V., Nie, P., Wang, C., Du, B., Lu, J., et al. (2021). Improved auditory function caused by music versus foreign language training at school age: is there a difference? *Cereb. Cortex* 32, 63–75. doi: 10.1093/cercor/bhab194

Virtala, P., Huotilainen, M., Partanen, E., and Tervaniemi, M. (2014). Musicianship facilitates the processing of Western music chords—an ERP and behavioral study. *Neuropsychologia* 61, 247–258. doi: 10.1016/j.neuropsychologia.2014.06.028

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children*, 4th Edn. San Antonio, TX: The Psychological Corporation.

Wendon, L. (2009). *Letterland ABC.* Cambridge: Letterland International Press.

Xu, Q., Tao, S., Li, S., Wang, W., Li, B., and Joshi, R. M. (2021). Who are the nonresponders to intervention among Chinese children learning English as a second language? *J. Educ. Psychol.* 113, 213–229.

Yee, T. W. (2020). The VGAM package for negative binomial regression. *Aust. N. Z. J. Stat.* 62, 116–131. doi: 10.1111/anzs.12283

Zhang, H. (2009). The revision of WISC-IV Chinese version. *Psychol. Sci.* 32, 1177–1179.

Zhao, Y.-S. (2008). "Music theory and solfeggio," in *The Examination Center, Ministry of Education, and the Central Conservatory of Music in China* (Beijing: People's Music Press).

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Neuroscience

SYSTEMATIC REVIEW
published: 27 January 2022
doi: 10.3389/fnins.2021.753356

Melodic Intonation Therapy on Non-fluent Aphasia After Stroke: A Systematic Review and Analysis on Clinical Trials

Xiaoying Zhang[1,2,3,4], *Jianjun Li*[1,2,3,4,5]* *and Yi Du*[6,7]

[1] School of Rehabilitation Medicine, Capital Medical University, Beijing, China, [2] Beijing Key Laboratory of Neural Injury and Rehabilitation, China Rehabilitation Research Center, Beijing, China, [3] Center of Neural Injury and Repair, Beijing Institute for Brain Disorders, Beijing, China, [4] Department of Psychology, Music Therapy Center, China Rehabilitation Research Center, Beijing, China, [5] Chinese Institute of Rehabilitation Science, Beijing, China, [6] Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences (CAS), Beijing, China, [7] Departments of Psychology, University of Chinese Academy of Sciences, Beijing, China

Melodic intonation therapy (MIT) is a melodic musical training method that could be combined with language rehabilitation. However, some of the existing literature focuses on theoretical mechanism research, while others only focus on clinical behavioral evidence. Few clinical experimental studies can combine the two for behavioral and mechanism analysis. This review aimed at systematizing recent results from studies that have delved explicitly into the MIT effect on non-fluent aphasia by their study design properties, summarizing the findings, and identifying knowledge gaps for future work. MIT clinical trials and case studies were retrieved and teased out the results to explore the validity and relevance of these results. These studies focused on MIT intervention for patients with non-fluent aphasia in stroke recovery period. After retrieving 128 MIT-related articles, 39 valid RCT studies and case reports were provided for analysis. Our summary shows that behavioral measurements at MIT are excessive and provide insufficient evidence of MRI imaging structure. This proves that MIT still needs many MRI studies to determine its clinical evidence and intervention targets. The strengthening of large-scale clinical evidence of imaging observations will result in the clear neural circuit prompts and prediction models proposed for the MIT treatment and its prognosis.

Keywords: melodic intonation therapy, music therapy, non-fluent aphasia, speech therapy, stroke

OPEN ACCESS

Edited by:
Mary Miu Yee Waye,
The Chinese University of Hong Kong,
Hong Kong SAR, China

Reviewed by:
Xiaoyu Xia,
The Military General Hospital of Beijing
PLA, China
Francesco Riganello,
Sant'Anna Crotone Institute, Italy

****Correspondence:***
Jianjun Li
lintaozixyz_1228@hotmail.com;
crrclijj@163.com

Specialty section:
This article was submitted to
Auditory Cognitive Neuroscience,
a section of the journal
Frontiers in Neuroscience

Received: *04 August 2021*
Accepted: *07 December 2021*
Published: *27 January 2022*

Citation:
Zhang X, Li J and Du Y (2022) Melodic Intonation Therapy on Non-fluent Aphasia After Stroke: A Systematic Review and Analysis on Clinical Trials.
Front. Neurosci. 15:753356.
doi: 10.3389/fnins.2021.753356

INTRODUCTION

Aphasia is a language disorder generally caused by stroke-related damage to the dominant hemisphere. It describes a multitude of acquired language impairment as a consequence of brain damage (Go et al., 2014; WHO, 2015; Benjamin et al., 2017; Koleck et al., 2017). In relation to localization, it is possible to make a division between fluent and non-fluent aphasia. Oral expression of non-fluent aphasia is characterized by low speech volume, lack of grammar, and pronunciation dysphonia. According to the American Hearing Language Association's classification of aphasia, the types of non-fluent aphasia include motor aphasia, complete aphasia, transcortical motor aphasia, and transcortical mixed aphasia (Kim et al., 2016; Gerstenecker and Lazar, 2019; Hoover, 2019). According to the survey data from WHO on stroke prevention in 2019, about 2.6 to 4.7 million people suffer from stroke-related aphasia yearly, significantly impacting their quality of life (WHO, 2015; Wang et al., 2019). Aphasia affects the patient's linguistic skills and daily communication. As the course of the disease is prolonged, it will also impede patients' quality of life.

Due to the lack of targeted surgery and efficacious treatment regimens, speech therapy is a general method to train patients with aphasia. The mechanism of speech therapy is mainly based on the language function centers located in the dominant left hemisphere (Kamath et al., 2019). Several studies have demonstrated that music therapy for non-fluent aphasia is used to treat patients who have lost their speaking ability after a stroke or accident. It is reported that the right hemispheric regions are more active during singing (Jeffries et al., 2003; Callan et al., 2006; Ozdemir et al., 2006). Music therapy involving melodic elements is deemed to be a potential treatment for non-fluent aphasia, as singing potentially activates patients' right hemisphere to compensate for their lesioned left hemisphere (Zipse et al., 2009; Schlaug et al., 2010). Aside from singing, many other music therapy techniques have also been attempted, and the effectiveness of some methods has been revealed.

Melodic intonation therapy (MIT) is one of the verified effective methods of aphasia by the American Academy of Neurology (AAN) (Helm-Estabrooks and Albert, 2004). MIT is an intonation-based treatment method for non-fluent or dysfluent aphasic patients developed in response to the observation that severely aphasic patients can often produce well-articulated, linguistically accurate words while singing but not during a speech (Albert et al., 1973; Sparks et al., 1974). MIT is a hierarchically structured treatment that utilizes intoned (sung) patterns that exaggerate the typical melodic content of speech across three levels of increasing difficulty. At the elementary level, patients need to complete 1–2 syllables of melodic intonation in oral expression, such as "hello," "thank you," "goodbye," etc. At the intermediate level, patients need to complete oral expressions of melodic intonation of 3–5 syllables, such as "I love you," "I am thirsty (hungry)," "I have to rest," etc. to express daily needs. At the advanced level, patients need to express sentences of 6–10 words or more, such as "I am going to train today," "It is 10 a.m. in the morning," etc. The original explanation of MIT is to utilize the musical and language output region in the right hemisphere, in which the mechanism differs from the left hemisphere (Albert et al., 1973; Sparks et al., 1974). An assumption raised by Albert and Sparks is that music can be effective by discovering music to language connections between the right and the left hemisphere in an interactive way or by using either reserved music/language functional area in the two hemispheres to speak. MIT combines melodic and rhythmic aspects of sentence intonation with language (Albert et al., 1973; Sparks et al., 1974; Sparks and Holland, 1976; Helm-Estabrooks et al., 1989; Cohen and Masse, 1993; Boucher et al., 2001; Norton et al., 2009). It can mobilize the auditory musical area on the right and the language area in the left hemisphere. The goal of MIT is namely to elicit the sound of the language (or spontaneous speech) by exaggerating the melody and rhythm of the language. The implementation process of MIT is musical, activating the right hemisphere mechanism that is not commonly used in daily language expression.

However, according to the currently published MIT studies, there is an excessive number of reviews and mechanism analysis studies. Still, there is a scarce number of randomized controlled trials (RCT), cross-over studies, cohort studies, and case studies. In experimental researches, the evidence is accentuated over language behavior measurements, and there are very few studies that use multimodal imaging observation to verify behavioral, neural mechanisms. In the assessment results of the language scale, the brain areas observed by MRI imaging are scattered, and the target areas of symptoms remain unclear. According to the results of existing mechanism analysis and scale evaluation, there are many possible narratives for the mechanism of MIT, but its underlying mechanism remains unclear as of yet (Breier et al., 2010; van de Sandt-Koenderman et al., 2010; Merrett et al., 2014; Zumbansen et al., 2014b). Therefore, the purpose of our review is to (1) retrieve the evidence and effectiveness of MIT for non-fluent aphasia after strokes, determine the superior performance of melody intonation therapy-related interventions in behavioral measurement results, and summarize our findings. (2) From a meager amount of MRI evidence, determine which areas the onset mechanism is more focused in, identify more targeted brain areas and circuits, and find a more feasible mechanism direction for the treatment of aphasia by MIT, thus providing the groundwork for future research.

MATERIALS AND METHODS

Selection of Studies

We have planned and analyzed literature from reviews, systematic reviews, randomized controlled trials (RCT), clinical-controlled trials (CCT), cross-over studies, cohort studies, self-control, and case studies, regarding aphasia and music therapy. A literature search was conducted on four electronic databases: PubMed, Bing Scholar, Google Scholar, and Medline. The included articles are in English, French, Italian, Spanish, German, Korean, and Japanese. The publication timeframe was from January 1970 to July 2021. The keywords of "stroke," "aphasia," "music," "melody," "rhythmic," "intonation," "melodic intonation therapy," "music therapy," "music and aphasia," and "rhythm and aphasia" were searched. The search was free and followed PRISMA's recommendations (Liberati et al., 2009; Higgins and Green, 2011), with a reference list of articles attached.

Randomized controlled trials (RCT), clinical-controlled trials (CCT), cross-over design, self-control, and case studies were subsequently recruited, with the omission of reviews. In accordance with the PICOS principle in evidence-based medicine, this review defines the criteria for inclusion.

(1) *Participants*: In participant's inclusion, all studies concerned only human adults (≥18 years) in stroke recovery period with non-fluent aphasia, including ischemic and hemorrhagic stroke, and the time since stroke was more than 2 weeks. (2) *Intervention*: The intervention group followed musical supported MIT such as melodic intonation therapy (MIT), modified MIT, rhythmic syllables therapy (RST), spoken

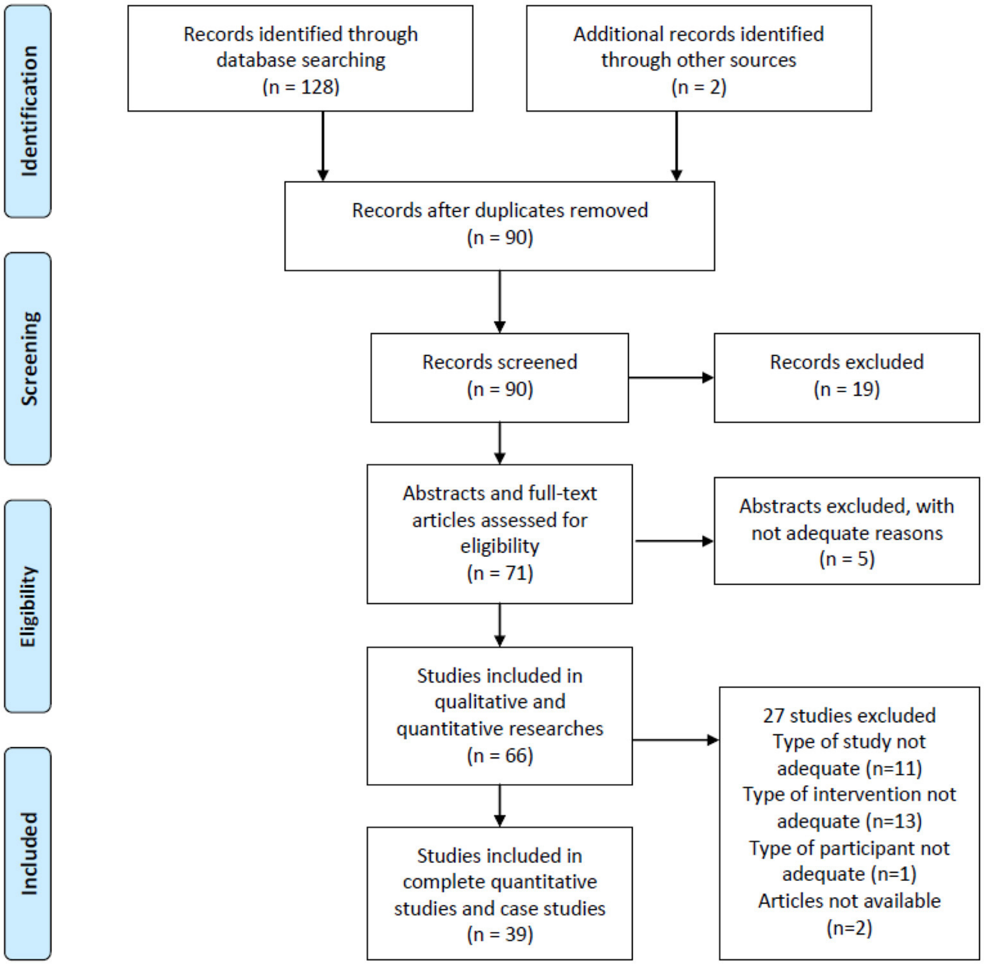

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 128)
Additional records identified through other sources (n = 2)
Records after duplicates removed (n = 90)
Records screened (n = 90)
Records excluded (n = 19)
Abstracts and full-text articles assessed for eligibility (n = 71)
Abstracts excluded, with not adequate reasons (n = 5)
Studies included in qualitative and quantitative researches (n = 66)
27 studies excluded
Type of study not adequate (n=11)
Type of intervention not adequate (n=13)
Type of participant not adequate (n=1)
Articles not available (n=2)
Studies included in complete quantitative studies and case studies (n = 39)

FIGURE 1 | Flow diagram of article identification and inclusion.

language stimuli, singing therapy (ST), rhythmic therapy (RT), prosody perception task (PPT), sung-spoken story recall task, melodic cueing, melodic singing, and rhythmic cueing. (3) *Comparison:* The MIT intervention dose ranges from 1 to 4 times per week, and the duration ranges from 1 to 12 weeks. The control group was followed by speech therapy or blank control in the same dose and duration. (4) *Outcomes*: Using behavioral evaluation scales and fMRI to evaluate the results, the primary outcomes with a $p < 0.5$ are meaningful. (5) *Study design*: Methods are a randomized controlled study of MIT and speech therapy, or a self-controlled study of MIT, or a cross-design study of modified MIT and speech therapy, case reports of MIT, etc.

We compared speech therapy and melodic intonation therapy, combined with commonly integrated rehabilitation, and evaluated clinical outcomes.

Data Sources and Search Strategy

After searching for relevant literature in 4 databases, a total of 128 works of literature about melodic intonation (induced) therapy were retrieved, and 2 was from another website. It was found that 90 articles were repeated in each database after reviewing the titles, indicating high reliability. After a quick review of the literature, 10 MIT literature reviews, 5 abstracts, and 4 qualitative analyses were excluded. The remaining 71 articles contain complete quantitative analysis and case studies. After careful examination of these articles, it was found that the data of 5 brief articles were published as spotlight, and 11 papers were presented as the original form; without statistical analysis, the statistical correlation could not be obtained. Thirteen articles did not belong to melodic intonation therapy and relative therapy. Three papers were for patients with a cognitive impairment not relevant to aphasia intervention. Finally, 39 quantitative experimental types of research and case studies of aphasia rehabilitation of typical MIT were identified. The risk of bias assessment was based on the handbook of Cochrane review methods (Higgins and Green, 2011; **Figure 1**).

The principle behind selecting these clinical studies as a systematic review is that these studies have applied MIT to clinical patients to observe the actual effects. Secondly, internationally standard measures were performed before and after the clinical trial to compare the results. Eleven of the experiments were accompanied by imaging tests. The intervention was a melodic musical form, accomplished by the therapist. Therefore, the above three points align with the therapy standards and principles proposed at the beginning of standard melodic intonation therapy.

MAIN RESULTS OF CLINICAL TRIALS OF MIT TO APHASIA

This review summarizes all MIT studies with non-fluent aphasia patients since 1970 (**Table 1**). Since MIT was established in the 1970s as a more effective supplementary treatment for non-fluency aphasia, clinical trials on MIT have gradually garnered widespread attention. MIT clinical trials have the following characteristics: (1) In the research before the Twentieth century, the behavioral observation records of patients with MIT were more detailed; (2) Comparative case studies, self-controlled studies, and small sample experiments were more numerous; (3) Most of them used subjective language assessment scales for result evaluation. After the Twentieth century, with the advancement of imaging medicine, researchers conducted large sample experiments while focusing on behavioral measurements. They were more concerned about the evidence yielded by the brain imaging structure. The assessment tool was taken as a classification feature. Twenty-two MIT clinical trials evaluated using language ability scales and 11 clinical trials using imaging measurements; all the 33 pieces of research are listed in **Table 1**.

The Effects of MIT on 15–40 Sample Trials: The Most Assessment Tools Are Subjective Measurement Scales

In these MIT clinical trials, 13 experiments use various language assessment scales for evaluation, accounting for the majority. Melodic interventions are the selected essential factors, but evaluation criteria are equally important. There are mainly two evaluation criteria in the quantitative studies, one is various standard language test scales, which include the Boston Diagnostic Aphasia Examination (BDAE), the Western Aphasia Battery (WAB) in different language versions, the Aphasia Quotient (AQ), the Aachen Aphasia Test (AAT), amongst others. The other is imaging check, which includes functional magnetic resonance imaging (fMRI), magnetic resonance (MR), and diffusion tensor imaging (DTI), which are usually applied in a one-time assessment.

RCT Studies Evaluated Using Standard Language Test Scales Showed Consistent Results in Behavioral Measurement Results (Without Imaging)

Of the more than 15 MIT RCT studies selected in this review, seven valuable clinical trials used the language assessment scale to evaluate the results. Conklyn et al. (2012), Lim et al. (2013), Van der Meulen et al. (2014), Van Der Meulen et al. (2016), Raglio et al. (2015), Kasdan and Kiran (2018), Haro-Martínez et al. (2019), Leo et al. (2019), and Zhang et al. (2016, 2021) all used various language scales to assess two groups of patients with aphasia. The results demonstrated that whether it was only one observation of the immediate treatment effect or the cumulative treatment effect for up to 12 weeks, compared with the speech therapy group, the MIT group was better in understanding (Haro-Martínez et al., 2019), retelling (Haro-Martínez et al., 2019), and oral task response time (Lim et al., 2013), and oral memory time and retelling phrase length (Kasdan and Kiran, 2018) have been markedly improved. Regarding spontaneous expression, most of the target languages trained by MIT are short sentences of varying lengths, while the content of melody training is fixed. Therefore, in addition to improving the level of training items, patients receiving MIT can also enhance the spontaneous speech of untrained items. This is particularly conspicuous in the test of story retelling (Van der Meulen et al., 2014; Van Der Meulen et al., 2016). These meaningful behavioral measurement results are reflected in the scores of different dimensions of various language test scales.

Among the specific results, Haro-Martínez et al. (2019) found that after MIT, the MIT group improved communicative activity log (CAL), but no significant difference was noted in comprehension and repetition. Leo et al. (2019) found that after singing melody in the MIT group, the aphasic patients recalled longer in the singing rather than the speaking task and also with chunk length in the singing task. Kasdan and Kiran (2018) compared 1-h immediate effect after MIT and then found that patients with standard MIT conspicuously improved phrase length. Zumbansen et al. (2014a) conducted a crossover study on 3 aphasia patients for 6 weeks to compare MIT. The results showed that all of the 3 patients in MIT training improved clarity of syllables significantly. Stahl et al. (2013) did a similar crossover study of 3 aphasia patients, and it turns out the MT group improved significantly in repetition. In 2014, Van der Meulen et al. (2014) and Van Der Meulen et al. (2016), conducted an MIT crossover study on 27 aphasia patients, among which 16 patients received MIT for 6 weeks, and 11 patients in the control group received MIT after weeks. It was revealed that compared to the control group, the MIT group improved the repetition (AAT) in both trained items and untrained items. He then ran the same MIT crossover study in 2016 and found MIT group improved repetition in trained items and spontaneous sentences in untrained items. Raglio's et al. RCT study (Conklyn et al., 2012; Lim et al., 2013; Raglio et al., 2015) also proved that MIT improved repetition, listening comprehension, spontaneous speech, naming, and responsive items 2–3 score. Vian (Vines et al., 2011) turns out that applying anodal-tDCS during MIT produced a significantly greater improvement in verbal fluency.

Case Studies and Small Sample Studies Have the Characteristics of Complete Specific Treatment Interventions

There are 6 clinical trials with sample sizes between 1 and 6 patients. These studies mostly use the patient's control or crossover design to observe the effectiveness of MIT intervention. Due to the small sample size, these studies reflect the characteristics of a more detailed record of the intervention process and a more evident division of music elements in MIT. In the MIT intervention conducted by Van der Meulen et al. (2012) for 2 patients, a dedicated MIT therapist carried out the implementation process. Although in the MIT study

TABLE 1 | The clinical trials of MIT for the aphasia of stroke.

Trial	Design	Sample	Participants	Music therapist involved	Blinding	Intervention	Duration and dose	Primary outcome	Measures	Main results
Researches measured by scales (26)										
Sparks et al. (1974)	Self-control	$n = 8$	8 Patients with severely impaired verbal output; good auditory comprehension; global aphasics; No improvement in verbal output for at least 6 months	No	No	Melodic intonation therapy (MIT) Programme	Daily therapy 3 months	Significant results	Boston Diagnostic Aphasia Examination (BDAE) (Goodglass and Kaplan, 1972)	Responsive naming–$F = 25.3$, df 1, 5; $P = 0.005$ Confrontation naming–$F = 7.9$, df 1, 5; $P = 0.038$ Phrase length–$F = 29.6$, df 1, 5; $P = 0.003$ Auditory commands–$F = 2.2$, df 1, 5; $P = 0.198$ Complex auditory material–$F = 0.8$, df 1, 5; $P = 0.396$ Reading comprehension, sentences-paragraphs–$F = 0.3$, df 1, 5; $P = 0.50$
Marshall and Holtzapple (1976)	Case study	$n = 4$	Patient 1: male; age: 49-years-time since stroke: 9 months Patient 2: male; age: 49-years-old time since injury: 25 months Patient 3: male; age: 53-years-old time since injury: 1 months Patient 4: male; age: 41-years-old time since injury: 1 months dignosis of left middle cerebral artery thrombosis with aphasia and severe apraxia.	No	No	Melodic intonation therapy (MIT) Modified melodic intonation therapy (MMIT)	60 min/day 3 days/week 3 months no return to normal speech	Significant results	Porch Index of Communication Ability (PICA) Overall Communicative Ability (OCA score)	Patient 1 has increased in PICA scores at 3 and 6 months post-MIT, and improved articulatory skills Patient 2 shows an increase of 8 scores (34–42) on the PICA Patient 3 and Patient 4 showed an improvement in verbal modality after MMIT
Goldfarb and Bader (1979)	Case study	$n = 1$	1 patient, male, 50-year-old with severe global aphasia after C for 10 years (left frontal thrombosis)	No	No	Melodic intonation therapy (MIT)	1 h/session 6 sessions/week 1 time in hospital 5 times at home	Significant results	Boston diagnostic aphasia examination (BDAE)	78% correct in speech mode; 84% correct in intonation without tapping mode; 92% correct in tonation plus tapping mode; and 88% correct without any hints
Popovic and Boniver (1992)	Self-control	$n = 80$	80 patient diagnosis with Broca aphasics and bucco-lingual apraxi	No	No	Melodic intonation therapy (MIT) in Romanian	60–120 min/session 7 sessions/week 2–4 weeks	Significant results	Romania aphasia test (RST)	MIT was considered an efficient method in the early stages of Broca aphasia with bucco-lingual apraxia

(Continued)

TABLE 1 | Continued

Trial	Design	Sample	Participants	Music therapist involved	Blinding	Intervention	Duration and dose	Primary outcome	Measures	Main results
Researches measured by scales (26)										
Springer et al. (1993)	Cross-over	$n = 12$	Group 1 (L/S Group) $n =$ age: 46.17 ± 12.84 years time since stroke: 14.17 ± 14.11 months Group 2 (S/L Group) $n = 6$ age: 41.83 ± 9.72 years time since stroke: 26.5 ± 22.75 months	No	No	Melodic intonation therapy (MIT): MIT's facilitation technique used in a different therapeutic program MIT: linguistically oriented approach (L) MIT: stimulation approach (S)	60 min/day 3–4 days/week 2 weeks	Significant results	Aachen aphasia test (AAT) Token test (TT) Communicative abilities in daily living (CADL) Amsterdam-Nijmegen everyday language test (ANELT)	There's significantly improvement for tasks with temporal items with spatial items ($p = 0.042$) and wh-words ($p = 0.071$); there's no significant differences in written multiple-choice task ($p = 0.583$) Non-parametric tests had significantly larger direct effects for the linguistically oriented learning approach (L), but significantly larger post-effect for the stimulation approach (S)
Baker (2000)	Case study	$n = 2$	Patient 1: female 32-year-old, trauma brain injury in left hemisphere time since injury: 9 months Patient 2: male, 30-years old, trauma brain injury in left hemisphere time since injury: 4 months	Yes	No	Modified melodic intonation therapy (MMIT): Specific musical line and accompaniment for each trained sentence	30 min/session 3–8 sessions/week 4–27 months	Significant results	Functional language of 180 words/phrases Functional language of 45 words	Patient 1 had acquired a functional language of 148 words/phrases Patient 2 was able to independently generate the 30 words
Bonakdarpour et al. (2003)	Self-control	$n = 7$	7 patients with non-fluent aphasia Mean age: 52.4 years (45–61 years) time since stroke: 35.43 months (14–58 months)	No	No	Melodic intonation therapy in Persian (MIT-P): Exaggeration of normal prosody	3–4 days/week 1 month	Significant results	Wilcoxon signed-rank test Farsi aphasia test (FAT) Brain CT scan for diagnosis	Wilcoxon signed rank test showed statistically significant improvement in phrase length ($p = 0.0125$); number of correct content units ($p = 0.0107$); confrontational naming ($p = 0.0312$); responsive naming ($p = 0.0107$); repetition ($p = 0.0084$); word discrimination ($p = 0.238$); commands ($p = 0.238$)

(Continued)

TABLE 1 | Continued

Trial	Design	Sample	Participants	Music therapist involved	Blinding	Intervention	Duration and dose	Primary outcome	Measures	Main results
Researches measured by scales (26)										
Wilson et al. (2006)	Case study	$n = 1$	1 patient, male, 48-years-old with a left middle cerebral artery tertiary stroke for 4 years	Yes	No	3 methods of trained items Method 1: Melodic intonation therapy (MIT) Method 2: Repetition training Method 3: Unrehearsed training	2 days/week 1 month	Significant results	Magnetic resonance imaging (MRI) for diagnosis Boston diagnostic aphasic examination (BDAE) Australian music examinations board (AMEB)	Compared to Method 2 and Method 3, Method 1 showed a significant main immediate effect of time, $F_{(1,56)} = 6.47$, $p < 0.05$, and phrase group, $F_{(2,56)} = 13.9$, $p < 0.001$, and a significant interaction between time and group, $F_{(2,56)} = 9.95$, $p < 0.001$. The results showed a significant long time effect of phrase group, $F_{(1,37)} = 5.08$, $p < 0.05$, and a significant interaction effect between time and group, $F_{(1,37)} = 5.4$, $p < 0.05$
Racette (2006)	Self-control	$n = 8$	4 Severe Broca's aphasia 4 Moderate to severe mixed aphasia age: 36–67 years (mean 51.63 years) (mean 8 years) time since stroke: 5–19 years	No	No	Experiment 1: Aphasic patients repeated and recalled familiar songs Experiment 2: Aphasic patients repeated and recalled lyrics from novel songs Experiment 3: With an auditory model while learning novel songs, aphasics repeated and recalled more words when singing than when speaking	Once experimental time, three times		Neuro-psychological battery of tests Standard non-colored Raven's matrices Tower of London Montrea battery for evaluation of amusia (MBEA)	Singing perse does not help aphasics to improve their speech, whether the songs were familiar (Experiment 1) or unfamiliar (Experiment 2) But in Experiment 3, with an auditory model while learning novel songs, aphasics is better than speaking and singing in experiment 1 and 2

(Continued)

TABLE 1 | Continued

Trial	Design	Sample	Participants	Music therapist involved	Blinding	Intervention	Duration and dose	Primary outcome	Measures	Main results
Researches measured by scales (26)										
Kim and Tomaino (2008)	Self-control	$n = 7$	5 Severe Non-fluent aphasia 2 Moderate aphasia time since stroke: 14.28 years	Yes	No	Music therapist supported music therapy Familiar songs singing Breathing into single-syllable sounds Musically assisted speech Rhythmic speech cueing, Vocal intonation Dynamically cued singing Oral motor exercises	30 min/session 3 times/weeks 4 weeks		Neuro-psychological battery of tests Video measurement	Speech and singing carefully enhance each patient's expectancy in achieving improved performance of word retrieval, prosody and articulation

Hough (2010)	Case study	$n = 1$	1 patient, male, 69-years-old with chronic Broca's aphasia after left cerebro-vascular accident of 4 years' duration	No	No	Modified melodic intonation therapy (MMIT)	3h sessions/week
							8 weeks
							Follow-up at 2–4 weeks
								Significant results	Western aphasia battery-revised (WAB-R);
									Aphasia quotient (AQ);
									Cortical quotient (CQ);
									American speech-language hearing Association functional assessment of communication skills (ASHA FACS)
										After MIT, the patient reached 75% accuracy on automatic phrases at 4 weeks; self-generated phrases was 55% at 8 weeks
										The results revealed a significant difference in the automatic phrase data between baseline and post-treatment data ($t = 18.7314$; df = 6.456; $p < 0.00001$)
										The results revealed a significant difference in the self-generated phrase data between baseline and post-treatment data ($t = 33.3729$; df = 10; $p < 0.00001$)
										AQ increased 13 scores and CQ increased 13.6 after MIT

Vines et al. (2011)	Self-control	$n = 6$	Median age: 30–81 years time since stroke: at least 1 year	No	No	Melodic intonation therapy (MIT) transcranial direct current stimulation (tDCS)	EG: 20 min/day CG: 20 min/day 3 days	Significant results	Boston diagnostic aphasia Examination (BDAE)	Applying anodal-tDCS during MIT produced a significantly greater improvement in verbal fluency

(Continued)

TABLE 1 | Continued

Trial	Design	Sample	Participants	Music therapist involved	Blinding	Intervention	Duration and dose	Primary outcome	Measures	Main results
Researches measured by scales (26)										
Van der Meulen et al. (2012)	Clinical trial no. NTR 1961	$n = 2$	Patient 1 age: 29 years time since stroke: 8 months Patient 2 age: 25 years time since stroke: 2 weeks	No	No	EG: melodic intonation therapy (MIT) CG: No language therapy	3–5 h/week 6 weeks	Significant results	Aachen aphasia test (AAT) Amsterdam nijmegen everyday Language test (ANELT) Sabadell story retell task	Patient 1 improved 5 scores in repetition and 5 scores in comprehension (AAT) Patient 2 improved 35 scores in repetition trained phrases; 50' in repetition; 7' in action naming and 9' in comprehension (AAT). 7 scores was improved in ANELT; 22.5 scores was improved in Sabadell
Conklyn et al. (2012)	RCT pilot	$n = 30$	EG $n = 10$, age: 56.8 ± 17.11 years time since stroke: 32.2 ± 93.42 days CG $n = 14$, age: 66.9 ± 11.77 years time since stroke: 28.4 ± 67.84 days	Yes	Single	EG: modified melodic intonation therapy (MMIT) CG: Music therapist discussed with the patient	15 min/session 3 sessions	Significant results	Western aphasia battery (WAB)	Compared to the control, MIT group adjusted total items 1–3 score ($p = 0.02$); 2–3 score ($p = 0.02$) and responsive items 2–3 score ($p = 0.02$)
Stahl et al. (2013)	Self-control Cross-over Cross-over	$n = 3$	Median age: 56.2 years time since stroke: 23.47 months	No	No	Formulaic singing therapy (MT) Rhythmic therapy (RT) Standard therapy (ST)	1 h/session 3sessions/week 6 weeks	Significant results	Aachen aphasia test (AAT) Sabadell story retell task	Compare to RT and ST, MT group improved significantly ($p = 0.001$) in repetition; MT group improved in spontaneous words and is stable after 3 months.
Lim et al. (2013)	CCT	$n = 21$	EG $n = 12$, age: 56.5 years time since stroke: 187 days CG $n = 9$, age: 62.7 years time since stroke: 2,473 days	No	No	EG: melodic intonation therapy (MIT) CG: speech language therapy (SLT)	2 × 60 min/week 4 weeks	Significant results	Western Aphasia Battery in Korean version (K-WAB) Aphasia quotient (AQ)	In chronic group, MIT improved AQ ($p = 0.126$); spontaneous speech ($p = 0.126$); comprehension ($p = 0.429$) and repetition ($p = 0.177$) In subacute group, MIT improved AQ ($p = 0.476$); comprehension ($p = 0.067$) and naming ($p = 0.352$)

(Continued)

TABLE 1 | Continued

Trial	Design	Sample	Participants	Music therapist involved	Blinding	Intervention	Duration and dose	Primary outcome	Measures	Main results
Researches measured by scales (26)										
Zumbansen et al. (2014a)	Self-control Cross-over	$n = 3$	Median age: 51.67 years time since stroke: 21.67 months	No	No	Melodic intonation therapy (MT) Rhythmic syllables therapy (RT) Spoken syllables therapy (ST) Patient 1: MT-RT-ST Patient 2: RT-ST-MT Patient 3: ST-MT-RT	1 h/session 3 sessions/week 6 weeks	Significant results in MIT	MT-86 aphasia battery Verbal fluency test Montreal battery of evaluation of musical abilities (MBEMA) Wechsler Adult intelligence scale–III (WAIS-III); Wechsler memory scale-III (WMS-III); Computer syllables tests (cordial analyseur)	Compare to RT and ST, patient 1 improved significantly ($Z = -2.101$, $p = 0.036$) in MIT; patient 2 improved in MIT ($Z = -2.017$, $p = 0.044$); patient 3 improved in MIT ($Z = -2.329$, $p = 0.024$)
Van der Meulen et al. (2014)	RCT Cross-over	$n = 27$	EG $n = 16$, age: 53.1 ± 12.0 years time since stroke: 9.3 ± 2.0 weeks CG $n = 11$, age: 52.0 ± 6.6 years time since stroke: 11.9 ± 5.9 weeks	No	No	EG: melodic intonation therapy (MIT) CG: followed by delayed MIT	5 h/week 6 weeks	Significant results	Aachen aphasia test (AAT) Amsterdam Nijmegen everyday language test (ANELT) Semantic association task (SAT) Sabadell story retell task MIT repetition	Compared to the control group, MIT improved Repetition (AAT) ($p = 0.05$); MIT-repetition ($p < 0.01$); trained items ($p < 0.01$); untrained items ($p = 0.25$) There is no significant difference in Sabadell ($p = 0.82$); ANELT ($p = 0.07$) and Naming (AAT) ($p = 0.10$)
Cortese et al. (2015)	Self-control	$n = 6$	Median age: 59.8 ± 9.3 years Range: 53–71 years time since stroke: 9 months	No	No	Melodic-rhythmic therapy in Italian	30–40 min/day 4 days/week 16 weeks	Significant results	Aachen aphasia test (AAT)	In Italian MRT, phonemic structure, speech automatism, prosody, communication, correct repetition, naming and comprehension improved ($p = 0.031$)

(Continued)

TABLE 1 | Continued

Trial	Design	Sample	Participants	Music therapist involved	Blinding	Intervention	Duration and dose	Primary outcome	Measures	Main results
Researches measured by scales (26)										
Raglio et al. (2015)	RCT	$n = 30$	EG $n = 10$, CG $n = 10$ chronic aphasia	No	No	EG: singing active music therapy CG: speech language therapy	30 min/day 15 weeks	Significant results	Aachen aphasia test (AAT) Short form health survey	The study shows a significant improvement in spontaneous speech in the experimental group (Aachener Aphasie subtest: $p = 0.020$; Cohen's $d = 0.35$); the 50% of the experimental group showed also an improvement in vitality scores of short form health survey (chi squared 4.114; $p = 0.043$)
Van Der Meulen et al. (2016)	RCT Cross-over	$n = 17$	EG $n = 10$, age: 58.1 ± 15.2 years time since stroke: 33.1 ± 19.4 months	No	No	EG: melodic intonation therapy (MIT) CG: no language therapy	EG: 5 h/week 1–6 weeks	Significant results	Aachen aphasia test (AAT) Amsterdam-Nijmegen everyday	1–6 weeks: compare to CG, EG (MIT) group improved in trained items ($p = 0.02$); untrained items ($p = 0.40$)
			CG $n = 10$, age: 63.6 ± 12.7 years time since stroke: 42.6 ± 23.7 months				CG: 5 h/week 7–12 weeks		Language test (ANELT) Correct information units (CIU) Semantic association test (SAT)	7–12 weeks: compare to CG, EG (MIT) group improved in trained items ($p < 0.01$); untrained items ($p < 0.01$)
Slavin and Fabus (2018)	Case study	$n = 1$	Age: 63 years old Time since stroke: 10 years	Yes	No	Melodic intonation therapy (MIT)	50 min/session 2 sessions/week 12 weeks	Significant results	Boston diagnostic aphasi Examination (BDAE), Apraxia battery for adults II edition	MIT improved auditory comprehension skills, answering questions, and repetition of BDAE after listening to paragraphs
Martínez et al. (2018), Trial no. NCT3433495	Randomized cross-over pilot trial	n=20	EG $n = 10$, age: 66.05 ± 14.9 years time since stroke: 18.9 ± 13.43 months EG $n = 10$, age: 61.4 ± 13.7 years time since stroke: 24.1 ± 16.35 months	No	No	EG: Spanish adaptation of melodic intonation therapy (S-MIT) CG: delayed MIT	30 min/session 12 sessions 6 weeks	Significant results in CAL	Boston diagnostic aphasia examination (BDAE) Communicative activity log (CAL)	Compared to the control group, S-MIT improved communicative activity log (CAL) ($p = 0.048$) There is no significant difference in comprehension ($p = 0.925$) and repetition ($p = 0.727$) of BDAE between two groups

(Continued)

TABLE 1 | Continued

Trial	Design	Sample	Participants	Music therapist involved	Blinding	Intervention	Duration and dose	Primary outcome	Measures	Main results
Researches measured by scales (26)										
Kasdan and Kiran (2018)	Self-control	$n = 40$	EG $n = 16$, age: 61.25 ± 10.19 years CG $n = 16$, age: 63.0 ± 10.26 years time since stroke: not specified	No	No	Songs Melodic intonation therapy (MIT)	1 h/session	Significant results	Western aphasia battery (WAB Aphasia quotient (AQ)	MIT improved a three-factor repeated measures, phrase length ($p < 0.001$); a between-subjects effect of group ($p < 0.001$)
Leo et al. (2019)	RCT	n=31	EG $n = 17$, age: 54.4 ± 11.3 years time since stroke: 3 weeks–6 months CG $n = 14$, age: 51.4 ± 17.7 years time since stroke: 3 weeks–6 months	No	No	Music perception task Prosody perception task Sung-spoken story recall task	1–1.5 h/day 3 weeks	Significant results	NIHSS score BDAE aphasia severity rating scale MBEA scale and rhythm RBMT story recall immediate	In the two tasks, the aphasic patients recalling longer in the sung than spoken task ($p = 0.013$); emotional prosody perception correlated significantly with the recall in the sung task ($p < 0.001$); and also with chunk length in the
Zhang et al. (2021)	RCT	$n = 40$	EG $n = 20$; age: 52.90 ± 9.08 years CG $n = 20$; age: 54.05 ± 10.81 years	Yes	No	EG: melodic intonation therapy in Chinese CG: Speech therapy in Chinese	0.5 h/day 5 times/week 8 weeks	Significant results	Boston diagnostic aphasia examination (BDAE) Hamilton anxiety scale (HAMA) Hamilton depression scale (HAMD)	In the spontaneous speech (information, $p = 0.0002$), the listening comprehension (true or false, $p = 0.0019$; word recognition,, $p = 0.0001$; and sequential order, $p = 0.0001$), fluency ($p = 0.0019$), repetitions ($p = 0.0019$), and naming ($p = 0.0001$) of the intervention group were significantly higher than the control group in terms of the cumulative effect of time and the difference between groups after 8 weeks

(Continued)

TABLE 1 | Continued

Trial	Design	Sample	Participants	Music therapist involved	Blinding	Intervention	Duration and dose	Primary outcome	Measures	Main results
Researches measured by imaging (13)										
Naeser and Helm-Estabrooks (1985)	Original	$n = 8$	Good response group (GR) $n = 4$ age: 49.5 ± 10.69 years Poor response group (PR) $n = 4$ age: 41.75 ± 14.91 years	No	No	Melodic intonation therapy (MIT)	Over 1–8weeks	Significant results	Boston diagnostic aphasia examination (BDAE) **CT scan**	GR cases improved in speech characteristics ratings for phrase length and grammatical form on the BDAE; the PR cases showed no improvement GR cases had lesions which involved Broca's area and white matter deep to it plus large superior lesion extension into peri ventricular white matter deep to the lower motor cortex area for face, and had no large lesion in Wernicke's area and no lesion in the temporal isthmus or the right hemisphere PR cases had bi-lateral lesions or lesion including Wernicke's area or the temporal isthmus
Laine et al. (1994)	Self-control pilot study	$n = 3$	Patient 1: Chronic global aphasia male, 58-years-old, 5 months after stroke Patient 2: chronic mixed non-fluent aphasia; male; 58-years-old; 16 months after stroke Patient 3: chronic Wernicke's aphasia; 62-years-old; male; 4 months after stroke	No	No	Repetition of words and sentences either with normal prosody or intoned (intoned vs. normal speech)	45 min/session 3 sessions/week 3.5 months	Significant results	Boston diagnostic aphasia examination (BDAE); **CT Scan single photon emission computed tomography (SPECT)**	Patient 1 showed a totally uniform pattern in the relative perfusion changes. His pattern indicated increased left hemisphere, not right hemisphere activation during MIT Patient 2 and 3 did not find evidence for increased right hemisphere activation during MIT

(Continued)

TABLE 1 | Continued

Trial	Design	Sample	Participants	Music therapist involved	Blinding	Intervention	Duration and dose	Primary outcome	Measures	Main results
Researches measured by imaging (13)										
Belin et al. (1996)	Self-control pilot study	$n = 7$	2 Chronic Broca's aphasia; 5 chronic global aphasia age: 40–58 years (mean, 49.7 years) time since stroke: 4 to 41 months; (mean ± SD, 19 ± 15 months)	No	No	Repetition of sentences either with normal prosody or intoned (normal speech vs. silence; intoned vs. normal speech)	the duration of MIT in French therapy ranged from 37–42 months	Significant results	Boston diagnostic aphasia examination (BDAE); Wilcoxon signed-rank test **positron emission tomography (PET)**	Significantly more words ($p < 0.03$, Wilcoxon's rank sum test) were correctly repeated with MIT (16.3 ± 8 words) than without MIT (12.4 ± 8 words) Two findings: 1st, simple passive (word hearing) and active (word repetition) verbal tasks performed without MIT resulted in abnormal activation of right hemisphere structures, homotopic to those normally activated in the intact left hemisphere. 2nd, word repetition performed with MIT reactivated Broca's area and the adjacent left prefrontal cortex
Schlaug et al. (2008)	Case study randomly assigned	$n = 2$	2 patients with severe non-fluent aphasia as the result of a left hemisphere ischemic stroke Patient 1: male; age 47; time since stroke: 13 months Patient 2: male; age 58; time since stroke: 12 months	No	No	Melodic intonation therapy (MIT) Speech repetition therapy (SRT) Patient 1 MIT vs. Patient 2 SRT	90 min/day 5 days/week Over 8 weeks totally 70 sessions	Significant results	Correct information units (CIUs) Activities of daily living (ADL) Boston diagnostic aphasia examination (BDAE); **Functional magnetic resonance imaging (fMRI)**	MIT-treated patient had greater improvement on all outcomes than the SRT treated patient Patient 1 showed significant fMRI changes in a right-hemisphere network involving the premotor, inferior frontal, and temporal lobes Patient 2 had changes in a left hemisphere network consisting of the inferior pre- and post-central gyrus and the superior temporal gyrus

(Continued)

TABLE 1 | Continued

Trial	Design	Sample	Participants	Music therapist involved	Blinding	Intervention	Duration and dose	Primary outcome	Measures	Main results
Researches measured by imaging (13)										
Schlaug et al. (2009)	Self-control pilot study	$n = 6$	Median age: not specified chronic aphasia patients time since stroke: at least 1 year	No	No	Melodic intonation therapy (MIT)	75 sessions	Significant results	**Magnetic resonance imaging (MRI); Diffusion tensor imaging (DTI)**	All six patients showed a significant increase in the absolute number of fibers in the right AF comparing post- vs. pre-treatment DTI studies (paired *t*-test, $p = 0.04$) and also an increase in the fiber length
Breier et al. (2010)	Case study	$n = 2$	Patient 1 age: 55 years old time since stroke: 5 years Patient 2 age: 49 years old time since stroke: 2 years	No	No Single	Melodic intonation therapy (MIT)	30 min/session 2 session/day 2 days/week 3 weeks	Controversial results	Action naming test **Magneto-encephalo-graphy (MEG)**	Patient 1 exhibited a significant increase in CIUs (>35%) after the first block of treatment. This improvement was maintained after the break Patient 1, who was improved in language function to MIT, exhibited a steady reduction in activation within the right hemisphere across the two therapy blocks, resulting in a strong left hemisphere lateralization of MEG activity Patient 2, who did not respond positively to MIT, exhibited increased right hemisphere activation after both blocks of therapy compared to baseline, resulting in a right hemisphere lateralization of MEG activity

(Continued)

TABLE 1 | Continued

Trial	Design	Sample	Participants	Music therapist involved	Blinding	Intervention	Duration and dose	Primary outcome	Measures	Main results
Researches measured by imaging (13)										
van de Sandt-Koenderman et al. (2010)	Case study	*n* = 1	1 patient with Broca's aphasia in the subacute stage post-stroke Age: 25 years old; female time since stroke: 2 weeks	Not	No	Melodic intonation therapy (MIT)	1 h/session 5 sessions/week 2–8 weeks	Significant results	Aachen aphasia test (AAT): CIUs/minute Story retelling; **functional magnetic resonance imaging (fMRI)**	AAT: spontaneous speech 1/5–3/5; repetition *T* = 39–>*T* = 47; naming *T* = 39–>*T* = 46; CIUs/min 22.5–>55 fMRI: left more than right IFG, left superior and middle temporal gyrus and perilesional region in the angular/ supermariginal gyus, left caudate nucleus, bilateral supplementary, cingulate and premotor areas, left prefrontal cortex.
Zipse et al. (2012)	Case study	*n* = 1	1 patient with stroke resulted in very large left-hemisphere lesion age: 11-year-old	No	No	Melodic intonation therapy (MIT)	1.5 h/session 5 sessions/week 16 weeks 80 sessions 120 h totally	Significant results	**Functional magnetic resonance imaging (fMRI)** **Diffusion tensor imaging (DTI)**	fMRI: There was an increase in activation in right supplementary motor areas after 40 sessions and higher levels of activation in the right posterior middle temporal gyrus (MTG), occipital cortex, and possibly cerebellum. It showed a strong increase in activation of right posterior middle frontal and inferior frontal areas DTI: Both the arcuate fasciculus (AF) and uncinate fasciculus (UF) increased in volume at the beginning, midpoint and the conclusion
Al-Janabi et al. (2014)	Case study	*n* = 2	Patient 1 age: 65 years old time since stroke: 18 months Patient 2 age: 49 years old time since stroke: 20 months	No	No	Melodic intonation therapy (MIT) Excitatory repetitive transcranial magnetic stimulation (rTMS)	20 min/day 6 days	Significant results	Western aphasia battery (WAB) **Functional magnetic resonance imaging (fMRI); 3T MR system**	Patient 1 revealed significant activity increase in left BA44, $t = 1.79$, $p < 0.05$ and decrease in right BA44, $t = 2.92$, $p < 0.01$ Patient 2 revealed significant activity increase in left BA44, $t = 1.77$, $p < 0.05$, right BA44, $t = 1.77$, $p < 0.05$, left BA45, $t = 3.51$, $p < 0.001$

(Continued)

TABLE 1 | Continued

Trial	Design	Sample	Participants	Music therapist involved	Blinding	Intervention	Duration and dose	Primary outcome	Measures	Main results
Researches measured by imaging (13)										
Jungblut et al. (2014)	Case study	$n = 3$	Patient 1 age: 53 years old time since stroke: 18 months Patient 2 age: 44 years old time since stroke: 18 months Patient 3 age: 44 years old time since stroke: 18 months	No	No	Melodic intonation therapy (MIT)	Once experimental time	Significant results	Hierarchical word list (HWL) **Functional magnetic resonance imaging (fMRI)**	pre and posttreatment assessments of patients' vocal rhythm production, language, and speech motor performance yielded significant improvements for all patients In the left superior temporal gyrus, whereas the reverse subtraction revealed either significant activation or right hemisphere activation
Orellana et al. (2014)	Same group self-control	$n = 20$	Median age: 43 years Range: 21–51 years time since stroke: not specific	No	No	Melodically intoned stimuli Spoken language stimuli	2 conditions were completed in 1 experiment 30 min totally	Significant results	Functional Magnetic resonance **imaging (fMRI); 3T MR system**	Compared to spoken items, melodic > spoken. For melodic intoned items, increased activation was seen left-lateralized in the SMG, IPL, middle and superior temporal gyrus, middle and superior frontal gyrus, Right-lateralized activation was seen in the insula, rolandic operculum, and pars opercularis of the inferior frontal gyrus

(Continued)

TABLE 1 | Continued

Trial	Design	Sample	Participants	Music therapist involved	Blinding	Intervention	Duration and dose	Primary outcome	Measures	Main results
Researches measured by imaging (13)										
Akanuma et al. (2015)	Self-control	$n = 10$	Median age: 63.7 years Months from onset: 111.5	No	No	singing melodically therapy	30 min/session 1 session/week 10 weeks	Significant results	**Positron emission tomography (PET)**	5 patients exhibited improvements after singing intervention; all exhibited intact right basal ganglia and left temporal lobes
Tabei et al. (2016)	Case study	$n = 1$	1 Patient, age: 48 years old time since stroke: 3 years	Yes	No	Japanese version of melodic intonation therapy (MIT-J)	45 min/day 9 days	Significant results	Japanese version of western aphasia battery (J-WAB) Naming of 90 words aphasia quotient (AQ) Japanese version of Raven's Colored Progressive matrices (J-RCPM) Benton visual retention test (BVRT) **Functional MRI 3.0-T MR scanner**	After MIT, the patient improved 4 points in spontaneous speech; 0.9 in auditory comprehension; 0.8 in repetition; 1.3 in naming of J-WEB; 14 in AQ; 6 in correct naming words and 1.78 seconds in response time fMRI showed a significant activation of medial frontal gyrus, inferior frontal gyrus, superior temporal gyrus, lentiform nucleus, and lingual gyrus of the right hemisphere

The experimental researches on MIT from 1970 to present. It is divided into two parts. The first part is the experimental researches that used scales to measure the results. The second part is the experimental researches that used imaging for verification. EG, experimentalgroup; CG, ControlGroup; RCT, randomisedcontrolled trial; DTI, diffusion tensor imaging; MR, magnetoencephalography; SMG, supramarginal gyrus; IPL, inferior parietal lobule; MEG, magnetoencephalogram; tDCS, transcranial direct-current stimulation; BA, brodmann area.

conducted by Racette (2006), Kim and Tomaino (2008), Stahl et al. (2013), Zumbansen et al. (2014b), and Cortese et al. (2015), the implementer of the intervention process was realized by a speech therapist. Still, because the case study can record the detailed procedure, they compared the difference between melody and rhythm and found that the melody is dominant. The prognostic score will display more positive results. In the case report by Slavin and Fabus (2018), the therapist trained in NMT who performed MIT treatment also showed positive results. Although the samples in the above studies are generally small, the results are similar to the RCT study of more than 15 people, and the intervention process tends to be more musical.

The Advantages and Disadvantages of RCT Studies of Using Medical Imaging or Computers for Evaluation

In the clinical trials reviewed, most of the studies using MRI have the following characteristics: (i) case studies are dominant; (ii) the number of subjects is inferior or equal to 6; (iii) in case of large sample size, MRI observation should only be used before and once after MIT intervention to provide an immediate comparison. The above three characteristics are in an either-or relationship and will not appear simultaneously in the same study. In addition, we also found that the number of MIT musical interventions directly leads to different imaging results.

Among the RCT studies searched for, eight types of research used MRI to compare the effect before and after treatment. Orellana et al. (2014) compared an immediate impact on 20 aphasia patients before and after once MIT. After the intervention, fMRI and 3T MR scans showed that MIT increased activation in the left-lateralized in the SMG, IPL, STG, and SFG. Right-lateralized activation was seen in the insula, rolandic operculum, and pars opercularis of the inferior frontal gyrus. Akanuma et al. (2015) used positron emission tomography (PET) to conduct a self-control study in 10 chronic aphasia patients. The results demonstrated that 5 patients exhibited improvements after singing intervention; all indicated intact right basal ganglia and left temporal lobes. Norton et al. (2009), Schlaug et al. (2009), and Zipse et al. (2012) performed DTI to analyze structural changes in both hemispheres in 7 patients before and after MIT intervention. It turns out that all 7 patients showed a substantial increase in the absolute number of fibers in the right arcuate fasciculus (AF) comparing post-vs. pre-treatment DTI studies (paired t-test, $p = 0.04$) and also an increase in the fiber length, although omitting to mention the professional music therapists. It is worth noting that their melodic intervention time all exceeded 8 weeks, 75 courses of treatment. Al-Janabi et al. (2014) observed patients with functional magnetic resonance imaging after 6 days of MIT intervention and found that the left BA44 and right BA44 of the patients who received MIT had a significant increase in the activity. But Breier et al. (2010) compared two patients with chronic aphasia and came up with contradictory results. He showed a steady decrease in activation in the right hemisphere of both treatment areas, resulting in strong left hemisphere lateralization of MEG activity. However, Jungblut used his case studies through fMRI to argue that the limitation of this study is that activation changes were not measured by image acquisition before and after treatment (Jungblut et al., 2014). Cortese et al. (2015) found that in Italian MIT, all phonemic structure, speech automatism, prosody, communication, correct repetition, naming, and comprehension improved, while the adaptation of the MIT in the French language was developed by Belin et al. (1991).

Case Studies

Because the case study method is more meticulous and concentrated, the examination and evaluation method of MRI plus scale is more common.

Van der Meulen et al. (2012) compared MIT interventions with those of two patients. After 6 weeks, patients with MIT improved 35 scores in repetition trained phrases, 50 scores in repetition, 7 scores in action naming, and 9 scores in comprehension (AAT). Seven scores were improved in Amsterdam Nijmegen Everyday Language Test (ANELT); 22.5 scores were improved in Sabadel Story Retell Task. Slavin and Fabus (2018) conducted a before-after MIT intervention in a 63-year-old man with chronic aphasia for 10 years. Unlike other studies, Slavin teamed up with a professional music therapist to intervene. The results found that MIT improved auditory comprehension skills, question answering, and repetition of BDAE after listening to the paragraphs. Breier et al. (2010) compared two patients with chronic aphasia with an average age of 53 and an average duration of 3.5 years. Using MR to observe hemisphere structural changes, patient 1 with MIT exhibited a significant increase in CIUs (>35%) after the first block of treatment. Patient 1 showed lateralization in the right hemisphere of MEG activity. Al-Janabi et al. (2014) used transcranial magnetic stimulation (rTMS) and MIT to intervene two aphasia patients with an average duration of 15 months and using MR to the before-and-after comparison. The results revealed that patients with MIT revealed significant activity increase in left BA44 and a decrease in right BA44. Patient 2 revealed significant activity increase in left BA44, right BA44, and left BA45. Tabei et al. (2016) used fMRI to observe a 48-year-old patient with a 3-year history of chronic aphasia before and after 9 days of intensive MIT. The results showed in fMRI that the patient had a significant activation of the medial frontal gyrus, inferior frontal gyrus, superior temporal gyrus, lentiform nucleus, and lingual gyrus of the right hemisphere.

In the Research Using fMRI Measurement, the Main Concentrated Region of Interest in Brain

Through summarizing the studies in **Table 1** which used fMRI to support MIT, we used the software BrainNet Viewer to locate the brain ROI (regions of interest). BrainNet Viewer is a brain network visualization tool for imaging connect omics. It can help researchers to visualize topological patterns of structural and to find functional brain networks derived from different imaging modalities (Xia et al., 2013). Using the BrainNet Viewer to locate the occurrence sites, it was found that all MIT-supported patients had more activation ROI in the right hemisphere than in the left hemisphere. The concentrated areas of ROI are the precentral gyrus, precentral sulcus, postcentral gyrus, middle frontal gyrus, superior temporal gyrus, superior temporal sulcus, middle temporal gyrus, inferior temporal sulcus, lingual gyrus, angular gyrus, etc. (**Figure 2**).

DISCUSSIONS

Our review selected 39 typical effective MIT experimental studies from 127 studies. Their common feature is the use of musical melody to intervene in aphasia, accompanied by effective evaluation. This analysis and discussion are based on the analysis of the intervention methods, evaluations, and effects of these studies.

Feasibility Differences in Measurement Methods Between Clinical Trials With More Subjects and Case Studies

In these RCT studies with relatively more subjects, we found that objective imaging observation was not used as a primary means of effective monitoring. There may be some correlations to the therapeutic way of MIT. The one-to-one treatment and evaluation method will increase the working load of clinicians. If every participant is involved in the medical imaging test, the clinical workload, patient compliance, and financial support will all influence factors. Therefore, in more than 6 subjects of RCT studies in the past 10 years, only two articles with imaging observations were found. However, all of the RCT findings, including the two objective tests, confirmed the effectiveness of the subjective measurement scale of MIT. Because MIT requires individualized intervention and a long course of treatment, language assessment scales are the most convenient way of assessment. Compared with the high-cost evaluation of functional MRI, the scale evaluation of more than 15 patients with aphasia is easy to operate on and easy to compare before and after. In these MIT studies using MRI detection, the changes in cortical white matter and fiber bundles are apparent, which provide substantial evidence for the therapeutic effect of MIT and lay a foundation for the study of neural mechanisms.

However, due to the time-consuming, labor-intensive, and cost-intensive MRI examinations, most of these MIT's RCT studies have the following shortcomings: (i) there are some studies (Orellana et al., 2014) that could perform long-term MIT intervention experiments, and the imaging examinations are meticulous. Still, the number of samples is too small. Most of the samples comprised 6 participants; (ii) although there are 4 studies (Schlaug et al., 2009; Stahl et al., 2013; Zumbansen et al., 2014a; Cortese et al., 2015) that can match the minimum number of statistical subjects, there is no long-term intervention for comparison; therefore, the cumulative effect cannot be observed. The only one-time immediate effect is not enough to explain the mechanism. Therefore, in the future, how to ensure that both the demand for sample size and the long-term intervention of MRI detection can be achieved is matter of pressing academic concern.

The Number of Musical Factors in the MIT Intervention Is Directly Linked to the Imaging Results

Previous literature has demonstrated several effective clinical results related to the recovery of musical melody-induced speech in the treatment of post-stroke aphasia. MIT (Albert et al., 1973), formally proposed by the American Academy of Neurology in 1973, is used to treat aphasia. In the early clinical treatment of non-fluency aphasia, Sparks et al. (1974) recorded the use of spectrum examples when training patients, with "Sprechgesang" as the core, requiring patients to follow the written melody. Zipse et al. (2012), Orellana et al. (2014), and Tabei et al. (2016), and others tend to use MIT treatment under more musical intervention, so their imaging results all show more features of active right hemisphere area. Moreover, although Zipse et al. (2012), Schlaug et al. (2014), Akanuma et al. (2015), and others used MIT recordings or provided MIT by general therapists, their intervention processes were all over 8 weeks. The natural melody factor in MIT makes the imaging results they obtained also reflected the active characteristics of the right hemisphere. Therefore, in the existing MIT experimental research, it is found that musical factors and the cumulative effect of time will directly affect the evidence that the right hemisphere of the brain participates in activities. Although rhythm is part of the music, as the rhythm is unpitched, we did not find a clear trend of activating the right hemisphere in the MIT intervention under the guidance of rhythm or language.

It is reported that the effects of the musical rhythm are observed in the left brain areas (Chen et al., 2008) and listening to musical rhythms recruits motor regions of the brain (Limb Charles et al., 2006; Limb et al., 2006; Thaut et al., 2014). However, these studies only focus on the music listening of healthy individuals or the rhythm perception of musicians. They are not the observation of MIT on patients with aphasia caused by stroke in the left hemisphere. Therefore, in the case of damage to the language center of the left hemisphere, patients treated with MIT can have correct oral speech output. This phenomenon confirms the mechanism of musical pitch from one side. But its brain mechanism still needs further study.

The Neural Mechanism of MIT Based on Music

In the evidence summarized in previous MIT experimental studies, we found that the ROIs activated by MIT were the central anterior gyrus, central anterior sulcus, central posterior gyrus, middle frontal gyrus, superior temporal gyrus, superior temporal sulcus, middle temporal gyrus, inferior temporal gyrus, lingual gyrus, and angular gyrus of the right hemisphere. These areas include the frontal motor cortex (including Broca's area and ventral anterior motor cortex), which connects speech sensation and output, auditory cortex (including superior temporal gyrus and middle temporal gyrus), and parietal cortex (including angular gyrus and gyrus). MIT based on music activities, that is, MIT provided by professional music therapists, whether extracting lyrics from familiar songs or learning new fixed-pitch short melody for patients, affects the white matter structure of

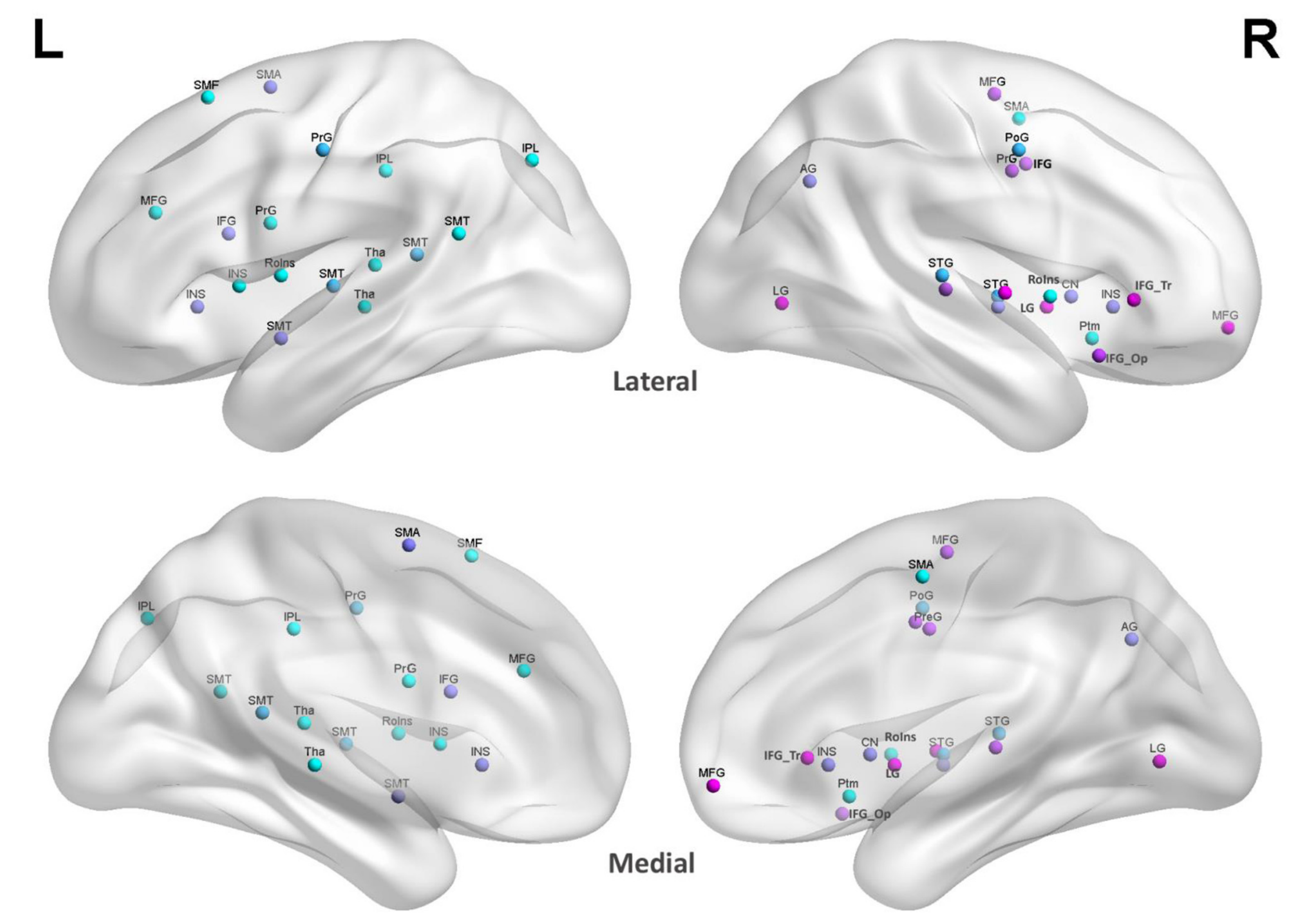


FIGURE 2 | ROI in MIT-supported studies. The site colors were randomly generated to distinguish different ROI regions at the same site to prevent color overlap. L, left hemisphere; R, right hemisphere; Lateral, lateral cerebral hemisphere; Medial, medial cerebral hemisphere; SMT, superior and middle temporal gyrus; SMA, supplementary motor area; SMF, superior and middle frontal gyrus; PrG, pre-central gyrus; PoG, post-central gyrus; IPL, inferior parietal lobule; IFG, inferior frontal gyrus; INS, insula; MFG, middle frontal gyrus; RoIns, rolandic operculum insula; Tha, thalamus; LG, lingual gyrus; AG, angular gyrus; LN, lentiform nucleus; IFG_Tr, inferior frontal gyrus pars triangularis; IFG_Op, inferior frontal gyrus pars opercularis.

the auditory-motor neural circuit compensation to promote the ability to encode and integrate verbal information. This trans-hemisphere "mirror effect" has an important mechanism for the language recovery of patients with aphasia.

Valuable Findings in Case Studies

It is found in literature retrieval that the evaluation methods of case studies are generally comprehensive and meticulous. Such qualitative studies reflecting the therapeutic effects of satisfactory MIT have more profound clinical implications for the brain regions it may activate. In the case reports retrieved in this paper, the evaluation criteria of early studies were generally international scales, mostly subjective scoring methods, and language recovery competence was based on scoring in different dimensions. In the recent 10 years of research, some medical imaging evidence of changes in brain structure at MIT to aphasia patients is easier to find in case reports (Schlaug et al., 2009; Al-Janabi et al., 2014; Tabei et al., 2016; Martzoukou et al., 2021). Besides, in the case study, whether the language assessment scale or fMRI was used, the subjective measurement and objective monitoring of patients have received sufficient concertation. Evidence of structural changes in patients' brain regions before and after also provides a factual neurological basis for MIT. It provides a realistic basis for the treatment of clinical aphasia.

The Importance of Music Therapist at MIT

In the literature we reviewed, only 5 studies mentioned the participation of music therapists. Although MIT originated in speech therapy, MIT's guidance is melodic. It should be necessary for a correct rehabilitative approach by MIT to have specific training. For instance, the accuracy of melodic language needs a musical or music therapy formation. The rest of the literature does not mention the credentials of speech therapists and whether they have music learning experience. In fact, in MIT, treatment performed by music therapists includes instrumental accompaniment, melodic guidance, and songs inducement.

Therefore, in the process of activating the vocabulary encoding of patients with aphasia, the instrument accompaniment, the professional, accurate melodic pitch, and the guidance from music therapists to play and sing are all combined to activate the melodic "lyrics" of the episodic memory network and promote the output of spoken language.

Expectations for Future MIT Development

Through MIT's RCT studies, the left and right brains were found to have different processing advantages. The functional areas responsible for music melody processing and memory retrieval are more concentrated in the auditory cortex of the right brain temporal lobe. Therefore, it is speculated that the left brain is more responsible for language functions. After damage sustained by the dominant hemisphere, MIT may activate the auditory cortex corresponding to music processing in the right brain and activate the right brain language motor area corresponding to the Broca's area of the left brain through the conduction of the right arcuate track to achieve compensation and guide the patient's language output, to achieve the purpose of language communication (Merrett et al., 2014). However, 90% of the literature we reviewed so far was RCT studies on Western language aphasia; only 10% of the literature comes from East Asian language aphasia (Japanese and Korean), while the MIT intervention in Chinese Mandarin aphasia trials has not been found in internationally registered clinical trials. Compared with Western languages, East Asian languages, as a tonal language (including four or more tones), have a more bilateral distribution of brain nerve circuits than Western languages represented by English (Liang and Du, 2018). However, despite this, the neural mechanism of the effect of MIT on East Asian languages has not been verified by a large sample of experiments.

It should be noted that, according to the high incidence of aphasia. However, relatively effective treatment methods were developed. A large amount of imaging evidence has not supported MIT, nor has it been endorsed by large cohort studies. This may be due to factors such as MIT's over-reliance on therapists, its unitary approach, lack of computerization, and individual patient differences. From existing evidence, MIT is effective and has positive results of scale testing. In the future, researchers should try the use of technology to develop music artificial intelligence evaluation and training tools, streamline and step the operation of MIT, reduce the human cost, and, on this basis, cooperate with imaging detection, and then conduct large sample experiments, so that the clinical and scientific value of MIT will be maximized in the future.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This research received a grant from Scientific Research Project of Establishment of the Winter Olympics Sports Injury Rehabilitation Diagnosis and Treatment System and Green Channel Demonstration, No. 2018YFF0301104 (to JL).

ACKNOWLEDGMENTS

We thank YD for the guidance, advice, encouragement, and assistance for this review, and Teng Wen-Jia for the analysis in BrainNet Viewer.

REFERENCES

Akanuma, K., Meguro, K., Satoh, M., Tashiro, M., and Itoh, M. (2015). Singing can improve speech function in aphasics associated with intact right basal ganglia and preserve right temporal glucose metabolism: Implications for singing therapy indication. *Int. J. Neurosci.* 126, 39–45. doi: 10.3109/00207454.2014.992068

Albert, M. L., Sparks, R. W., and Helm, N. A. (1973). Melodic intonation therapy for aphasia. *Arch. Neurol.* 29, 130–131. doi: 10.1001/archneur.1973.00490260074018

Al-Janabi, S., Nickels, L. A., Sowman, P. F., Burianová, H., Merrett, D., and Thompson, B. (2014). Augmenting melodic intonation therapy with non-invasive brain stimulation to treat impaired left-hemisphere function two case studies. *Front. Psychol.* 5:37. doi: 10.3389/fpsyg.2014.00037

Baker, F. (2000). *Modifying the Melodic Intonation Therapy Program for Adults With Severe Non-fluent Aphasia. Music Therapy Perspectives (2000), Vol. 18.* New York, NY: American Music Therapy Association.

Belin, P., Van Eeckhout, P., Zilbovicius, M., Remy, P., François, C., Guillaume, S., et al. (1996). Recovery from nonfluent aphasia after melodic intonation therapy: a PET study. *Neurology* 47, 1504-1511. doi: 10.1212/wnl.47.6.1504

Belin, P., Van Eeckhout, P., Zilbovicius, M., Remy, P., FranFois, C., Guillaume, S., et al. (1991). Recovery from nonfluent aphasia after melodic intonation therapy. *Hum. Mov. Sci.* 10, 315–334.

Benjamin, E. J., Blaha, M. J., Chiuve, S. E., Cushman, M., Das, S. R., Deo, R., et al. (2017). Heart disease and stroke statistics-—2017 update: a report from the American heart association. *Circulation* 135, e146–e603. doi: 10.1161/CIR.0000000000000485

Bonakdarpour, B., Eftekharzadeh, A., and Ashayeri, H. (2003). Melodic intonation therapy in Persian aphasic patients. *Aphasiology.* 17, 75–95. doi: 10.1080/729254891

Boucher, V., Garcia, L. J., Fleurant, J., and Paradis, J. (2001). Variable efficacy of rhythm and tone in melody-based interventions: implications for the assumption of a right-hemisphere facilitation in nonfluent aphasia. *Aphasiology* 15, 131–149. doi: 10.1080/02687040042000098

Breier, J. I., Randle, S., Maher, L. M., and Papanicolaou, A. C. (2010). Changes in maps of language activity activation following melodic intonation therapy using magnetoencephalography: two case studies. *J. Clin. Exp. Neuropsychol.* 32, 309–314. doi: 10.1080/13803390903029293

Callan, D. E., Tsytsarev, V., Hanakawa, T., Callan, A. M., Katsuhara, M., Fukuyama, H., et al. (2006). Song and speech: brain regions involved with perception and covert production. *Neuroimage* 31, 1327–1342. doi: 10.1016/j.neuroimage.2006.01.036

Chen, J. L., Penhune, V. B., and Zatorre, R. J. (2008). Moving on time: brain network for auditory-motor synchronization is modulated by rhythm complexity and musical training. *J. Cogn. Neurosci.* 20, 226–239. doi: 10.1162/jocn.2008.20018

Cohen, N. S., and Masse, R. (1993). The application of singing and rhythmic instruction as a therapeutic intervention for persons with neurogenic communication disorders. *J. Music Ther*. 30, 81–99. doi: 10.1093/jmt/30.2.81

Conklyn, D., Novak, E., Boissy, A., Bethoux, F., and Chemali, K. (2012).The effects of modified melodic intonation therapy on nonfluent aphasia: a pilot study. *J. Speech Lang. Hear. Res.* 55:1463. doi: 10.1044/1092-4388(2012/11-0105)

Cortese, M., Riganello, F., and Arcuri, F. (2015). Rehabilitation of aphasia: application of melodic-rhythmic therapy to the Italian language. *Front. Hum. Neurosci.* 9:520. doi: 10.3389/fnhum.2015.00520

Gerstenecker, A., and Lazar, R. (2019). Language recovery following stroke. *Clin. Neuropsychol*. 33, 928–947. doi: 10.1080/13854046.2018.1562093

Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Blaha, M. J., et al. (2014). Executive summary: Heart disease and stroke statistics – 2014 update: a report from the American heart association. *Circulation* 129, 399–410. doi: 10.1161/01.cir.0000442015.53336.12

Goldfarb, R., and Bader, E. (1979). Espousing melodic intonation therapy in aphasia rehabilitation: a case study. *Int. J. Rehabil. Res.* 2, 333–42. doi: 10.1097/00004356-197909000-00002

Goodglass, H., and Kaplan, E. (1972). *The Assessment of Aphasia and Related Disorders*. London: Henry Klimpton.

Haro-Martínez, A. M., Lubrini, G., Madero-Jarabo, R., Díez-Tejedor, E., and Fuentes, B. (2019). Melodic intonation therapy in post-stroke nonfluent aphasia: a randomized pilot trial. *Clin. Rehabil*. 33, 44–53. doi: 10.1177/0269215518791004

Helm-Estabrooks, N., and Albert, M. (2004). *Manual of Aphasia and Aphasia Therapy*, 2nd Edn. Austin, TX: PRO-ED, Inc.

Helm-Estabrooks, N., Nicholas, M., and Morgan, A. (1989). *Melodic Intonation Therapy.* Austin, TX: PRO-ED, Inc.

Higgins, J. P. T., and Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*. The Cochrane Collaboration. Available online at: www. handbook.Cochrane.org.

Hoover, J. (2019). Heatstroke. *N. Engl. J. Med.* 381, 2449–2459. doi: 10.1056/NEJMc1909690

Hough, M. S. (2010). Melodic intonation therapy and aphasia: another variation on a theme. *Aphasiology* 24, 775–786. doi: 10.1080/02687030903501941

Jeffries, K. J., Fritz, J. B., and Braun, A. R. (2003). Words in melody: an H(2)15O PET study of brain activation during the singing and speaking. *Neuroreport* 14, 749–754. doi: 10.1097/00001756-200304150-00018

Jungblut, M., Huber, W., Mais, C., and Schnitker, R. (2014). Paving the way for speech: voice-training-induced plasticity in chronic aphasia and apraxia of speech—three single cases. *Neural Plast.* 2014, 1–14. doi: 10.1155/2014/841982

Kamath, V., Sutherland, E. R., and Chaney, G.-A. (2019). A meta-analysis of neuropsychological functioning in the logopenic variant of primary progressive aphasia: comparison with the semantic and non-fluent variants. *J. Int. Neuropsychol. Soc.* 26, 322–330. doi: 10.1017/S1355617719001115

Kasdan, A., and Kiran, S. (2018). Please don't stop the music: song completion in patients with aphasia. *J. Commun. Disord.* 75, 72–86. doi: 10.1016/j.jcomdis.2018.06.005

Kim, G., Min, D., Lee, E. O., and Kang, E. K. (2016). Impact of co-occurring dysarthria and aphasia on functional recovery in post-stroke patients. *Ann. Rehabil. Med*. 40, 1010–1017. doi: 10.5535/arm.2016.40.6.1010

Kim, M., and Tomaino, C. M. (2008). Protocol evaluation for effective music therapy for persons with nonfluent aphasia. *Top. Stroke Rehabil.* 15, 555–569. doi: 10.1310/tsr1506-555

Koleck, M., Gana, K., Lucot, C., Darrigrand, B., Mazaux, J. M., and Glize, B. (2017). Quality of life in aphasic patients 1 year after a first stroke. *Qual. Life Res.* 26, 45–54. doi: 10.1007/s11136-016-1361-z

Laine, M., Tuomainen, J., and Ahonen, A. (1994). *Changes in hemispheric brain perfusion elicited by Melodic Intonation Therapy: A preliminary experiment with single photon emission computed tomography (SPECT)*. Cham: Scandinavian University Press.

Leo, V., Sihvonen, A. J., Linnavalli, T., Tervaniemi, M., Laine, M., Soinila, S., et al. (2019). Cognitive and neural mechanisms underlying the mnemonic effect of songs after stroke. *NeuroImage Clin.* 24:101948. doi: 10.1016/j.nicl.2019.101948

Liang, B., and Du, Y. (2018). The functional neuroanatomy of lexical tone perception: an activation likelihood estimation meta-analysis. *Front. Neurosci.* 12:495. doi: 10.3389/fnins.2018.00495

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 339:b2700. doi: 10.1136/bmj.b2700

Lim, K. B., Kim, Y. K., Lee, H. J., Yoo, J., Hwang, J. Y., Kim, J. A., et al. (2013). The therapeutic effect of neurologic music therapy and speech-language therapy in post-stroke aphasic patients. *Ann. Rehabil. Med.* 37:556. doi: 10.5535/arm.2013.37.4.556

Limb Charles, J., Stefan, K., Eric, O. B., Sherin, R., and Allen, B. R. (2006). Left hemispheric lateralization of brain activity during passive rhythm perception in musicians. *Anat. Rec. A Discov. Mol. Cell. Evol. Biol.* 288, 382–389. doi: 10.1002/ar.a.20298

Limb, C. J., Kemeny, S., Ortigoza, E. B., Rouhani, S., and Braun, A. R. (2006). Left hemispheric lateralization of brain activity during passive rhythm perception in musicians. *Cereb. Cortex* 18, 2844–2854.

Marshall, N., and Holtzapple, P. (1976). *Medodic Intonation Therapy: variations on a theme*. Audiology and Speech pathology service veterans administration hospital, Bew Orleans, Louisianna.

Martínez, A. H., Lubrini, G., Madero-Jarabo, R., Díez-Tejedor, E., and Fuentes, B. (2018). Melodic intonation therapy in post-stroke nonfluent aphasia: a randomized pilot trial. *Clin. Rehabili.* 33, 44–53.

Martzoukou, M., Nousia, A., Nasios, G., and Tsiouris, S. (2021). Adaptation of melodic intonation therapy to greek: a clinical study in broca's aphasia with brain perfusion spect validation. *Front. Aging Neurosci*. 13:664581. doi: 10.3389/fnagi.2021.664581

Merrett, D. L., Peretz, I., and Wilson, S. J. (2014). Neurobiological, cognitive, and emotional mechanisms in melodic intonation therapy. *Front. Hum. Neurosci.* 8:401. doi: 10.3389/fnhum.2014.00401

Naeser, M. A., and Helm-Estabrooks, N. (1985). Ct scan lesion localization and response to melodic intonation therapy with nonfluent aphasia cases. *Cortex* 21, 203–223.

Norton, A., Zipse, L., Marchina, S., and Schlaug, G. (2009). Melodic intonation therapy. *Ann. N. Y. Acad. Sci.* 1169, 431–436. doi: 10.1111/j.1749-6632.2009.04859.x

Orellana, M. C. P., van de Sandt-Koenderman, M. E., Saliasi, E., van der Meulen, I., Klip, S., van der Lugt, A., et al. (2014). Insight into the neurophysiological processes of melodically intoned language with functional MRI. *Brain Behav*. 4, 615–625. doi: 10.1002/brb3.245

Ozdemir, E., Norton, A., and Schlaug, G. (2006). Shared and distinct neural correlates of singing and speaking. *Neuroimage* 33, 628–635. doi: 10.1016/j.neuroimage.2006.07.013

Popovic, A., and Boniver, J. (1992). Cause of death determined at autopsy in the University Hospital of Liège. *Developments from 1878 to 1986. Rev. Med. Liege.* 47, 618–23.

Racette, A. (2006). Making non-fluent aphasics speak: sing along! *Brain* 129, 2571–2584. doi: 10.1093/brain/awl250

Raglio, A., Oasi, O., Gianotti, M., Rossi, A., Goulene, K., and Stramba-Badiale, M. (2015). Improvement of spontaneous language in stroke patients with chronic aphasia treated with music therapy: a randomized controlled trial. *Int. J. Neurosci.* 126, 235–242. doi: 10.3109/00207454.2015.1010647

Schlaug, G., Marchina, S., and Kumar, S. (2014). Study design for the fostering eating after stroke with transcranial direct current stimulation trial: a randomized controlled intervention for improving Dysphagia after acute ischemic stroke. *J. Stroke Cerebrovasc. Dis.* 24, 511–520. doi: 10.1016/j.jstrokecerebrovasdis.2014.09.027

Schlaug, G., Marchina, S., and Norton, A. (2008). From singing to speaking: why singing may lead to recovery of expressive language function in patients with broca's aphasia. *Music Percep*. 25, 315–323.

Schlaug, G., Marchina, S., and Norton, A. (2009). Evidence for plasticity in white-matter tracts of patients with chronic Broca's aphasia undergoing intense intonation-based speech therapy. *Ann. N. Y. Acad. Sci.* 1169, 385–394. doi: 10.1111/j.1749-6632.2009.04587.x

Schlaug, G., Norton, A., and Marchina, S. (2010). From singing to speaking: facilitating recovery from non-fluent aphasia. *Fut. Neurol.* 5, 657–665. doi: 10.2217/fnl.10.44

Slavin, D., and Fabus, R. (2018). A case study using a multimodal approach to melodic intonation therapy. *Am. J. Speech Lang. Pathol.* 27, 1352–1362. doi: 10.1044/2018_AJSLP-17-0030

Sparks, R. W., Helm, N. A., and Albert, M. L. (1974). Aphasia rehabilitation resulting from melodic intonation therapy. *Cortex* 10, 303–316. doi: 10.1016/S0010-9452(74)80024-9

Sparks, R. W., and Holland, A. L. (1976). Method: melodic intonation therapy for aphasia. *J. Speech Hear. Disord*. 41, 287–297. doi: 10.1044/jshd.4103.287

Springer, L., Willmes, K., and Haag, E. (1993). Training in the use of wh-questions and prepositions in dialogues: A comparison of two different approaches in aphasia therapy. *Aphasiology* 7, 251–270, doi: 10.1080/02687039308249509

Stahl, B., Henseler, I., Turner, R., Geyer, S., and Kotz, S. A. (2013). How to engage the right brain hemisphere in aphasics without even singing: evidence for two paths of speech recovery. *Front. Hum. Neurosci.* 7:35. doi: 10.3389/fnhum.2013.00035

Tabei, K. I., Satoh, M., Nakano, C., Ito, A., Shimoji, Y., Kida, H., et al. (2016). Improved neural processing efficiency in a chronic aphasia patient following melodic intonation therapy: a neuropsychological and functional MRI study. *Front. Neurol.* 7:148. doi: 10.3389/fneur.2016.00148

Thaut, M. H., Trimarchi, P. D., and Parsons, L. M. (2014). Human brain basis of musical rhythm perception: common and distinct neural substrates for meter, tempo, and pattern. *Brain Sci.* 4, 428–452. doi: 10.3390/brainsci4020428

van de Sandt-Koenderman, M., Smits, M., van der Meulen, I., Visch-Brink, E., van der Lugt, A., and Ribbers, G. (2010). A case study of melodic intonation therapy (MIT) in the subacute stage of aphasia: early re-re activation of left hemisphere structures. *Proc. Soc. Behav. Sci.* 6, 241–243. doi: 10.1016/j.sbspro.2010.08.121

Van Der Meulen, I., De Sandt-Koenderman, V., Mieke, W. M. E., Heijenbrok, M. H., Visch-Brink, E., and Ribbers, G. M. (2016). Melodic intonation therapy in chronic aphasia: evidence from a pilot randomized controlled trial. *Front. Hum. Neurosci.* 10:533. doi: 10.3389/fnhum.2016.00533

Van der Meulen, I., Sandt-Koenderman, V., and Ribbers, G. M. (2012). Melodic intonation therapy: present controversies and future opportunities. *Arch. Phys. Med. Rehabil.* 93, S46–S52. doi: 10.1016/j.apmr.2011.05.029

Van der Meulen, I., van de Sandt-Koenderman, W. M. E., Heijenbrok-Kal, M. H., Visch-Brink, E. G., and Ribbers, G. M. (2014). The efficacy and timing of melodic intonation therapy in subacute aphasia. *Neurorehabil. Neural Repair* 28, 536–544. doi: 10.1177/1545968313517753

Vines, B. W., Norton, A. C., and Schlaug, G. (2011). Non-invasive brain stimulation enhances the effects of melodic intonation therapy. *Front. Psychol.* 2:230. doi: 10.3389/fpsyg.2011.00230

Wang, L. D., Liu, J. M., Yang, Y., Peng, B., and Wang, Y. L. (2019). Stroke prevention in china still faces huge challenges-summary of "stroke prevention report 2018 in China. *China Circul. J.* 34, 105–119.

WHO (2015). *World Report on Aging and Health.* Available online at: http://apps.who.int/iris/bitstream/10665/186463/1/9789240694811_eng.pdf?ua=1 (accessed April 11, 2017).

Wilson, SJ., Parsons, K., and Reutens, D. C. (2006). *Preserved Singing In Aphasia: A Case Study Of The Efficacy Of Melodic Intonation Therapy*. Music Perception: An Interdisciplinary Journal, Vol. 24, No. 1. p. 23–36.

Xia, M., Wang, J., and He, Y. (2013). BrainNet viewer: a network visualization tool for human brain connectomics. *PLoS ONE* 8:e68910. doi: 10.1371/journal.pone.0068910

Zhang, X., Wang, C., and Liu, S. (2016). Case study of melodic intonation therapy and therapeutic singing in the treatment of motor aphasia after stroke. *Chin. J. Stroke* 11, 791–794. doi: 10.3969/j.issn.1673-5765.2016.09.016

Zhang, X. Y., Yu, W. Y., Teng, W. J., Lu, M. Y., Wu, X. L., Yang, Y. Q., et al. (2021). Effectiveness of melodic intonation therapy in chinese mandarin on non-fluent aphasia in patients after stroke: a randomized control trial. *Front. Neurosci.* 15:648724. doi: 10.3389/fnins.2021.648724

Zipse, L., Norton, A., Marchina, S., and Schlaug, G. (2009). Singing versus speaking in nonfluent aphasia. *Neuroimage* 47:S119. doi: 10.1016/S1053-8119(09)71121-8

Zipse, L., Norton, A., Marchina, S., and Schlaug, G. (2012). When right is all that is left: plasticity of righthemisphere tracts in a young aphasic patient. *Ann. N. Y. Acad. Sci.* 1252, 237–245. doi: 10.1111/j.1749-6632.2012.06454.x

Zumbansen, A., Peretz, I., and Harbert, S. (2014a). The combination of rhythm and pitch can account for the beneficial effect of melodic intonation therapy on connected speech improvements in Brocaâ€™s aphasia. *Front. Hum. Neurosci.* 8:592. doi: 10.3389/fnhum.2014.00592

Zumbansen, A., Peretz, I., and Hébert, S. (2014b). Melodic intonation therapy: back to basics for future research. *Front. Neurol.* 5:7. doi: 10.3389/fneur.2014.00007

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

ORIGINAL RESEARCH
published: 28 February 2022
doi: 10.3389/fpsyg.2022.801263

Music Perception Abilities and Ambiguous Word Learning: Is There Cross-Domain Transfer in Nonmusicians?

Eline A. Smit[1,2]*, *Andrew J. Milne*[1] *and Paola Escudero*[1,2]

[1]*The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, NSW, Australia,*
[2]*ARC Centre of Excellence for the Dynamics of Language, Canberra, ACT, Australia*

OPEN ACCESS

Edited by:
Caicai Zhang,
The Hong Kong Polytechnic University, Hong Kong SAR, China

Reviewed by:
Francesca Talamini,
University of Padua, Italy
Mireille Besson,
UMR7291 Laboratoire de Neurosciences Cognitives (LNC), France

***Correspondence:**
Eline A. Smit
e.smit@westernsydney.edu.au

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

Received: *25 October 2021*
Accepted: *08 February 2022*
Published: *28 February 2022*

Citation:
Smit EA, Milne AJ and Escudero P (2022) Music Perception Abilities and Ambiguous Word Learning: Is There Cross-Domain Transfer in Nonmusicians? Front. Psychol. 13:801263. doi: 10.3389/fpsyg.2022.801263

Perception of music and speech is based on similar auditory skills, and it is often suggested that those with enhanced music perception skills may perceive and learn novel words more easily. The current study tested whether music perception abilities are associated with novel word learning in an ambiguous learning scenario. Using a cross-situational word learning (CSWL) task, nonmusician adults were exposed to word-object pairings between eight novel words and visual referents. Novel words were either non-minimal pairs differing in all sounds or minimal pairs differing in their initial consonant or vowel. In order to be successful in this task, learners need to be able to correctly encode the phonological details of the novel words and have sufficient auditory working memory to remember the correct word-object pairings. Using the Mistuning Perception Test (MPT) and the Melodic Discrimination Test (MDT), we measured learners' pitch perception and auditory working memory. We predicted that those with higher MPT and MDT values would perform better in the CSWL task and in particular for novel words with high phonological overlap (i.e., minimal pairs). We found that higher musical perception skills led to higher accuracy for non-minimal pairs and minimal pairs differing in their initial consonant. Interestingly, this was not the case for vowel minimal pairs. We discuss the results in relation to theories of second language word learning such as the Second Language Perception model (L2LP).

Keywords: music perception, pitch, phonological processing, cross-situational word learning, auditory perception

INTRODUCTION

Music and language are universal to humans (Patel, 2003) and the connection between the two has been an object of research for centuries, with early ideas even suggesting that music is a spin-off of language in evolution (Pinker, 1997). While the precise origins of music and language remain unclear, there are many parallels that can be drawn between the two. Both use a rule-based hierarchical structure organized into discrete elements and sequences (Tervaniemi et al., 1999; Tervaniemi, 2001; Patel, 2003; Degé and Schwarzer, 2011; Burnham et al., 2015), such as syllables, words, and sentences for language and single notes, intervals, chords, and musical phrases for music (Ong et al., 2016). When focusing on the acoustic characteristics of music and speech sounds, similarities can be found in the reliance on segments of rhythm and harmony alternated with silence, pitch, acoustic envelope, duration, and fundamental frequency (Varnet et al., 2015). In order to understand music and speech, a listener needs to categorize sounds into meaningful units. For speech, perceptual skills are needed to distinguish sounds into separate vowels or consonants and for music into pitches (Hallam, 2017). The auditory skills needed to process language are similar to those needed to discriminate between rhythms (Lamb and Gregory, 1993), harmonies, and melodies (Barwick et al., 1989; Lamb and Gregory, 1993; Anvari et al., 2002). Numerous studies support the overlap of auditory processes involved in music and speech perception (Overy, 2003; Tallal and Gaab, 2006; Patel and Iversen, 2007; Sammler et al., 2007; Wong and Perrachione, 2007; Chandrasekaran et al., 2009; Kraus and Chandrasekaran, 2010; Besson et al., 2011; Rogalsky et al., 2011; Schulze et al., 2011; Bidelman et al., 2013; Gordon et al., 2015; Kraus and White-Schwoch, 2017) and individuals with musical training appear to be advantaged in these shared processes (Krishnan et al., 2005; Bigand and Poulin-Charronnat, 2006; Krizman et al., 2012; White-Schwoch et al., 2013; Elmer et al., 2014).

Those that are expert listeners in either music or language have been found to show *cross-domain transfer* (Ong et al., 2016), where an advantage is found for perception in the other domain; for example, in word segmentation (François et al., 2013), syllabic perception (Musacchia et al., 2007; Ott et al., 2011; Elmer et al., 2012; Kühnis et al., 2013; Chobert et al., 2014; Bidelman and Alain, 2015), receptive and productive phonological skills at the word, sentence and passage level (Slevc and Miyake, 2006), and word dictation (Talamini et al., 2018). It is suggested that long-term expertise in music, which is gained by years of practice, has led to a fine-tuning of the auditory system (Strait and Kraus, 2011a,b), as evidenced by enhanced neural responses to changes in acoustic elements, such as pitch, intensity, and voice onset time (Schön et al., 2004; Magne et al., 2006; Jentschke and Koelsch, 2009; Marie et al., 2011a,b). Musicians indeed show enhanced cortical processing of pitch in speech compared to nonmusicians (Magne et al., 2006; Besson et al., 2007; Musacchia et al., 2007; Kraus and Chandrasekaran, 2010). These and numerous other studies support the idea of cross-domain transfer between music and speech perception (see Hallam, 2017 for an extensive list). The present study focuses on the potential auditory processing advantages in pitch perception and auditory working memory (Ott et al., 2011; Kühnis et al., 2013; Pinheiro et al., 2015; Dittinger et al., 2016, 2017, 2019) associated with music perception skills. Many examples of the effect of music training on speech processing have been reported. For instance, training in music has been associated with phonological perception in the native language (L1; Zuk et al., 2013) and with fluency in a second language (L2; Swaminathan and Gopinath, 2013; Yang et al., 2014). As well, longitudinal studies in children's speech perception found positive effects of music training (Moreno et al., 2009; Degé and Schwarzer, 2011; François et al., 2013; Thomson et al., 2013). Regarding the transfer of music experience to word learning, Dittinger et al. (2016, 2017, 2019) presented listeners with unfamiliar Thai monosyllabic words and familiar visual referents during a learning phase and tested them on their ability to match the words with their corresponding visual objects. Overall, they found that both music training led to higher accuracy in both young adults and children. Additionally, a longitudinal effect of music training was shown, as musicians had the same advantage when tested 5 months later (Dittinger et al., 2016).

However, counter-examples to a positive association between music training and speech perception also exist (Ruggles et al., 2014; Boebinger et al., 2015; Swaminathan and Schellenberg, 2017; Stewart and Pittman, 2021). For instance, Swaminathan and Schellenberg (2017) found that rhythm perception skills predicted English listeners' discrimination of Zulu phonemic contrasts, but only for contrasts that closely resembled English phonemic contrasts. The authors found no association between other music perception skills, such as melody perception or general music training and non-native speech perception, suggesting that an effect of rhythm rather than pitch is related to participants' native language background rather than their music skills. Specifically, unlike for tonal languages, English does not contrast pitch for signaling lexical meaning; hence, it is likely that listeners focus on other cues, such as temporal cues, to distinguish one word from another.

Apart from the ability to perceive novel or familiar phonological contrasts, another important component involved in speech processing, including novel word learning, is working memory. Working memory, which is a short-term memory involved in immediate conscious perceptual and linguistic processing, plays an important role in novel word learning (Gathercole et al., 1997; Warmington et al., 2019). Mixed results have been found regarding a musician's advantage in working memory, with some studies finding no difference between musicians and nonmusicians (Hansen et al., 2012), whereas others find improved auditory and verbal working memory for musicians compared to nonmusicians (Parbery-Clark et al., 2011; Bergman Nutley et al., 2014). A meta-analysis conducted by Talamini et al. (2017) on different types of memory found a medium effect size for short-term and working memory with musicians performing better than nonmusicians, depending on the type of stimulus used.

Most studies examining the link between speech processing and musical abilities have compared professional musicians to nonmusicians (see Zhu et al., 2021), with a large focus on explicit tasks when comparing linguistic and musical abilities (e.g., Dittinger et al., 2016, 2017, 2019). In such tasks, there is no ambiguity during learning, but the link between words and meaning in daily life is much more ambiguous without immediate clear connections, with studies showing that pairing between words and their referent objects are learned by tracking co-occurrences through repeated exposure (e.g., Smith and Yu, 2008; Escudero et al., 2016b; Mulak et al., 2019). Very little is known about the role of musical abilities for ambiguous word learning scenarios, which are most common in everyday life of word learning (Tuninetti et al., 2020). In the realm of music perception, recent studies have shown that musical elements, such as musical grammar

(Loui et al., 2010), harmony (Jonaitis and Saffran, 2009), musical expectation (Pearce et al., 2010), and novel pitch distributions from unfamiliar musical scales (Ong et al., 2017a; Leung and Dean, 2018), can be learned through statistical learning. Statistical learning is a domain-general learning mechanism leading to the acquisition of statistical regularities in (in this case auditory) input. This type of learning may lead to cross-domain transfer between music and language due to learners showing sensitivity toward particular acoustic cues (e.g., pitch; Ong et al., 2016) which may result in improved ambiguous word learning. Despite the potential effect of music abilities on ambiguous word learning and the many types of learners considered in statistical word learning studies (such as young infants, children and adults, and L2 learners Yu and Smith, 2007; Smith and Yu, 2008; Suanda et al., 2014; Escudero et al., 2016b,c; Mulak et al., 2019), participants' musical experience or expertise have yet to investigated. In sum, it has been established that music and language rely on similar general auditory processing skills and, although results are mixed, the majority of studies finds an advantage for music training on auditory and speech perception. By testing whether music abilities in a nonmusician population can help ambiguous word learning, we can further unravel more influences of music on language learning than previously shown.

The current study tests the effect of specific music perception abilities on statistical learning of novel words in a nonmusician adult population. We tested musical abilities through two adaptive psychometric tests targeting specific music perception skills, namely, the ability to perceive fine-pitch mistuning, through the Mistuning Perception Test (MPT; Larrouy-Maestri et al., 2018, 2019), and the ability to discriminate between pitch sequences, through the Melodic Discrimination Test (MDT; Harrison et al., 2017; Harrison and Müllensiefen, 2018). The MPT is an adaptive psychometric test measuring sensitivity to intonation accuracy in vocal musical performance (Larrouy-Maestri et al., 2018, 2019). Perception of vocal mistuning is a core musical ability, as evidenced by its high correlation with other musical traits (Law and Zentner, 2012; Kunert et al., 2016; Larrouy-Maestri et al., 2019), and its importance when judging the quality of a musical performance (Larrouy-Maestri et al., 2019). The MDT aims to test melodic working memory, as it requires melodies to be held in auditory working memory in order for participants to compare and discriminate them correctly (Dowling, 1978; Harrison et al., 2017; Harrison and Müllensiefen, 2018). To do well in these tasks, specific auditory processing skills, in particular pitch perception and auditory working memory, are required. A recent large-scale study across thousands of speakers of tonal, pitch-accented, and non-tonal languages using these two tasks (and a beat alignment task) has shown that language experience shapes music perception ability (Liu et al., 2021). Here, we test the opposite, namely, whether the same music perception skills help with language learning, and specifically when learning novel words with different degrees of phonological overlap. Our specific focus is on pitch processing abilities but acknowledge that rhythm processing is also an important component in music and language processing (see Swaminathan and Schellenberg, 2017).

To test whether pitch perception and auditory working memory are helpful when learning words in ambiguous scenarios, we used a cross-situational word learning (CSWL) paradigm in which meanings of new words are learned through multiple exposures over time without explicit instruction, where learning of word-object pairings can only take place through their statistical co-occurrences (e.g., Escudero et al., under review; Yu and Smith, 2007; Kachergis et al., 2010; Smith and Smith, 2012; Escudero et al., 2016a,b, 2021; Mulak et al., 2019; Tuninetti et al., 2020). Early CSWL experiments focused on words with very little phonological overlap (e.g., Smith and Yu, 2008; Vlach and Johnson, 2013), where a listener can rely on other cues to learn the novel words and does not have to focus on the fine phonological details of each word (Escudero et al., 2016b). Therefore, (Escudero et al., 2016a,b) and Mulak et al. (2019) studied CSWL of monosyllabic non-minimal and minimal pairs, differing only in one vowel or consonant, to test whether listeners can encode sufficient phonological detail in a short time to learn these difficult phonological contrasts. It was found that accurate phonological encoding of vowel and consonant contrasts predicts high performance in CSWL tasks (Escudero et al., 2016a; Mulak et al., 2019).

In the present study, we thus tested whether musical ability impacts word learning of phonologically overlapping words using Escudero et al. (2016b) and Mulak et al. (2019)'s CSWL paradigm. Overall, we hypothesize that those with stronger musical abilities are better at perceiving speech sounds due to enhanced pitch perception and working memory, and that will be reflected in higher accuracy overall in the CSWL task. We may also see differences in how well vowels and consonants are learned, due to higher acoustic variability in vowels compared to consonants (Ong et al., 2015), which may favor learners with stronger pitch perception skills.

MATERIALS AND METHODS

Participants

Fifty-four participants took part in the study and were tested online, which is our common practice since the start of the COVID-19 pandemic, using our validated online testing protocols (Escudero et al., 2021). In Escudero et al. (2021), we compared online and face-to-face testing using the same CSWL design and online testing results were found to be very similar to results from the laboratory. Ten participants were excluded from the analysis due to technical difficulties, mostly internet dropouts during the experiment or excessive environmental noise, leading to a total participant sample of 44 ($M_{age} = 26.79$, $SD_{age} = 11.12$, 33 females). Participants were recruited through the Western Sydney University's online research participation system (SONA) or *via* word-of-mouth and participation was rewarded with course credit for the former and voluntary for the latter. Written informed consent was obtained online from all participants prior to the start of the experiment, and the study was approved by the Western Sydney University Human Research Ethics Committee (H11022).

Materials

Questionnaires

The questionnaires conducted at the beginning of the experiment consisted of two parts: a language and a musical background questionnaire. The language background questionnaire consisted of questions aimed to get detailed information regarding participants native (and other) language, as well as the language background of their parents/caretakers. The musical background questionnaire is the Goldsmiths Musical Sophistication Index (GMSI; Müllensiefen et al., 2014), which aims to collect wide-range data related to one's engagement with music (e.g., music listening and music performance behavior). Both questionnaires were administered through Qualtrics (Qualtrics, Provo, UT). From the GMSI, 23 participants indicated having zero years of experience with playing an instrument, and seven had 10 or more years of experience. From the language questionnaire, we found that 17 were Australian English monolinguals and 27 were bi- or multilinguals.

Cross-Situational Word Learning

All words and visual referents have been used in prior CSWL studies (Vlach and Sandhofer, 2014; Escudero et al., 2016a,b; Mulak et al., 2019; Escudero et al., under review). Novel words consisted of eight monosyllabic nonsense words recorded by a female native speaker of Australian English and followed a consonant-vowel-consonant (CVC) structure while adhering to English phonotactics. The stimuli were produced in *infant-directed speech* (IDS) as we are replicating previous studies that used IDS to compare adult and infant listeners and included two tokens for each word to match prosodic contours across all stimuli (Escudero et al., 2016a,b).

The eight words were combined into minimal pair sets to form specific consonant or vowel minimal pairs or non-minimal pairs. The two types of minimal pairs featured words that either differed in their initial consonant (consMPs; e.g., BON-TON) or in their vowel (vowelMPs; e.g., DIT-DUT). Non-minimal pairs were formed by pairing two words from each of the two minimal pair types in random order (nonMPs; e.g., BON-DIT).

Every novel word was randomly paired with a color picture of a novel item, which is not readily identifiable as a real-world object. These word-referent pairings were the same for all participants. An overview of the novel words and visual referents is presented in **Figure 1**.

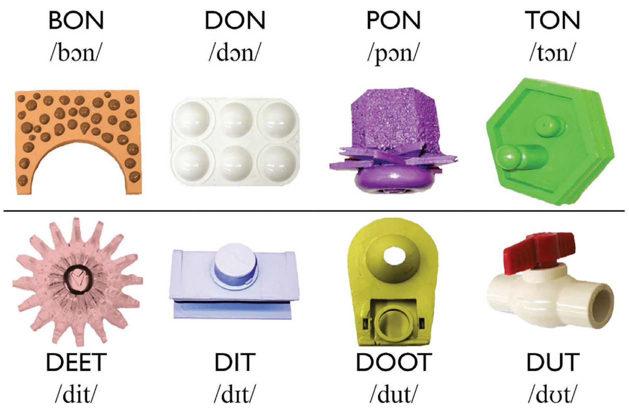


FIGURE 1 | The eight novel words and their visual referents. The four words in the top row are minimally different in their initial consonant, whereas the words on the bottom are minimally different in their vowel. The vowel used for the consonant minimal pairs is/O/as in POT. Vowels used for the vowel minimal pairs are/i/as in BEAT, /I/as in BIT, /u/as in BOOT, and/U/as in PUT.

Mistuning Perception Test

The MPT, which is an adaptive psychometric test, uses short excerpts (6–12 s) of musical stimuli from pop music performances which are representative of real-life music and are therefore ecologically valid (from MedleyDB; Bittner et al., 2014). The test highly correlates with low- and high-level pitch perception abilities, such as pitch discrimination and melody discrimination, and thus provides an assessment of important pitch processing abilities (Larrouy-Maestri et al., 2019). In a two-alternative forced-choice task, participants were presented with a pitch-shifted version (out-of-tune) and the normal version (in-tune) of a stimulus and were asked to indicate which version was out-of-tune. Pitch shifting varied from 10 cents to 100 cents, sharp, and flat (for more details about the construction of the MPT, see Larrouy-Maestri et al., 2019). Before starting the task, participants received an example of an out-of-tune and an in-tune version. A demo of the experiment can be found on https://shiny.gold-msi.org/longgold_demo/?test=MPT.

Melodic Discrimination Test

Similar to the MPT, the MDT is also an adaptive psychometric test. The MDT is developed to test one's ability to discriminate between two melodies (Harrison et al., 2017; Harrison and Müllensiefen, 2018). Participants are presented with a three-alternative forced-choice (3-AFC) paradigm where they listen to three different versions of the same melody, each with a different pitch height (musical transposition), and with one containing an *altered* note produced by changing its relative pitch compared to the base melody (Harrison et al., 2017), resulting in a pitch height change for one note compared to the other melodies. Each melody can be altered using four pre-determined constraints: (1) melodies with five notes or fewer cannot have the first nor last note altered, (2) melodies with six notes or longer cannot have the first two nor last two notes altered, (3) the note cannot be altered by more than six semitones, and (4) the altered not must be between an eight note and a dotted half note in length (see Harrison et al., 2017). Participants are asked to indicate which of the three melodies are the odd one out. Participants heard an implementation of the MDT with 20 items (see doi:10.5281/zenodo.1300951) using the shiny package in R (Chang et al., 2020) which uses an adaptive item selection procedure with each participant's performance level determining the level of difficulty of item presentation. Performance level is estimated using Item Response Theory (de Ayala, 2009). A demo of the experiment can be found on https://shiny.gold-msi.org/longgold_demo/?test=MDT. Tests scores for both the MDT as the MPT are computed as intermediate and final abilities with weighted-likelihood estimation (Warm, 1989) and using Urry's rule for item selection (Magis and Raîche, 2012).

Procedure

We followed our adult online testing protocol, which was validated in Escudero et al. (2021), for details please see on https://osf.io/nwr5d/. In short, participants signed up for a timeslot on SONA after which they received an email with specific instructions for the experiment (e.g., wearing headphones and participating from a silent study space with no background noise was required) and an invitation for a Zoom call. Participants unable to meet the participation requirements were excluded from the analysis (see Section "Participants"). During the Zoom call, participants were first familiarized with the procedure and then sent links to the consent forms, background questionnaires, and the experiment. During the experiment, they were asked to share their screen and computer audio throughout the entire video call, apart from when filling out the questionnaire to ensure privacy. Participants' screen and audio sharing enabled experimenter's verification of appropriate auditory stimuli presentation and participants' attention. The experimenter was on mute and with their video off during the experiment to avoid experimenter bias.

Participants first completed the language and musical background questionnaires and were then instructed to start the CSWL task. The CSWL task consisted of a learning and a test phase set up in PsychoPy 3 (Peirce, 2007; Peirce et al., 2019) hosted on Pavlovia.org. Following previous CSWL studies, minimal instruction was provided (i.e., "Please listen to the sounds and look at the images") prior to the learning phase. During the learning phase, participants saw 24 trials each consisting of two images accompanied by auditory representations of two words without indication of which word corresponded to which image. The visual referents were presented first for 0.5 s before the onset of the first word. Both words lasted for 1 s and were followed by a 0.5 s inter-stimuli interval (ISI). After this, a 2 s inter-trial interval (IT) consisting of a blank screen was then presented, leading to a total trial time of 5 s. The learning phase was directly followed by a test phase of 24 trials, for which participants were told that they would be tested on what they have learned and to indicate their answers by pressing specific keys on the keyboard. Every test trial presented two possible visual referents simultaneously on the screen for 3 s. During this, participants heard one spoken target word four times (with alternating tokens of the words) and were then asked to indicate which visual referent (the left or the right one) corresponded with the target word by pressing a key on the keyboard any time after the onset of the target word. Trial order was randomized across all participants. The presentation of left and right of the visual referents was counterbalanced and resulted in two between-subject learning conditions. A blank screen of 2 s was presented in between trials. Directly after the CSWL task, participants completed the MDT and the MPT task to measure their music perception abilities.

STATISTICAL ANALYSIS

We used a Bayesian Item Response Theory (IRT) model to analyze accuracy. IRT models are particularly useful for predicting the probability of an accurate answer depending on an item's difficulty, its discriminability, a participant's latent ability, and a specified guessing parameter (Bürkner, 2020), which provides a lower bound for the model's predictions. The statistical analyses were run in the statistical program R (R Core Team, 2020) with the brms package using Stan (Bürkner, 2017, 2018; R Core Team, 2020).

We used approximate leave-one-out (LOO) cross-validation to find the model that generalizes best to out-of-sample data. Additionally including GMSI or participant's language background did not improve the out-of-sample predictions of the model.

The best model included only the interaction between Pair type and MPT. However, as we are interested in both MPT and MDT as main factors, we will report the next best model. The difference in the LOOIC values for these two models is negligible. Prior to fitting the models, we tested for correlation between MPT, MDT, and GMSI. MPT and MDT were moderately positively correlated, $r(1054) = 0.39$, $p < 0.005$; MPT and GMSI were moderately positively correlated, $r(1054) = 0.30$; and MDT and GMSI were weakly positively correlated, $r(1054) = 0.11$.

Accuracy was modeled as a binary response variable, with 0 for inaccurate and 1 for accurate. We used a 4-parameter non-linear logistic model (4PL, Agresti, 2010) on the Bernoulli distribution with an item, a person and a guessing parameter. The discriminability parameter is removed. The item parameter models the difficulty of the tested items (in this case the pair types); the person parameter models the individual ability of each participant. The guessing parameter represents the probability of being accurate if participants were only guessing (Bürkner, 2020). All of our trials are binary forced choice; hence, we use a fixed guessing parameter of 0.5. An advantage of using IRT for modeling binary accuracy responses is that this probability can be taken into account as a type of baseline in the model, meaning that the model's estimates of the underlying probability of being correct will not fall below the 0.5 threshold. We did not include a discrimination parameter, as all tested items are very similar.

The categorical variable Pair type was turned into a factor and modeled using dummy coding, which is the default in R. For MPT and MDT, we are using the raw data scores, as recommended by the experiment designers (MPT: Larrouy-Maestri et al., 2018, 2019; MDT: Harrison et al., 2017; Harrison and Müllensiefen, 2018), which were computed from the underlying item response models. These scores range from −4 to +4. GMSI was scaled and centered to a previously determined population mean from Harrison and Müllensiefen (2018).

For the 3-PL IRT accuracy model, we included separate priors for the item, person and guessing parameters. As detailed below, all such priors were weakly informative in that they weakly favor an effect of zero size and disfavor unfeasibly large effects. The following model formula (including priors) was run in R:

```
Accuracy ~ 0.5 + 0.5 * inv_logit(eta),
Eta ~ 1 + Pairtype * (MDT ability + MPT
ability) + (1|item) + (1|participant),
nl = TRUE)
family <- brmsfamily("bernoulli,"
link = "identitiy").
priors<-
prior("normal (0,5)," class = "b," nlpar = "eta") +
prior("constant(1)," class = "sd,"
group = "participant,"  nlpar = "eta") +
prior("normal(0,3)," class = "sd," group = "item,"
nlpar = "eta").
```

An important aspect of Bayesian regression is that it calculates the whole posterior distribution of each effect, which allows for the calculation of credibility intervals. In contrast with frequentist confidence intervals, credibility intervals indicate the 95% certainty that reported effect falls within the range of the interval (Smit et al., 2019). Evidence for a hypothesized effect will be assessed through evidence ratios, which quantify the likelihood of a tested hypothesis against its alternative (Bürkner, 2017, 2018). We consider evidence ratios of >10 to be strong evidence and above >30 to be very strong evidence [see Jeffreys (1998), as cited by Kruschke (2018)]. For directional hypotheses, where the predicted direction of an effect is given, effects with evidence ratios of >19 are roughly similar to an alpha of 0.05 in null-hypothesis significance testing (NHST; Makowski et al., 2019; Milne and Herff, 2020).

We expect that high musical perception abilities transfer to stronger phonological processing which subsequently translates to higher performance in the CSWL task (as evidenced by higher accuracy), compared to those with less musical perception abilities. With regards to the three tested pair types, we expect them to follow the same pattern as in previous CSWL studies, namely, a higher performance for nonMPs and consMPs and lower performance for vowelMPs (Escudero et al., 2016a). Additionally, we were interested in the differences between the moderations of MPT and MDT per pair type. As the MPT tests for perception of fine-pitch changes, one might expect participants with higher MPT scores to learn vowel contrasts more easily due to the acoustic similarities between musical pitch and vowels. As MDT measures auditory short-term memory, we expect high MDT scores to positively correlate with accuracy in general.

RESULTS

Figure 2 shows the overall percentage of accurate responses per pair type. Performance across pair types appears to be very similar and participants were able to learn all pair types during the task, as evidence by performance being significantly above chance (see **Figure 2**). Accuracy for these learners is similar, albeit a little lower, to that found in a previous study (between 0.60 and 0.70 for all pair types) using the exact same design and online testing methodology (Escudero et al., 2021).

Hypothesis tests run on the results from the multilevel Bayesian model show strong evidence that for participants with average MDT and MPT, accuracy for consMPs is lower than for nonMPs (see **Table 1**, hypothesis 1). We did not find sufficient evidence to support a difference between the other pair types (hypotheses 2 and 3). We then tested whether performance per pair type is moderated by MPT and MDT ability. As shown in **Figure 3**, mean accuracy for nonMPs does not appear to be moderated by MPT ability, whereas for consMPs, higher MPT ability leads to higher accuracy, which was not expected. Also unexpectedly, the opposite occurs for vowelMPs, where higher MPT ability negatively impacts performance. As per our predictions, for MDT ability (see **Figure 4**), we see that higher scores generally lead to improved accuracy, especially for nonMPs and vowelMPs. However, important to note is that, as visualized by the colored ribbons in **Figures 3, 4**, the slopes' credibility intervals are highly overlapping, which indicates that the evidence for these differences might not be decisive. Therefore, we conducted hypothesis testing to confirm this (see hypotheses 4–6 for MPT ability and 10–12 for MDT ability in **Table 1**). As can be seen in **Table 1**, MDT ability influences accuracy in the expected direction (i.e., higher MDT leads to higher accuracy) for all pair types, but unexpectedly, MPT has a negative effect on accuracy for vowelMPs.

Regarding the extent to which the effect of MPT and MDT differs by pair type, unexpectedly, we find very strong evidence that MPT ability has a stronger impact on accuracy for consMPs than for nonMPs and vowelMPs (see **Table 1**; hypotheses 7 and 9) and strong evidence for nonMPs compared to vowelMPs (see **Table 1**; hypothesis 8). Thus, not only does MPT negatively influence the learning of vowelMPs as shown in hypothesis 4, but it also impacts the learning of vowelMPs less strongly than the learning of nonMPs and consMPs. Our finding of strong evidence suggesting that MDT ability has a stronger impact on accuracy for nonMPs and vowelMPs compared to consMPs (see **Table 1**; hypotheses 13 and 15) was also unexpected, as we thought MDT would influence the learning of all pair types equally.

DISCUSSION

In this study, we tested whether music perception abilities impact the learning of novel word pairs in a CSWL paradigm that provides no explicit instruction during the learning phase. Overall, we found that participants were able to learn all novel word-object pairings regardless of the phonological overlap between the novel words, mostly replicating (albeit a little lower) previous reported results using the same online protocol (Escudero et al., 2021). That is, overall accuracy was comparable for novel words that had large phonological differences, forming non-minimal pairs (nonMPs), and for words that differed in a single consonant (consMPs) or a single vowel (vowelMPs). Regarding the relation between accuracy and music perception abilities, participants with average MPT and MDT had similar word learning scores across pair types, with performance for consMP probably being slightly lower than for the other pair types. Crucially, we found unexpected results for how MPT and MDT influenced

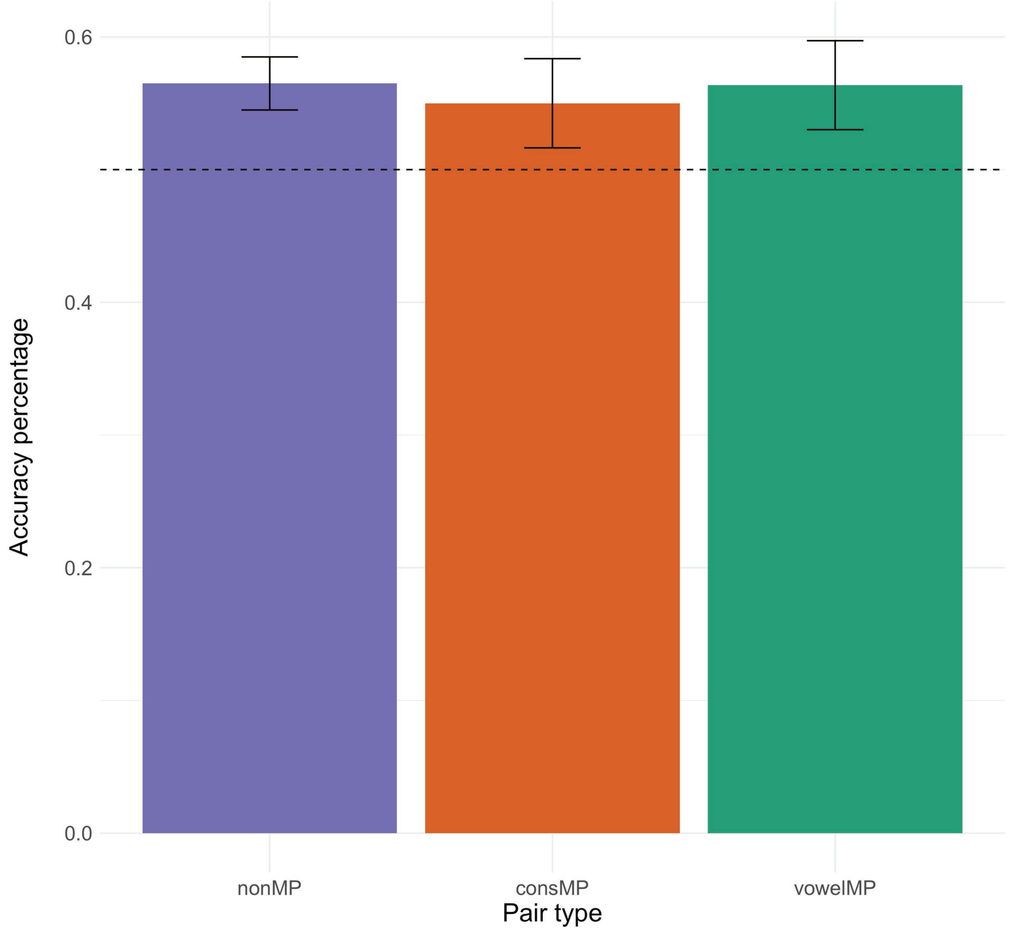


FIGURE 2 | Mean accuracy (in percentage) per pair type. Error bars represent the standard error over the mean accuracy responses per pair type. The dotted line represents accuracy by chance.

word learning performance in nonMPs versus consonant and vowelMPs, which we discuss below.

As mentioned above, although we expected higher MPT participants to learn vowel contrasts more easily due to the acoustic similarities between pitched musical sounds and vowels (consonants do not have a clear pitch), we found the opposite effect. It appears that stimuli containing variable pitch information (such as vowels) pose extra difficulty for listeners who are more attuned to such information. A plausible explanation for these results is proposed by Ong et al. (2017b) who suggest that listeners' experience is important for their ability to learn new acoustic cues, whether this experience is linguistic (through a native language that distinguishes lexical tone contrasts, such as Cantonese, Mandarin, or Thai) or musical. In a *distributional learning* (a form of statistical learning) experiment of nonnative lexical tones, they found that listeners without music or tonal language experience were able to discriminate lexical tones from ambiguous versions of the target tones after a short exposure (Ong et al., 2015). In a follow-up study, they found mixed results for *pitch experts*, who they define as listeners with extensive experience with pitch either through a tonal language or through musical training. Those with a tonal language background were able to learn non-native lexical tones distributionally but those with a musical background were not. This was unexpected as musical training has been found to have a positive effect on statistical learning (e.g., François et al., 2013; Chobert et al., 2014), and musicians were expected to perform better due to an improved ability to extract regularities from the input. These results led Ong and colleagues to conclude that domain-specific experience with pitch influences the ability to learn non-native lexical tones distributionally (Ong et al., 2017b), indicating no cross-domain transfer of music and linguistic abilities in distributional learning.

Ong and colleagues discussed their results in relation to the Second Language Perception (L2LP) model (Escudero, 2005; van Leussen and Escudero, 2015; Elvin and Escudero, 2019; Elvin et al., 2020, 2021; Yazawa et al., 2020), suggesting that the tonal language speakers only had to shift their category boundaries to the novel tonal categories, whereas the musicians had to create new categories, which is more difficult (Ong et al., 2017b). Another possible explanation is that musicians did not consider the stimuli as speech tones and thus may have processed them as musical stimuli resulting in them not learning the tonal categories (Ong et al., 2017b), but this argument assumes that musical pitch cannot be learned distributionally. In a different study, Ong et al. (2017a) tested distributional learning of musical pitch with nonmusicians and showed that they were able to acquire pitch from a novel musical system in this manner. This may be different for musicians, who were found to outperform nonmusicians in the discrimination and identification of Cantonese lexical tones (Ong et al., 2020).

TABLE 1 | Hypothesis testing—accuracy model.

Hypothesis tests	Estimate	Est. Error	[90% CI]	Evid. Ratio	Post. Prob
For average MDT and MPT ability:					
1. nonMP–consMP > 0	−1.83	1.67	[−4.88, 0.25]	11.11	0.92
2. nonMP–vowelMP > 0	−0.63	1.80	[−3.90, 1.21]	0.64	0.39
3. vowelMP–consMP > 0	1.20	2.47	[−2.63, 4.87]	3.27	3.27
***MPT ability > 0* in the following conditions and contrasts:**					
4. nonMP	0.41	0.70	[−0.50, 1.76]	2.47	0.71
5. consMP	2.97	1.48	[0.80, 5.55]	91.78	0.99
6. vowelMP	−0.88	0.85	[−2.08, 0.17]	12.10	0.92
7. consMP–nonMP	2.55	1.53	[0.28, 5.20]	30.61	0.97
8. nonMP–vowelMP	1.30	1.05	[−0.07, 3.04]	16.37	0.94
9. consMP–vowelMP	3.85	1.69	[1.38, 6.68]	92.75	0.99
***MDT ability > 0* in the following conditions and contrasts:**					
10. nonMP	0.95	0.42	[0.28, 1.63]	78.30	0.99
11. consMP	−0.08	0.84	[−1.45, 1.11]	0.92	0.48
12. vowelMP	1.16	0.93	[−0.07, 2.65]	16.33	0.94
13. nonMP–consMP	1.04	0.89	[−0.24, 2.52]	10.06	0.91
14. vowelMP–nonMP	0.21	0.98	[−1.78, 1.14]	1.44	0.59
15. vowelMP–consMP	1.24	1.23	[−0.54, 3.29]	7.63	0.88

Estimate = mean of the effect's posterior distribution. Estimate error = standard deviation of the posterior distribution. 90% CI = 90% credibility intervals. Evidence ratio = the posterior probability under the hypothesis against its alternative.

From studies on distributional learning of pitch and lexical tones, it can be concluded that cross-domain transfer between speech and music largely depends on the listener's musical or linguistic experience (Ong et al., 2015, 2016, 2017a,b, 2020). Nonmusicians without tonal language experience can learn novel pitch contrasts in both the speech and the music domain, but the situation is more complex for pitch experts, suggesting that those with extensive music experience may struggle more than those with tonal experience. However, an important difference between Ong et al.'s studies and the current study is that they tested listeners at both ends of the experience spectrum, while we tested listeners ranging from the lower to middle end of the music experience spectrum based on their music perception skills. By using music perception tasks, we were able to classify participants using a continuous predictor rather than splitting them into groups, which allowed us to uncover more detailed information about what happens with speech learning as music perception skills increase. A further difference is in the stimuli used, as the lexical and musical tones used in Ong et al. (2015, 2016, 2017a,b, 2020) contained many variable pitches along a continuum, while our stimuli had limited and uncontrolled pitch variation. Specifically, we focused

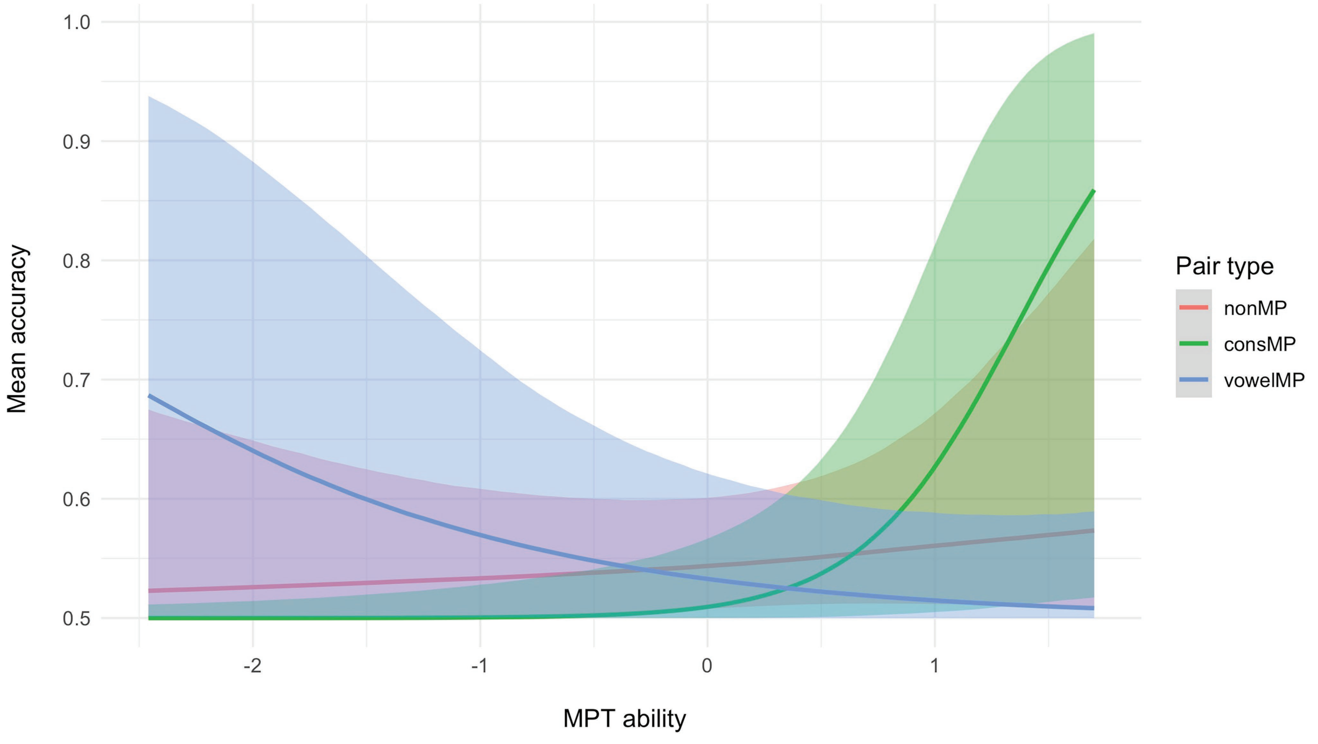


FIGURE 3 | Conditional effects of MPT ability and pair type on mean accuracy with 95% credibility intervals.

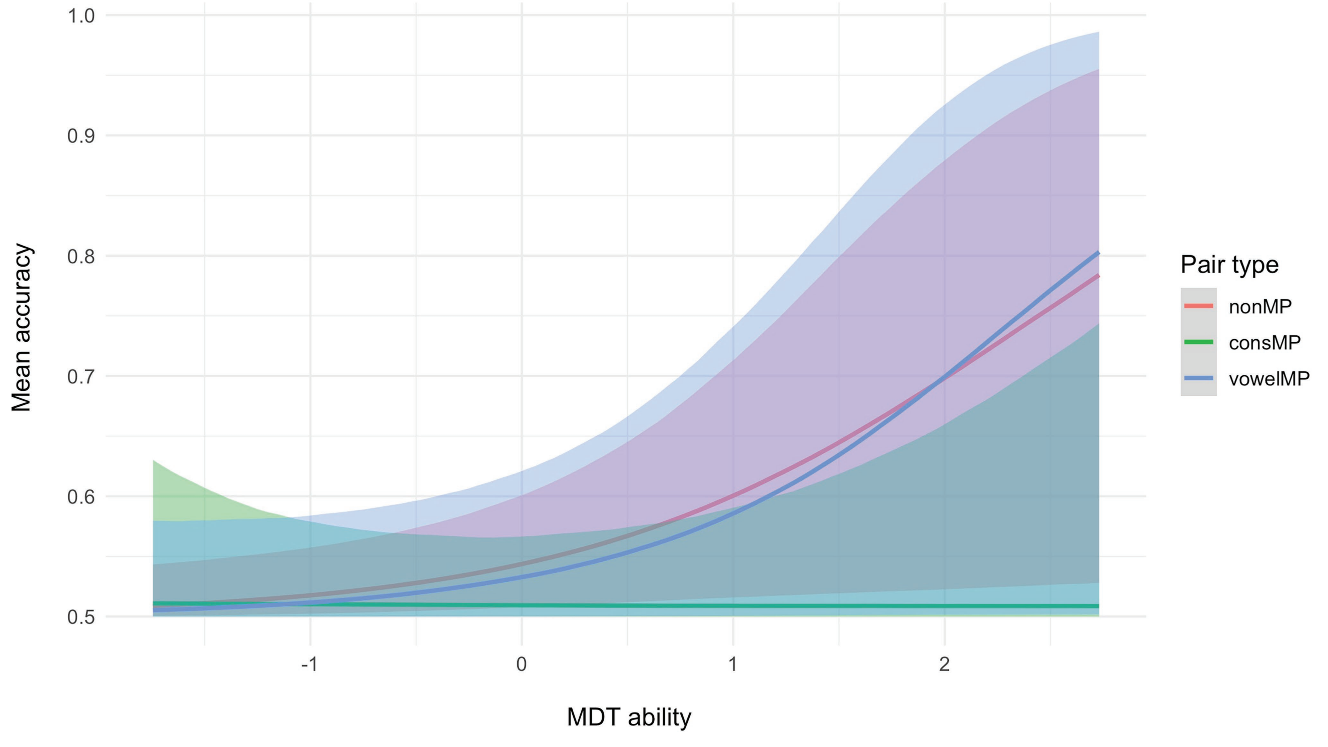


FIGURE 4 | Conditional effects of MDT ability and pair type on mean accuracy with 95% credibility intervals.

on word learning of naturally produced novel words, where pitch variability was not consistent among the different words and pair types. Thus, listeners in the present study may have used other acoustic cues that are not pitch-related to discriminate and learn the novel words.

Given that listeners with strong pitch perception abilities are more likely to use pitch as a cue to discriminate between stimuli (Perfors and Ong, 2012; Ong et al., 2017b, 2020), our vowelMP stimuli may have been particularly challenging for them due to the use of infant-directed speech (IDS). IDS is the speech style or register typically used by mothers and caregivers when speaking to babies and is characterized by the use of larger pitch variations. Many studies have shown that IDS can facilitate word learning in infants (Ma et al., 2011; Graf Estes and Hurley, 2013) and adults (Golinkoff and Alioto, 1995) due to higher salience leading to enhanced attentional processing (Golinkoff and Alioto, 1995; Kuhl et al., 1997; Houston-Price and Law, 2013; Ellis, 2016). Despite it facilitating infant and adult speech learning, IDS may have a negative effect for those with strong musical perception abilities as they might think they are hearing different words due to varying pitch contours when only one word is presented. Unexpectedly, MPT ability affected learning of cMPS and nonMPs more than vMPs. As vMPs naturally contain more pitch variation, those were expected to be the most difficult to learn, hence the influence of IDS is likely stronger on cMPS and nonMPs than on vMPs. A similar result of hearing multiple words instead of one due to the use of IDS has been found in a prior CSWL study (Escudero et al., under review), where the target population consisted of native Mandarin speakers who were L2 English learners. Specifically, word pairs containing non-native vowel contrasts with IDS pitch fluctuations were difficult to learn for L1 Mandarin L2 English learners.

Thus, in populations where pitch variations indicate different lexical meanings, such as native speakers of Mandarin (Han, 2018), IDS can be problematic and impair word learning as participants might perceive multiple categories where only one is presented (Escudero and Boersma, 2002; Elvin et al., 2014; van Leussen and Escudero, 2015). The impact of a learner's native language on novel language learning has been explained by L2 speech theories (e.g., Flege, 1995; Escudero, 2005; Best and Tyler, 2007; van Leussen and Escudero, 2015). In particular, the L2LP model (Escudero, 2005; van Leussen and Escudero, 2015; Elvin and Escudero, 2019; Elvin et al., 2020, 2021; Yazawa et al., 2020) proposes three learning problems when L1 and L2 categories differ in number or in phonetic realization. This model is the only one that handles lexical development and word learning with consideration of hearing more differences than produced in the target language as a learning problem (van Leussen and Escudero, 2015; Escudero and Hayes-Harb, 2021). Specifically, listeners can categorize binary L2 contrasts into more than two L1 categories, which is referred to as Multiple Category Assimilation (MCA, L2LP; Escudero and Boersma, 2002) and can lead to a *subset problem* (Escudero and Boersma, 2002; Escudero, 2005; Elvin and Escudero, 2014, 2019). A subset problem occurs when an L2 category does not exist in a listener's L1 but is acoustically similar to two or more separate L1 categories and thus is perceived as more than one L1 sound, with no overt information from the target language that will allow the learner to stop hearing the extra category or stop activating *irrelevant* or *spurious* lexical items (Escudero and Boersma, 2002; Escudero, 2005; Elvin and Escudero, 2014, 2019).

With regard to our CSWL task, we expect that using *adult-directed speech* (ADS) without these additional pitch fluctuations would improve learning for the nonMPs and consMPs for tonal language speakers, but not for vowelMPs. When using IDS, nonmusicians and non-tonal speakers show a pattern where performance is lowest for pair types with the highest pitch variability (i.e., vowelMPs). The use of IDS, which adds even more pitch variability than naturally present in the vowelMPs, seems to pose problems for learners who are not music experts but have some music perception skills. For tonal language speakers, the use of IDS poses problems in general as they consistently use pitch information to discriminate between all pair types. If pitch variability is the main predictor for performance in this CSWL task, then music experts (i.e., musicians) should struggle more with the vowelMPs than the nonmusicians tested here but should perform better for the nonMPs and consMPs than the tonal language speakers discussed earlier in Escudero et al. (under review).

Regarding the results for MDT, although not decisive, the evidence suggests that MDT ability more strongly influences accuracy for nonMPs and vowelMPs compared to consMPs. The MDT ability test focuses heavily on auditory short-term memory (Dowling, 1978; Harrison et al., 2017; Harrison and Müllensiefen, 2018). It has been suggested that auditory short-term memory for consonants is distinct from that for vowels (Pisoni, 1975), as explained by the cue-duration hypothesis (Pisoni, 1973), which suggests that the acoustic features needed to discriminate between two different consonants are shorter and thus less well represented in auditory short-term memory than those of vowels (Chen et al., 2020). As well, seminal studies on speech sounds have suggested that consonants may be stored differently in short-term memory compared to vowels (Crowder, 1971, 1973a,b), with the idea that vowels are processed at an earlier stage compared to consonants (Crowder and Morton, 1969). It is possible that a different type of auditory memory is activated for nonMPs, which does not rely as strongly on the discrimination of the acoustic features of the stimuli than what is needed to distinguish between phonologically overlapping stimuli. As similarly suggested in Escudero et al. (2021), this could be tested using time-sensitive neurophysiological methods, such as electroencephalography (EEG).

Some limitations of this study must be noted. Even though we tested for perceptual skills, it is possible that accuracy also depends on other skills, such as how well a listener is able to do crossmodal associations. Likewise, it is possible that general cognitive abilities may impact the learning of novel words in an ambiguous word learning paradigm. As we find some differences between accuracy for the different pair types in the current study and prior CSWL studies using the same paradigm (Escudero et al., 2016; Mulak et al., 2019), it might seem that individual differences, such as the ability to do crossmodal associations or general cognitive abilities, may be the cause of these differences. However, there are other possible sources between the current study and prior CSWL results that might have led to the differences between studies, such as the number of trials and the number of responses used in the learning and test phases. We are currently replicating learning and testing phases from those previous studies using online testing to see if the number of trials is the source of the difference. If this is not the case, future studies can then look further into other possible sources, such as general cognitive abilities. Regarding the use of IDS, it is an empirical question whether adults in general will perform better with stimuli characterized by shorter durations, and non-enhanced differences between vowels and neutral prosodic contours (such as ADS). On the contrary, we found that enhanced vowel differences that are similar to those typical of IDS facilitate phonetic discrimination for adults listeners (Escudero et al., 2011; Escudero and Williams, 2014). Additionally, there is a possibility that the degree of novelty of the auditory and visual stimuli impacts accuracy responses. Even though language background did not have an influence on accuracy, future studies could consider implementing measuring participants' familiarity with the stimuli. Another possible limitation is that we did not collect information regarding participants' headphones. However, we did check whether participants were able to hear the stimuli and were wearing headphones, as part of our pre-registered protocol.

Overall, the results show that the tested music perception abilities impact the learning of words that differ in a single consonant or vowel differently and in complex ways. Pitch perception is an important factor for novel word learning, to the extent that those with stronger pitch perception skills are better at distinguishing consonant contrasts, and apparently *too* good at distinguishing vowel contrasts. Using stimuli produced in adult-directed-speech, our follow-up research will establish whether the negative correlation between pitch perception and accuracy in words distinguished by a single vowel is due to our use of IDS and its concomitant large pitch variations. We also find that consonants and vowels are learned differently for those with melodic discrimination skills, reflected in improved auditory short-term memory. In contrast to MPT, an increase in MDT leads to better learning of words distinguished by a single vowel than those distinguished by a single consonant, which may be connected to better auditory short-term memory for vowels. The contrasting results for the two tested music perception skills may reflect different stages of processing. Our results have one clear implication for theories of cross-domain transfer between music and language: considering populations along the entire spectrum of musicality and linguistic pitch experiences is the only way to uncover exactly where and when problems with word learning occur.

CONCLUSION

We tested whether specific music perception abilities impact learning of minimal pair types in adults that have not been selected for their musical abilities. Using a CSWL paradigm, we have shown that pitch perception and auditory working memory affect the learning of vowel and consonant minimal word pairs, but vowels and consonants are impacted differently. We suggest this may be due to the pitch fluctuations of the specific characteristic of stimuli, namely, words produced in infant-directed speech (IDS). Similar to the patterns observed in native speakers of tonal languages, this type of speech register may lead to the listeners' perception of more distinctions than intended. In future studies, we aim to test the role of IDS compared to adult-directed speech, how specific levels of training in music impact performance in CSWL, and the differential storage of vowels versus consonants.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Western Sydney University Human Research Ethics Committee (H11022). The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ES and PE conceived the initial experiments. ES was responsible for overseeing data collection and wrote the initial draft. ES and AM analyzed the data. ES, AM, and PE wrote the paper. All authors contributed to the article and approved the submitted version.

FUNDING

Data collection was funded by a Transdisciplinary & Innovation Grant from the Australian Research Centre for the Dynamics of Language (project number TIG1112020) awarded to ES. PE's and ES' work and the article publication fees were funded by an Australian Research Council Future Fellowship (FT160100514) awarded to PE. AM's work was funded by an Australian Research Council Discovery Early Career Researcher Award (project number DE170100353).

ACKNOWLEDGMENTS

We would like to acknowledge and thank Deeahn Sako, Madeleine Leehy, and Christopher Piller for their help with data collection and the participants for their time and participation.

REFERENCES

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. New York: John Wiley & Sons.

Anvari, S. H., Trainor, L. J., Woodside, J., and Levy, B. A. (2002). Relations among musical skills, phonological processing, and early reading ability in pre-school children. *J. Exp. Psychol.* 83, 111–130. doi: 10.1016/s0022-0965(02)00124-8

Barwick, J., Valentine, E., West, R., and Wilding, J. (1989). Relations between reading and musical abilities. *Br. J. Educ. Psychol.* 59, 253–257. doi: 10.1111/j.2044-8279.1989.tb03097.x

Bergman Nutley, S., Darki, F., and Klingberg, T. (2014). Music practice is associated with development of working memory during childhood and adolescence. *Front. Hum. Neurosci.* 7:926. doi: 10.3389/fnhum.2013.00926

Besson, M., Chobert, J., and Marie, C. (2011). Transfer of training between music and speech: common processing, attention, and memory. *Front. Psychol.* 2:94. doi: 10.3389/fpsyg.2011.00094

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. *Restor. Neurol. Neurosci.* 25, 399–410. doi: 10.1371/journal.pone.0089642

Best, C. T., and Tyler, M. D. (2007). "Nonnative and second-language speech perception: commonalities and complementaries," in *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*. eds. O.-S. Bohn and M. J. Munro (Amsterdam: John Benjamins), 13–34.

Bidelman, G. M., and Alain, C. (2015). Musical training orchestrates coordinated neuroplasticity in auditory brainstem and cortex to counteract age-related declines in categorical vowel perception. *J. Neurosci.* 35, 1240–1249. doi: 10.1523/jNEUROSCIE.3292-14.2015

Bidelman, G. M., Hutka, S., and Moreno, S. (2013). Tone language speakers and musicians share enhanced perceptual and cognitive abilities for musical pitch: evidence for bidirectionality between the domains of language and music. *PLoS One* 8:e60676. doi: 10.1371/journal.pone.0060676

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

Bittner, R., Salamon, J., Tierney, M., Mauch, M., Cannam, C., and Bello, J. (2014). MedleyDB: a multitrack dataset for annotation-intensive MIR research. Paper presented at the International Society for Music Information Retrieval (ISMIR), Taipei, Taiwan.

Boebinger, D., Evans, S., Rosen, S., Lima, C. F., Manly, T., and Scott, S. K. (2015). Musicians and non-musicians are equally adept at perceiving masked speech. *J. Acoust. Soc. Am.* 137, 378–387. doi: 10.1121/1.4904537

Bürkner, P.-C. (2017). Brms: an R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* 80, 1–28. doi: 10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *R Journal* 10, 395–411. doi: 10.32614/RJ-2018-017

Bürkner, P.-C. (2020). Bayesian Item Response Modeling in R with brms and Stan. *arXiv* [Epub ahead of print]

Burnham, D., Brooker, R., and Reid, A. (2015). The effects of absolute pitch ability and musical training on lexical tone perception. *Psychol. Music* 43, 881–897. doi: 10.1177/0305735614546359

Chandrasekaran, B., Krishnan, A., and Gandour, J. T. (2009). Relative influence of musical and linguistic experience on early cortical processing of pitch contours. *Brain Lang.* 108, 1–9. doi: 10.1016/j.bandl.2008.02.001

Chang, W., Cheng, J., Allaire, J. J., Xi, Y., and McPherson, J. (2020). Shiny: web application framework for R. *Tech. Innov. Stat. Educ.* 1:7492. doi: 10.5070/T591027492

Chen, S., Zhu, Y., Wayland, R., and Yang, Y. (2020). How musical experience affects tone perception efficiency by musicians of tonal and non-tonal speakers? *PLoS One* 15:e0232514. doi: 10.1371/journal.pone.0232514

Chobert, J., Francois, C., Velay, J. L., and Besson, M. (2014). Twelve months of active musical training in 8- to 10-year-old children enhances the preattentive processing of syllabic duration and voice onset time. *Cereb. Cortex* 24, 956–967. doi: 10.1093/cercor/bhs377

Crowder, R. G. (1971). The sound of vowels and consonants in immediate memory. *J. Verb. Learni. Behav.* 10, 587–596. doi: 10.1016/S0022-5371(71)80063-4

Crowder, R. G. (1973a). Representation of speech sounds in precategorical acoustic storage. *J. Exp. Psychol.* 98, 14–24. doi: 10.1037/h0034286

Crowder, R. G. (1973b). Precategorical acoustic storage for vowels of short and long duration. *Percept. Psychophys.* 13, 502–506. doi: 10.3758/BF03205809

Crowder, R. G., and Morton, J. (1969). Precategorical acoustic storage (PAS). *Percept. Psychophys.* 5, 365–373. doi: 10.3758/BF03210660

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York, NY: The Guilford Press.

Degé, F., and Schwarzer, G. (2011). The effect of a music program on phonological awareness in preschoolers. *Front. Psychol.* 2:124. doi: 10.3389/fpsyg.2011.00124

Dittinger, E., Barbaroux, M., D'Imperio, M., Jäncke, L., Elmer, S., and Besson, M. (2016). Professional music training and novel word learning: from faster semantic encoding to longer-lasting word representations. *J. Cogn. Neurosci.* 28, 1584–1602. doi: 10.1162/jocn_a_00997

Dittinger, E., Chobert, J., Ziegler, J. C., and Besson, M. (2017). Fast brain plasticity during word learning in musically-trained children. *Front. Hum. Neurosci.* 11:233. doi: 10.3389/fnum.2017.00233

Dittinger, E., Scherer, J., Jäncke, L., Besson, M., and Elmer, S. (2019). Testing the influence of musical expertise on novel word learning across the lifespan using a cross-sectional approach in children, young adults and older adults. *Brain Lang.* 198:104678. doi: 10.1016/j.bandl.2019.104678

Dowling, W. J. (1978). Scale and contour: two components of a theory of memory for melodies. *Psychol. Rev.* 85, 341–354. doi: 10.1037/0033-295X.85.4.341

Ellis, N. C. (2016). Salience, cognition, language complexity, and complex adaptive systems. *Stud. Second. Lang. Acquis.* 38, 341–351. doi: 10.1017/S027226311600005X

Elmer, S., Klein, C., Kühnis, J., Liem, F., Meyer, M., and Jäncke, L. (2014). Music and language expertise influence the categorization in musically trained and untrained subjects. *Cereb. Cortex* 22, 650–658. doi: 10.1093/cercor/bhr142

Elmer, S., Meyer, M., and Jäncke, L. (2012). Neurofunctional and behavioral correlates of phonetic and temporal categorization in musically trained and untrained subjects. *Cereb. Cortex* 22, 650–658. doi: 10.1093/cercor/bhr142

Elvin, J., and Escudero, P. (2014). "Perception of Brazilian Portuguese Vowels by Australian English and Spanish Listeners," in *Proceedings of the International Symposium on the Acquisition of Second Language Speech Concordia Working Papers in Applied Linguistics*. 5, 145–156.

Elvin, J., and Escudero, P. (2019). "Cross-linguistic influence in second language speech: implications for learning and teaching," in *Cross-Linguistic Influence: From Empirical Evidence to Classroom Practice*. eds. M. Juncal Gutierrez-Mangado, M. Martínez-Adrián and F. Gallardo-del-Puerto (Cham: Springer), 1–20.

Elvin, J., Escudero, P., and Vasiliev, P. (2014). Spanish is better than English for discriminating Portuguese vowels: acoustic similarity versus vowel inventory. *Front. Psychol.* 5:1188. doi: 10.3389/fpsyg.2014.01188

Elvin, J., Williams, D., and Escudero, P. (2020). "Learning to perceive, produce and recognise words in a non-native language," in *Linguistic Approaches to Portuguese as an Additional Language*. eds. K. V. Molsing, C. B. L. Perna and A. M. T. Ibaños (Amsterdam: John Benjamins Publishing Company).

Elvin, J., Williams, D., Shaw, J. A., Best, C. T., and Escudero, P. (2021). The role of acoustic similarity and non-native categorisation in predicting non-native discrimination: Brazilian Portuguese vowels by English vs. Spanish listeners. *Languages* 6:44. doi: 10.3390/languages6010044

Escudero, P., Benders, T., and Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. *J. Acoust. Soc. Am.* 130, EL206–EL212. doi: 10.1121/1.3629144

Escudero, P., and Boersma, P. (2002). "The subset problem in L2 perceptual development: multiple- category assimilation by Dutch learners of Spanish." in *Proceedings of the 26th Annual Boston University Conference on Language Development*. eds. B. Skarabela, S. Fish, and A. H.–J. Do. November 2–4, 2001. Somerville, MA: Cascadilla Press, 208–219.

Escudero, P., and Hayes-Harb, R. (2021). The ontogenesis model may provide a useful guiding framework, but lacks explanatory power for the nature and development of L2 lexical representation. *Biling. Lang. Congn.* 1–2. doi: 10.1017/S1366728921000602

Escudero, P., Mulak, K. E., Fu, C. S., and Singh, L. (2016a). More limitations to monolingualism: bilinguals outperform monolinguals in implicit word learning. *Front. Psychol.* 7:1218. doi: 10.3389/fpsyg.2016.01218

Escudero, P., Mulak, K. E., and Vlach, H. A. (2016b). Cross-situational word learning of minimal word pairs. *Cogn. Sci.* 40, 455–465. doi: 10.1111/cogs.12243

Escudero, P., Mulak, K. E., and Vlach, H. A. (2016c). Infants encode phonetic detail during cross-situational word learning. *Front. Psychol.* 7:1419. doi: 10.3389/fpsyg.2016.01419

Escudero, P., Smit, E. A., and Angwin, A. (2021). Investigating orthographic versus auditory cross-situational word learning with online and lab-based research. *PsyArXive*. doi: 10.31234/osf.io/tpn5e [Epub ahead of print]

Escudero, P. (2005). *Linguistic Perception and Second Language Acquisition: Explaining the Attainment of Optimal Phonological Categorization*. LOT Dissertation Series 113, Utrecht University.

Escudero, P., and Williams, D. (2014). Distributional learning has immediate and long-lasting effects. *Cognition* 133, 408–413. doi: 10.1016/j.cognition.2014.07.002

Flege, J. E. (1995). Second Language Speech Learning Theory, Findings, and Problems.

François, C., Chobert, J., Besson, M., and Schön, D. (2013). Music training for the development of speech segmentation. *Cereb. Cortex* 23, 2038–2043. doi: 10.1093/cercor/bhs180

Gathercole, S. E., Hitch, G. J., and Marin, A. J. (1997). Phonological short-term memory and new word learning in children. *Dev. Psychol.* 33, 966–979. doi: 10.1037/0012-1649.33.6.966

Golinkoff, R. M., and Alioto, A. (1995). Infant-directed speech facilitates lexical learning in adults hearing Chinese: implications for language acquisition. *J. Child Lang.* 22, 703–726. doi: 10.1017/S0305000900010011

Gordon, R. L., Shivers, C. M., Wieland, E. A., Kotz, S. A., Yoder, P. J., and McAuley, J. D. (2015). Musical rhythm discrimination explains individual differences in grammar skills in children. *Dev. Sci.* 18, 635–644. doi: 10.1111/desc.12230

Graf Estes, K., and Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy* 18, 797–824. doi: 10.1111/infa.12006

Hallam, S. (2017). The impact of making music on aural perception and language skills: a research synthesis. *Lond. Rev. Educ.* 15, 388–406. doi: 10.18546/LRE.15.3.05

Han, M., de Jong, N. H., and Kager, R. (2018). Lexical tones in mandarin Chinese infant-directed speech: age-related changes in the second year of life. *Front. Psychol.* 9:434. doi: 10.3389/fpsyg.2018.00434

Hansen, M., Wallentin, M., and Vuust, P. (2012). Working memory and musical competence of musicians and nonmusicians. *Psychol. Music* 41, 779–793. doi: 10.1177/0305735612452186

Harrison, P. M. C., Collins, T., and Müllensiefen, D. (2017). Applying modern psychometric techniques to melodic discrimination testing: item response theory, computerised adaptive testing, and automatic item generation. *Sci. Rep.* 7:3618. doi: 10.1038/s41598-017-03586-z

Harrison, P. M. C., and Müllensiefen, D. (2018). Melodic discrimination test (MDT), psychTestR implementation. *Zenodo*. doi: 10.5281/zenodo.1300950

Houston-Price, C., and Law, B. (2013). "How experiences with words supply all the tools in the toddler's word – learning toolbox," in *Theoretical and Computational Models of Word Learning: Trends in Psychology and Artificial Intelligence*. eds. L. Gogate and G. Hollich (Hershey, PA: IGI Global), 81–108.

Jeffreys, H. (1998). *The Theory of Probability*. England: OUP Oxford.

Jentschke, S., and Koelsch, S. (2009). Musical training modulates the development of syntax processing in children. *NeuroImage* 47, 735–744. doi: 10.1016/j.neuroimage.2009.04.090

Jonaitis, E. M., and Saffran, J. R. (2009). Learning harmony: the role of serial statistics. *Cogn. Sci.* 33, 951–968. doi: 10.1111/j.1551-6709.2009.01036.x

Kachergis, G., Yu, C., and Shiffrin, R. M. (2010). "Cross-situational statistical learning: implicit or intentional?" in *Proceedings of the Annual Meeting of the Cognitive Science Society*. *Vol 32*. August 11–14, 2010.

Kraus, N., and Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nat. Rev. Neurosci.* 11, 599–605. doi: 10.1038/nrn2882

Kraus, N., and White-Schwoch, T. (2017). Neurobiology of everyday communication: what have we learned from music? *Neuroscientist* 23, 287–298. doi: 10.1177/1073858416653593

Krishnan, A., Xu, Y., Gandour, J., and Cariani, P. (2005). Encoding of pitch in the human brainstem is sensitive to language experience. *Cogn. Brain Res.* 25, 161–168. doi: 10.1016/j.cogbrainres.2005.05.004

Krizman, J., Marian, V., Shook, A., Skoe, E., and Kraus, N. (2012). Subcortical encoding of sound in enhanced in bilinguals and relates to executive function advantages. *Proc. Natl. Acad. Sci.* 109, 7877–7881. doi: 10.1073/pnas.1201575109

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Adv. Methods Pract. Psychol. Sci.* 1, 270–280. doi: 10.1177/2515245918771304

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., et al. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science* 277, 684–686. doi: 10.1126/science.277.5326.684

Kühnis, J., Elmer, S., Meyer, M., and Jäncke, L. (2013). The encoding of vowels and temporal speech cues in the auditory cortex of professional musicians: an EEG study. *Neuropsychologia* 51, 1608–1618. doi: 10.1016/j.neuropsychologia.2013.04.007

Kunert, R., Willems, R. M., and Hagoort, P. (2016). An independent psychometric evaluation of the PROMS measure of music perception skills. *PLoS One* 11:e0159103. doi: 10.1371/journal.pone.0159103

Lamb, S. J., and Gregory, A. H. (1993). The relationship between music and reading in beginning readers. *J. Educ. Psychol.* 13, 19–27. doi: 10.1080/0144341930130103

Larrouy-Maestri, P., Harrison, P. M. C., and Müllensiefen, D. (2018). Mistuning perception test, psychTestR implementation. *Zenodo*. doi: 10.5281/zenodo.1415363

Larrouy-Maestri, P., Harrison, P. M. C., and Müllensiefen, D. (2019). The mistuning perception test: A new measurement instrument. *Behav. Res. Methods* 51, 663–675. doi: 10.3578/s13428-019-01225-1

Law, L. N. C., and Zentner, M. (2012). Assessing musical abilities objectively: construction and validation of the profile of music perception skills. *PLoS One* 7:e52508. doi: 10.1371/journal.pone.0052508

Leung, Y., and Dean, R. T. (2018). Learning unfamiliar pitch intervals: a novel paradigm for demonstrating the learning of statistical associations between musical pitches. *PLoS One* 13:e0203026. doi: 10.1371/journal.pone.0203026

Liu, J., Hilton, C. B., Bergelson, E., and Mehr, S. A. (2021). Language experience shapes music processing across 40 tonal, pitch-accented, and non-tonal languages. *bioRxiv*. doi: 10.1101/2021.10.18.464888 [Epub ahead of print]

Loui, P., Wessel, D. L., and Hudson Kam, C. L. (2010). Human rapidly learn grammatical structure in a new musical scale. *Music. Percept.* 27, 377–388. doi: 10.1525/mp.2010.27.5.377

Ma, W., Golinkoff, R. M., Houston, D. M., and Hirsh-Pasek, K. (2011). Word learning in infant- and adult-directed speech. *Lang. Learn. Dev.* 7, 185–201. doi: 10.1080/15475441.2011.579839

Magis, D., and Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *J. Stat. Softw.* 48:i08. doi: 10.18637/jss.v048.i08

Magne, C., Schön, D., and Besson, M. (2006). Musician children detect pitch violations in both music and language better than nonmusician children: behavioral and electrophysiological approaches. *J. Cogn. Neurosci.* 18, 199–211. doi: 10.1162/jocn.2006.18.2.199

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., and Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Front. Psychol.* 10:2767. doi: 10.3389/fpsyg.2019.02767

Marie, C., Delogu, F., Lampis, G., Belardinelli, M. O., and Besson, M. (2011a). Influence of musical expertise on segmental and tonal processing in mandarin Chinese. *J. Cogn. Neurosci.* 23, 2701–2715. doi: 10.1162/jocn.2010.21585

Marie, C., Magne, C., and Besson, M. (2011b). Musicians and the metric structure of words. *J. Cogn. Neurosci.* 23, 294–305. doi: 10.1162/jocn.2010.21413

Milne, A. J., and Herff, S. A. (2020). The perceptual relevance of balance, evenness, and entropy in musical rhythms. *Cognition* 203:104233. doi: 10.1016/j.cognition.2020.104233

Moreno, S., Marques, C., Santos, A., Santos, M., Castro, S. L., and Besson, M. (2009). Musical training influences linguistic abilities in 8-year-old children: more evidence for brain plasticity. *Cereb. Cortex* 19, 712–723. doi: 10.1093/cercor/bhn120

Mulak, K. E., Vlach, H. A., and Escudero, P. (2019). Cross-situational learning of phonologically overlapping words across degrees of ambiguity. *Cogn. Sci.* 43:e12731. doi: 10.1111/cogs.12731

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:e101091. doi: 10.1371/journal.pone.0089642

Musacchia, G., Sams, M., Skoe, E., and Kraus, N. (2007). Musicians have enhanced subcortical auditory and audiovisual processing of speech and music. *Proc. Natl. Acad. Sci. U. S. A.* 104, 15894–15898. doi: 10.1073/pnas.0701498104

Ong, J. H., Burnham, D., and Escudero, P. (2015). Distributional learning of lexical tones: a comparison of attended vs. unattended listening. *PLoS One* 10:e0133446. doi: 10.1371/journal.pone.0133446

Ong, J. H., Burnham, D., Escudero, P., and Stevens, C. J. (2017b). Effect of linguistic and musical experience on distributional learning of nonnative lexical tones. *J. Speech Lang. Hear. Res.* 60, 2769–2780. doi: 10.1044/2016_JSLHR-S-16-0080

Ong, J. H., Burnham, D., and Stevens, C. J. (2017a). Learning novel musical pitch via distributional learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 43, 150–157. doi: 10.1037/xlm0000286

Ong, J. H., Burnham, D., Stevens, C. J., and Escudero, P. (2016). Naïve learners show cross-domain transfer after distributional learning: the case of lexical and musical pitch. *Front. Psychol.* 7:1189. doi: 10.3389/fpsyg.2016.01189

Ong, J. H., Wong, P. C. M., and Liu, F. (2020). Musicians show enhanced perception, but not production, of native lexical tones. *J. Acoust. Soc. Am.* 148, 3443–3454. doi: 10.1121/10.0002776

Ott, C. G. M., Lnager, N., Oeschlin, M. S., Meyer, M., and Jäncke, L. (2011). Processing of voiced and unvoiced acoustic stimuli in musicians. *Front. Psychol.* 2:195. doi: 10.3389/fpsyg.2011.00195

Overy, K. (2003). Dyslexia and music: from timing deficits to musical intervention. *Ann. N. Y. Acad. Sci.* 999, 497–505. doi: 10.1196/annals.1284.060

Parbery-Clark, A., Strait, D. L., Anderson, S., Hittner, E., and Kraus, N. (2011). Musical experience and the aging auditory system: implications for cognitive abilities and hearing speech in noise. *PLoS One* 6:e18082. doi: 10.1371/journal.pone.0018082

Patel, A. D. (2003). Language, music, syntax and the brain. *Nat. Neurosci.* 6, 674–681. doi: 10.1038/nn1082

Patel, A. D., and Iversen, J. R. (2007). The linguistic benefits of musical abilities. *Trends Cogn. Sci.* 11, 369–372. doi: 10.1016/j.tics.2007.08.003

Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., and Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage* 50, 302–313. doi: 10.1016/j.neuroimage.2009.12.019

Peirce, J. W. (2007). PsychoPy-psychophysics software in python. *J. Neurosci. Methods* 162, 8–13. doi: 10.1016/j.jneumeth.2006.11.017

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., et al. (2019). PsychoPy2: experiments in behavior made easy. *Behav. Res. Methods* 51, 195–203. doi: 10.3758/s13428-01801193-y

Perfors, A., and Ong, J. H. (2012). "Musicians Are Better at Learning Non-native Sound Contrasts Even in Non-tonal Languages," in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. eds. N. Miyake, D. Peebles and R. P. Cooper. August 1-4, 2012. Cognitive Science Society, 839–844.

Pinheiro, A. P., Vasconcelos, M., Dias, M., Arrais, N., and Gonçalves, Ó. F. (2015). The music of language: An ERP investigation of the effects of musical training on emotional prosody processing. *Brain Lang.* 140, 24–34. doi: 10.1016/j.bandl.2014.10.009

Pinker, S. (1997). *How the Mind Works*. New York: Norton.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Percept. Psychophys.* 13, 253–260. doi: 10.3758/BF03214136

Pisoni, D. B. (1975). Auditory short-term memory and vowel perception. *Mem. Cogn.* 3, 7–18. doi: 10.3758/BF03198202

R Core Team (2020). R: A Language and Environment for Statistical Computing [Computer Software Manual]. Vienna, Austria. Available at: https://www.R-project.org/

Rogalsky, C., Rong, F., Saberi, K., and Hickok, G. (2011). Functional anatomy of language and music perception: temporal and structural factors investigated using functional magnetic resonance imaging. *J. Neurosci.* 31, 3843–3852. doi: 10.1523/JNEUROSCI.4515-10.2011

Ruggles, D. R., Freyman, R. L., and Oxenham, A. J. (2014). Influence of musical training on understanding voiced and whispered speech in noise. *PLoS One* 9:e86980. doi: 10.1371/journal.pone.0086980

Sammler, D., Grigutsch, M., Fritz, T., and Koelsch, S. (2007). Music and emotion: electrophysiological correlates of the processing of pleasant and unpleasant music. *Psychophysiology* 44, 293–304. doi: 10.1111/j.1469-8986.2007.00497.x

Schön, D., Magne, C., and Besson, M. (2004). The music of speech: music training facilitates pitch processing in both music and language. *Psychophysiology* 41, 341–349. doi: 10.1111/1469-8986.00172.x

Schulze, K., Zysset, S., Mueller, K., Friederici, A. D., and Koelsch, S. (2011). Neuroarchitecture of verbal and tonal working memory in nonmusicians and musicians. *Hum. Brain Mapp.* 32, 771–783. doi: 10.1002/hbm.21060

Slevc, L. R., and Miyake, A. (2006). Individual differences in second-language proficiency: does musical ability matter? *Psychol. Sci.* 17, 675–681. doi: 10.1111/j.1467-9280.2006.01765.x

Smit, E. A., Milne, A. J., Dean, R. T., and Weidemann, G. (2019). Perception of affect in unfamiliar musical chords. *PLoS One* 14:e0218570. doi: 10.1371/journal.pone.0218570

Smith, A. D. M., and Smith, K. (2012). "Cross-situational learning," in *Encyclopedia of the Sciences of Learning*. ed. N. M. Seel (Boston, MA: Springer US), 864–866.

Smith, L. B., and Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition* 106, 1558–1568. doi: 10.1016/j.cognition.2007.06.010

Stewart, E. C., and Pittman, A. L. (2021). Learning and retention of novel words in musicians and nonmusicians. *J. Speech Lang. Hear. Res.* 64, 2870–2884. doi: 10.1044/2021_JSLHR-20-00482

Strait, D. L., and Kraus, N. (2011a). Can you hear me now? Musical training shapes functional brain networks for selective auditory attention and hearing speech in noise. *Front. Psychol.* 2:113. doi: 10.3389/fpsyg.2011.00113

Strait, D. L., and Kraus, N. (2011b). Playing music for a smarter ear: cognitive, perceptual and neurobiological evidence. *Music. Percept.* 29, 133–146. doi: 10.1525/mp.2011.29.2.133

Strange, W. (ed.) (1999). "Second language speech learning theory, findings, and problems," in *Speech perception and linguistic experience: issues in cross-language research*. Timonium, MD: York Press, 229-273.

Suanda, S. H., Mugwanya, N., and Namy, L. L. (2014). Cross-situational statistical word learning in young children. *J. Exp. Child Psychol.* 126, 395–411. doi: 10.1016/j.jecp.2014.06.003

Swaminathan, S., and Gopinath, J. K. (2013). Music training and second-language English comprehension and vocabulary skills in Indian children. *Psychol. Stud.* 58, 164–170. doi: 10.1007/s12646-013-0180-3

Swaminathan, S., and Schellenberg, E. G. (2017). Musical competence and phoneme perception in a foreign language. *Psychon. Bull. Rev.* 24, 1929–1934. doi: 10.3758/s13423-017-1244-5

Talamini, F., Altoè, G., Carretti, B., and Grassi, M. (2017). Musicians have better memory than nonmusicians: a meta-analysis. *PLoS One* 12:e0186773. doi: 10.1371/journal.pone.0186773

Talamini, F., Grassi, M., Toffalini, E., Santoni, R., and Carretti, B. (2018). Learning a second language: can music aptitude or music training have a role? *Learn. Individ. Differ.* 64, 1–7. doi: 10.1016/j.lindif.2018.04.003

Tallal, P., and Gaab, N. (2006). Dynamic auditory processing, musical experience and language development. *Trends Neurosci.* 29, 382–390. doi: 10.1016/j.tins.2006.06.003

Tervaniemi, M. (2001). "Musical sound processing in the human brain: evidence from electric and magnetic recordings," in *The Biological Foundations of Music. Vol. 930*. eds. R. J. Zatorre and I. Peretz (New York, NY: New York Academy of Sciences), 259–272.

Tervaniemi, M., Kujala, A., Alho, K., Virtanen, J., Ilmoniemi, R. J., and Näätänen, R. (1999). Functional specialization of the human auditory cortex in processing phonetic and musical sounds: A magnetoencephalographic (MEG) study. *NeuroImage* 9, 330–336. doi: 10.1006/nimg.1999.0405

Thomson, J. M., Leong, V., and Goswami, U. (2013). Auditory processing interventions and developmental dyslexia: A comparison of phonemic and rhythmic approaches. *Read. Writ.* 26, 139–161. doi: 10.1007/s11145-012-9359-6

Tuninetti, A., Mulak, K., and Escudero, P. (2020). Cross-situational word learning in two foreign languages: effects of native and perceptual difficulty. *Front. Commun.* 5:602471. doi: 10.3389/fcomm.2020.602471

van Leussen, J.-W., and Escudero, P. (2015). Learning to perceive and recognize a second language: The L2LP model revised. *Front. Psychol.* 6:1000. doi: 10.3389/fpsyg.2015.01000

Varnet, L., Wang, T., Peter, C., Meunier, F., and Hoen, M. (2015). How musical expertise shapes speech perception: evidence from auditory classification images. *Sci. Rep.* 5:14489. doi: 10.1038/srep14489

Vlach, H. A., and Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition* 127, 375–382. doi: 10.1016/j.cognition.2013.02.015

Vlach, H. A., and Sandhofer, C. M. (2014). Retrieval dynamics and retention in cross-situational statistical word learning. *Cogn. Sci.* 38, 757–774. doi: 10.1111/cogs.12092

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika* 54, 427–450. doi: 10.1007/BF02294627

Warmington, M. A., Kandru-Pothineni, S., and Hitch, G. J. (2019). Novel-word learning, executive control and working memory: a bilingual advantage. *Bilingualism* 22, 763–782. doi: 10.1017/S136672891800041X

White-Schwoch, T., Carr, K. W., Anderson, S., Strait, D. L., and Kraus, N. (2013). Older adults benefit from music training early in life: Biological evidence for long-term training-driven plasticity. *J. Neurosci.* 33, 17667–17674. doi: 10.1523/JNEUROSCI.2560-13.2013

Wong, P. C. M., and Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Appl. Psycholinguist.* 28, 565–585. doi: 10.1017/S0142716407070312

Yang, H., Ma, W., Gong, D., Hu, J., and Yao, D. (2014). A longitudinal study on children's music training experience and academic development. *Sci. Rep.* 4:5854. doi: 10.1038/srep05854

Yazawa, K., Whang, J., Kondo, M., and Escudero, P. (2020). Language-dependent cue weighting: An investigation of perception modes in L2 learning. *Second. Lang. Res.* 36, 557–581. doi: 10.1177/0267658319832645

Yu, C., and Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychol. Sci.* 18, 414–420. doi: 10.1111/j.1467-9280.2007.01915.x

Zhu, J., Chen, X., and Yang, Y. (2021). Effects of amateur musical experience on categorical perception of lexical tones by native Chinese adults: an ERP study. *Front. Psychol.* 12:611189. doi: 10.3389/fpsyg.2021.611189

Zuk, J., Ozernov-Palchik, O., Kim, H., Lakshminarayanan, K., Gabrieli, J. D. E., Tallal, P., et al. (2013). Enhanced syllable discrimination thresholds in musicians. *PLoS One* 8:e80546. doi: 10.1371/journal.pone.0080546

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

HYPOTHESIS AND THEORY
published: 22 April 2022
doi: 10.3389/fpsyg.2022.786899

Understanding Design Features of Music and Language: The Choric/Dialogic Distinction

Felix Haiduk[1]* *and W. Tecumseh Fitch*[1,2]*

[1]*Department of Behavioral and Cognitive Biology, University of Vienna, Vienna, Austria,* [2]*Vienna Cognitive Science Hub, University of Vienna, Vienna, Austria*

Music and spoken language share certain characteristics: both consist of sequences of acoustic elements that are combinatorically combined, and these elements partition the same continuous acoustic dimensions (frequency, formant space and duration). However, the resulting categories differ sharply: scale tones and note durations of small integer ratios appear in music, while speech uses phonemes, lexical tone, and non-isochronous durations. Why did music and language diverge into the two systems we have today, differing in these specific features? We propose a framework based on information theory and a reverse-engineering perspective, suggesting that design features of music and language are a response to their differential deployment along three different continuous dimensions. These include the familiar propositional-aesthetic ('goal') and repetitive-novel ('novelty') dimensions, and a dialogic-choric ('interactivity') dimension that is our focus here. Specifically, we hypothesize that music exhibits specializations enhancing coherent production by several individuals concurrently—the 'choric' context. In contrast, language is specialized for exchange in tightly coordinated turn-taking—'dialogic' contexts. We examine the evidence for our framework, both from humans and non-human animals, and conclude that many proposed design features of music and language follow naturally from their use in distinct dialogic and choric communicative contexts. Furthermore, the hybrid nature of intermediate systems like poetry, chant, or solo lament follows from their deployment in the less typical interactive context.

Keywords: language, music, information theory, choric, dialogic, animal communication

Edited by:
William Shiyuan Wang,
The Hong Kong Polytechnic University, Hong Kong SAR, China

Reviewed by:
Reyna L. Gordon,
Vanderbilt University, United States
Peter Cariani,
Boston University, United States

***Correspondence:**
Felix Haiduk
felix.haiduk@univie.ac.at
W. Tecumseh Fitch
tecumseh.fitch@univie.ac.at

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

Received: *30 September 2021*
Accepted: *22 February 2022*
Published: *22 April 2022*

Citation:
Haiduk F and Fitch WT (2022) Understanding Design Features of Music and Language: The Choric/Dialogic Distinction. Front. Psychol. 13:786899. doi: 10.3389/fpsyg.2022.786899

INTRODUCTION

Music and language are two human cognitive and communicative systems that are similar in a variety of ways: the vocal-auditory domain is typically the primary modality, but it is not the only one (writing, sign, or dance are others). Both utilise the same vocal apparatus, and similar motor systems and perceptual physiology. Their respective neural underpinnings have major shared portions. Both consist of elements combined in a hierarchical manner by certain, culture-specific rules. Both systems are learned, but have biological components shared with other species. Despite these many similarities, this paper is concerned with the differences between the two systems. Why should two universal human systems, that share so much, nonetheless exhibit consistent differences?

It is clear that there is a great variety of music and language within and across cultures, and what is termed 'music' varies within a culture (see Trehub et al., 2015; Thompson et al., 2019), fulfilling a broad range of psychological purposes that influence their acoustic features. For example, while dance music will engage motor systems, lullabies are used for soothing infants, and this translates into consistent acoustic differences cross-culturally (Mehr et al., 2018). Similarly, language changes when playing with young infants, reciting a mantra in a ritual, or engaging in political discussions. However, despite this variety, certain features seem to differentiate many instances of music and language (which we will term 'typical' in this paper). Hockett (1960) and Fitch (2006) termed these prototypical properties 'design features' of language and music, respectively (see **Table 1**). Nonetheless, borders between language and music are not clear cut (as in the case of poetry or religious chanting), and particular instantiations of music and language can be 'more musical' or 'more linguistic' than prototypical instances.

In this paper, we propose a framework that aims to explain design features differentiating music and language as responses along three continuous dimensions. (1) the **goal** of the linguistic or musical act, with a more propositional or more aesthetic focus; (2) the repetitiveness or **novelty** of the events within a linguistic or musical sequence and (3) crucially, the **interaction** and temporal coordination between individuals participating in linguistic or musical acts, the poles of which we term 'choric' and 'dialogic'. While the first two dimensions are widely recognised and discussed in comparisons of music and language, the last one has more often been neglected. We think that the interaction dimension is a crucial addition for understanding design feature differences between language and music, because

TABLE 1 | Design features differing between language and music, updated from Fitch (2006).

Design Feature	Language		Music		Definition
	V	S	V	I	
Vocal auditory channel	+	–	+	–	Signal sequences are patterns of sounds produced by the vocal tract and articulators
Broadcast transmission	+	+?	+	+	Signal sequences are detectable by anyone within given distance/line of sight
Rapid fading	+	+	+	+	Signal sequences dissipate when signalling stops
Interchangeability	+	+	+	–	Individuals can be both sender and receiver
Total feedback	+	+[1]	+	+?[1]	Senders themselves perceive what they signal
Specialisation	+	+	+	+	A signal sequence does not directly trigger a specific behaviour in the receiver
Productivity	+	+	+	+	Ability to produce novel signal sequences
Discreteness	+	+	+	+	Signalling units are functionally distinct
Cultural transmission	+	+	+	+	The signalling system is transmitted between individuals *via* learning and teaching
Movement[2]	+	+	+	+	Movements of body (–parts) accompany movements that create the signal itself
Transposability	+	+	+	+	The relationships between signal units rather than absolute features identify a signal sequence (a sentence is considered the same regardless of who spoke/signed it, a melody regardless of instrument, voice or absolute pitch)
Duality of Patterning	+	+	–	–	Signal sequences can be analysed both as units of signalling (cenemes) and meaning-bearing units (pleremes)
Generativity	+	+	+	+	Signal units are recombined according to rules
Semanticity	+	+	–	–	Fixed associations exist between meaning-bearing units and states or properties of the world/environment
Arbitrariness	+	+	–	–	The content of most meaning-bearing units is unrelated to features of signalling units
Displacement	+	+	–	–	Meaning-bearing units refer to entities outside their spatial and temporal context
Discrete pitches	–	–	+	+	Allowed pitches are based on a scale of tones related by intervals
Isochronic	–	–	+	+	Regular periodic pulse providing a reference framework for other temporal features of the signal sequence
Performative context	–	–	+	+	Classes of signal sequences (e.g. songs or styles) recur in specific social contexts
Repeatable (repertoire)	–	–	+	+	Signal sequences are distinguishable (pieces), exactly repeatable and repeated in certain contexts
A-referentially expressive	–	+?	+	+	Higher order relations of a signal sequence are cognitively mapped to movement and affective responses

These design features concern speech (including sign) or musical acts that we label as 'typical', e.g., spoken conversations or musical ensemble playing. V=vocal, S=signed, I=instrumental.

[1] *Sensorimotor.*

[2] *Added by the authors.*

major acoustic differences between spoken language and music are rooted in social cognition and interaction.

We derive these dimensions by applying a 'reverse-engineering' approach, based on information theory, starting from the observed design features. This framework supports predictions about the changes in design features expected for 'nontypical' instances of music and language, thus laying the foundations for a more fine-grained and continuous analysis of music and language when used for different psychological and social purposes.

Our comparison of music and language focusses on social interactions and starts from the auditory domain, based on the premise that written communication is a derived form. However, both systems go beyond the purely acoustic domain (Cross et al., 2013; Levinson and Holler, 2014; Honing, 2018). For example, both music and language incorporate body movement in the form of dance and co-speech gestures, or mime and sign languages (which are typically silent). Although our framework takes the auditory domain as a starting point, we expect that it can also be applied more generally to movement-based communication, predicting changes in movement-based communication and incorporating movement into speech or song. We thus think our framework might also be useful for analysing animal communication, both acoustic and multimodal.

Our framework is in principle compatible with various hypotheses about the evolutionary relation of language and music (see Cross et al., 2013). We assume only that variation in acoustics occurs based on social and perceptual goals, pointing at fundamental relevant traits, but remain agnostic with regards to the evolutionary processes involved (biological and/or cultural) and/or the origin states of language and music (e.g., from a common audio-vocal precursor system, as Darwin, 1871 proposed). However, our framework does assume a pivotal role for audio-vocal communication at some point in evolution, thus incorporating the phylogenetically unusual trait of vocal learning (Fitch, 2006; Jarvis, 2019), which is shared by both systems. Crucially, our framework avoids dichotomous conceptions of music and language as either fully distinct or fully indissociable faculties. This notion of the differences along a continuum follows naturally both from neural evidence and from the existence of styles intermediate between music and language (poetry, rap, lament and others).

The paper is structured as follows: the first section presents the conceptual and theoretical foundations for our framework: a reverse engineering approach allows us to derive three dimensions from design features differing between language and music. The three dimensions described—goal, novelty, and interactivity—create a space within which both prototypical and non-canonical forms of both music and language can be situated. Information theory makes these design features predictable. The three sections that follow discuss the three dimensions in more detail, arguing that the characteristic design features of music and language can be understood as a function of their deployment within this three-dimensional space. The last section opens the door to comparative cognition, arguing that some vocal communication in non-human animals can also be fruitfully understood using our framework, and ends with predictions and suggestions for questions to be addressed in future empirical research.

CONCEPTUAL AND THEORETICAL FRAMEWORK

Design Stance and Reverse Engineering

In addition to investigations of neural and cognitive processes, individual development and cultural specifics, a deeper understanding of both language and music and how they differ requires inquiry into their evolutionary origin(s). Various hypotheses have been proposed regarding the origin of music, often concerned with finding an adaptive value (see Mehr et al., 2021; Savage et al., 2021 for recent reviews of the debate). We will not focus on possible adaptive values in this paper, nor will we investigate the causal roles of the many possible evolutionary, cultural or developmental processes involved. Rather, we will take a design stance and a 'reverse engineering' approach, using the design features proposed by Hockett (1960) and Fitch (2006; see **Table 1**) as a starting point for our framework.

The 'design' stance has a long tradition in biology and relies on the idea that under certain constraints evolutionary processes act to refine and optimize traits as would an engineer (Hockett, 1960; Krebs and Davies, 1997; Maynard Smith, 2000; Csete and Doyle, 2002; Richardson, 2003; Tooby and Cosmides, 2005); the use of 'design' in this context implies natural selection and has no association with unscientific notions of 'intelligent design'. This allows us to ask what constraints on concrete linguistic or musical acts could plausibly yield the observed design features differentiating music and language. We will conceptualise these constraints as poles of continuous dimensions, creating a multidimensional conceptual space. Crucially, this continuous space allows us to predict how design features of non-typical instances of language and music, such as poetry or rap, should vary as a response of their deployment along the dimensions proposed.

First, note that the kind of elements that make up language and music differ. Language consists of phonemes that are the building blocks for meaning-bearing units like morphemes and words, which in turn are combined to yield sentences. This organisation rests on the need to convey propositional meaning, which is a key characteristic of prototypical language use, but not of prototypical music. Accordingly, the design feature of semanticity and those derived from it (arbitrariness, displacement and duality of patterning) discriminate prototypical language from prototypical music. Although sung music that uses lyrics is common, there is no requirement to perceive lyrics in order to recognise a sound sequence as music, and much music is purely instrumental. Music instead has stronger links to movement, and to emotional and aesthetic appraisal (Huron, 2006; see also Thompson et al., 2019 for cross-cultural perspectives). Fitch (2006) subsumes the expressive mappings of musical form to movement and emotions under the design feature 'a-referentially expressive'.

These contrasting design feature differences between prototypical language and music suggest a trade-off between a primary goal of conveying semantic meaning for language (which we term 'propositional') and a goal of aesthetic appraisal (in a broad sense, see Huron, 2016) for music. We suggest that many observed design feature differences can be explained by interlocutors following either aesthetic or propositional goals. Notably, both aesthetic and propositional goals require predictive cognitive processes, but in different ways, as reviewed below. But simply categorizing music as aesthetic and language as propositional is also incomplete—some ways of speaking also pursue aesthetic goals, as in poetry, while some music has propositionality, like humming 'Happy Birthday', to indicate gift-giving, or songs mimicking birdsong. It is thus useful to conceive of music and language as lying on a propositional-aesthetic continuum, where language typically tends towards the propositional side while music tends towards the aesthetic side, but with some instances between these poles. We will term this continuous axis the '**goal**' dimension.

A second dimension further partially differentiates language and music. Conversational language typically conveys a large amount of novel semantic information (Grice, 1975) and exact repetition is unusual. Music in contrast is typically characterised by repetition at multiple levels, from single tones or chords, motifs, and melodies, up to repeated performances of entire musical pieces. This is supported by two further contrasting design features: while language has gliding intonation, flexible lexical tone and continuously variable syllable durations, music typically consists of tones of fixed pitches organised in scales, and is prototypically characterised by rigorous timing based on isochronous meter (for exceptions see Savage et al., 2015). Thus, in music, both the temporal and the spectral acoustic dimensions relate their elements by small integer ratios. Repeatability is further related to the design feature of performative context, where certain kinds of music are repeated in specific cultural situations (e.g., lullabies to soothing babies). Repeating the same phrases does occur in language, but mostly in specific cultural situations like religious or artistic acts (e.g., prayers or poems). Typically, however, repetition is uncharacteristic of everyday conversations but abundant in music making (Savage et al., 2015). This repetitive-novel continuum is thus another dimension where music and language have a different focus, although again certain instances occupy the middle ground along the continuum. We will call this the '**novelty**' dimension.

Both the goal and the novelty dimensions are widely known and discussed, and both involve predictive cognitive processes. However, although language and music can be deployed at several points along these dimensions, the predictive cues they provide differ (e.g., music has a much smaller set of possible temporal and frequency constituents than speech). We will argue that these differences make sense only when a third dimension is added, involving the timing of individual performances in a dyad or group. As Brown (2007) has argued, an important difference between music and language is their temporal coordination. Language prototypically exhibits sequential turn-taking, where speakers typically have little overlap in their utterances. In music, simultaneity is both possible and typical: music is often performed by several people simultaneously. We will adapt the term 'concurrent' to refer to individuals simultaneously performing (vocalising, playing), specifically when these signals are coupled (causally related) and coordinated (thus excluding two unconnected conversations at the same party). Concurrence does not necessarily imply the same events happening at precisely the same time (which we term 'synchronous', following Ravignani et al., 2014). We dub the end of this dimension that involves turn-taking and alternation '**dialogic**', and the pole featuring concurrent performance '**choric**' (from the Greek *choros* meaning 'chorus'). We choose these novel terms to specifically imply joint action: deliberate coordination within a common representational framework (see Sebanz and Knoblich, 2009, 2021; for music, e.g., Keller et al., 2014; for language, e.g., Tomasello, 2010). While turn-taking requires cues to predict the end of the current speaker's phrase, concurrence requires much more fine-grained ongoing predictions about subsequent events in a vocal sequence. Again, this **choric/dialogic axis** defines a continuum, and there are intermediate cases of dialogic form in music, for example exchanging solos in jazz or call-and-response songs, and concurrence in language, such as group chanting or recitation. We call this axis the '**interactivity**' dimension.

The purpose of these three dimensions (see **Table 2**) is to conceptualise a continuous space that can account for both prototypical instances of language and music and instances that are not considered typical, and to explain their design features as a consequence of the deployment along these dimensions.

Hockett's design stance as applied to language has been criticised for neglecting cognition, being biased concerning the modality of transmission (auditory-vocal) and focussing on surface aspects of the linguistic code rather than its content (Wacewicz and Żywiczyński, 2015). However, these criticisms are less telling regarding music, and our approach attempts to overcome any such limitations. For example, we start our comparison of language and music assuming an auditory-vocal modality, but emphasize that it can also be applied to signed languages or mime, and incorporate facial expressions, gestures and body language, as long as information trajectories can be measured in the target domain (see **Table 3**). Crucially, cognition plays a central role in our framework, *via* the mutually predictive role of the participants in temporally unfolding musical and linguistic acts, which require complex multi-time scale cognitive processes.

Information Theory

All three of our dimensions centrally involve predictive processes. In language (goal dimension), inferring propositional meanings involves prediction at the level of semantics, while aesthetic experiences exploit the interplay between fulfilment of expectations and deviation from predictions. Repetition entails high predictability, while novelty implies low predictability. Finally, for the interactivity dimension, coordinating events in time between several individuals requires either prediction or reaction, with prediction being faster and more flexible.

TABLE 2 | Overview of the three proposed dimensions of our framework, with examples from music, language, and animal communication.

Dimension	Pole 1		Pole 2	
	Name	**Example**	**Name**	**Example**
Goal	Propositional	Discussing the week's events with a friend Singing 'Happy Birthday'	Aesthetic	Shakespeare sonnets Listening to your favourite Beatles album
Novelty	Novelty	Listening to a conference talk Variation and recombination of melodic modules in BaAka music (Lewis, 2021)	Repetition	Word repetition for emphasis ('I did not break the dish. I did not break the dish. I repeat, I did not break the dish') Choruses in songs
Interactivity	Choric	Religious ensemble chanting Ensemble music Plain-tailed wren mating display (within sex)	Dialogic	Conversational speech Call-and-response song Animal antiphonal calling

TABLE 3 | Assumptions and measures of information theory.

Assumption	Measure/method	References
Information is an adequate model of prediction, plausible to happen in the brain	Predictive coding and similar accounts	Friston, 2010; McDonnell et al., 2011; Pearce and Wiggins, 2012; Crupi et al., 2018; Koelsch et al., 2019
Entropy and information can be measured at multiple levels of the signal sequence concurrently, and their interaction can be modelled	Models based on statistical learning and using a multiple viewpoint approach	Pearce and Wiggins, 2012; Forth et al., 2016; see also Rohrmeier and Koelsch, 2012
The information/entropy trajectories of the different levels can be compared	Mutual information measures for multivariate time series (transfer entropy, partial information decomposition, etc.)	Hlaváčková-Schindler et al., 2007; Williams and Beer, 2010 (preprint); Williams and Beer, 2011 (preprint)
Context (e.g., discourse context, conceptual knowledge, etc.) can be modelled using information theory	Conditional entropy (e.g., with *n*-gram models)	Piantadosi et al., 2012; Mahowald et al., 2013; see also Kuperberg and Jaeger, 2016; see also Venhuizen et al., 2019
Information theory can be applied to both discrete or continuous (or discretisable) sequences, e.g., for body movement and gesturing	Discretisation of continuous signals	Glowinski et al., 2013; Zbili and Rama, 2021
	Sample entropy, multiscale entropy	Glowinski et al., 2010; Glowinski and Mancini, 2011

Prediction involves estimates of probability: at any given point during the musical or linguistic act, possible subsequent events are assigned a probability given the current context, influencing the perceiver's expectations about what happens next. Predictions always have a degree of uncertainty, allowing some possibility of other events to occur instead (even a highly familiar event may be corrupted by noise or mistakes). Thus, in order to support successful prediction, a signaller should decrease the uncertainty about subsequent events. Signals can in this way be analysed in terms of the change in their information over time.

The scientific field dealing with reduction in uncertainty is information theory, originally formulated by Shannon (1948), and we will use information theory as our theoretical foundation when analysing the deployment of language and music along the goal, novelty, and interactivity dimensions. The common currency is information, which is simply reduction in uncertainty, quantified in bits. If an event in a sequence is highly predictable, that event's information content—should it occur—is low. Unexpected events are surprising and have a high information content, hence information content is also termed 'surprisal'. Information theory has developed considerably since Shannon's fundamental insights, and now provides a rich toolbox for analysing a variety of phenomena (see **Table 3**; Crupi et al., 2018). In a crucial addition, the uncertainty of predictions themselves, i.e., the confidence in or precision of one's own predictions (Koelsch et al., 2019), can also be quantified as the expected value of the information, or entropy (see, e.g., Hansen and Pearce, 2014). For concision, **Table 3** lists some of the central assumptions we will adopt, and provides references to the methods and measures used to implement information theory in our framework.

Computational models have been successfully used to manipulate and analyse the information dynamics of sequences (e.g., Hansen et al., 2021). Most such models are probabilistic: they can capture multiple streams of musical features (see **Table 3**), and relying on the Markov assumption (see Rohrmeier and Koelsch, 2012), they predict local dependencies. However, predictions for musical and linguistic sequences can span more than just the next event, especially when syntax or harmonic schemas are considered (Rohrmeier and Koelsch, 2012), indicating the need for hierarchical processing across multiple related time scales (see Zuidema et al., 2018). As long as predictions for events with given probabilities are generated these can in principle be used for measuring information and entropy. Our framework will be discussed based on the prediction of the next, discrete event in a sequence, acknowledging that specific models and measures will need to take long-distance dependencies into account.

With these preliminaries in hand, we now turn to a more detailed consideration of the three axes of our framework, applying them to both prototypical song or speech, but also considering atypical or intermediate cases like poetry.

THE 'GOAL' DIMENSION: PROPOSITIONAL-AESTHETIC DIFFERENCES

The goal dimension concerns the broader purpose of linguistic or musical sequence productions, whether to convey semantic messages, or to elicit and modulate aesthetic responses in a broad sense (including emotional appraisal, pleasure, movement expressiveness, etc., see Huron, 2016). Both poles of this continuum involve predictions at multiple levels, but the poles differ in how the levels interact.

Propositionality in Language

The main goal of linguistic acts is arguably to convey propositional meaning: they enable a comprehender to infer the message the speaker intends to convey (Seifert et al., 2013; Kuperberg and Jaeger, 2016). Although speech acts can often convey social relationship, status, sex, origin, etc., paralinguistically (Ladd, 2014), propositionality is nonetheless at the core of language. From an information theoretical perspective this entails reduction of uncertainty about the propositional content transmitted using the current context.

Applying a framework of reverse-engineering and information theory to language, Mahowald et al. (2020) argue that word length, word frequencies, and sequences of phonemes are all designed to optimise the lexicon in order to efficiently communicate, by optimally balancing complexity and informativity. This holds true over a wide variety of languages, and involves tight interactions between multiple linguistic levels. Using a comprehension model that implements both linguistic experience and world knowledge, Venhuizen et al. (2019) showed that entropy reduction is high in propositional words (reducing uncertainty in meaning), and surprisal (information) decreases towards the end of sentences, when the intended message becomes incrementally clearer. However, linguistic sequences involve multiple levels of representation (semantic, syntactic, phonological, etc.), and prediction takes place at all levels (Levinson and Torreira, 2015; Kuperberg and Jaeger, 2016 for a multimodal perspective see Holler and Levinson, 2019; for a critical review see Huettig and Mani, 2016). These levels have also been shown to interact. The hypothesis of uniform information density of a communicative act suggests a constant information rate per unit time (see, e.g., Aylett and Turk, 2004; Piantadosi et al., 2011), and studies show that speakers can actively manipulate information rate at different levels by altering for example phonetic cues, syntactic cues or word length (Mahowald et al., 2013). Specifically, enhanced prosodic prominence or longer durations are used when syllables cannot be predicted well (that is when entropy is high) based on syntactic, semantic or pragmatic contexts (Aylett and Turk, 2004). Comprehenders also use the current context for disambiguation to infer the conveyed message, and higher predictability given current contextual information yields shorter word lengths (Piantadosi et al., 2012; Gibson et al., 2019). This body of language research shows the direct interaction of information and acoustic features given a propositional goal, but also illustrates how conversational situations can be naturally implemented in an information-theoretic framework.

Thus, it appears that propositionality, specifically prediction and inference of encoded messages, profoundly affects the design of languages. The meaning-bearing level is of primary importance, and variations in predictability at the propositional level are balanced by changes in elements within non-propositional supporting levels, like phonology and word choice. These elements vary to enhance predictability (e.g., from context) or to alter the information rate (e.g., by changing in duration), supporting successful decoding of the propositional message. Thus, part of the attested prosodic variability of speech, e.g., in syllable duration or voice pitch, is an effective response that allows variable rates of information and predictability at lower levels, in support of the propositional goal.

Aesthetics and Reward in Music

Key components of the human reward system relate to prediction (expectancy) and surprise (expectancy violation; Schultz et al., 1997). When an outcome is better than expected, dopamine release is increased, resulting in positive emotions and supporting positive reinforcement learning. Worse than expected outcomes lead to decreased dopaminergic firing, negative emotions and learned avoidance. Dopaminergic firing also predicts the timing of rewarding events (Hollerman and Schultz, 1998). The difference between expected and actual outcomes is termed reward prediction error (Schultz, 2017), and involves predictions about how rewarding a future event will be, as distinguished from sensory predictions about which event will occur (de Fleurian et al., 2019; Koelsch et al., 2019). The extent to which these two predictive contexts—reward prediction error and sensory prediction error—provide appropriate explanatory frameworks for musical pleasure is debated (Colombo and Wright, 2017; Hansen et al., 2017; de Fleurian et al., 2019), but fundamental to either account is the ability to make predictions regarding sequences of sonic events. This ongoing or 'on-line' predictive processing is reflected in many theories of musical meaning based on tension-relaxation dynamics (e.g., Meyer, 1956; Narmour, 1990; Huron, 2006; Lerdahl and Krumhansl, 2007; see also Rohrmeier and Koelsch, 2012). However, we note that not all kinds of music rely on expectancy dynamics in order to fulfil their purposes (e.g., Musique concrète).

What design features allow a sequence of sounds to generate expectations, hence to be predictable, but also allow pleasant surprises and (reward) prediction errors? To generate expectations, there must be stable probabilistic relations between elements of a sound sequence, so the probability of particular events occurring concurrently or adjacent to another sound should be higher than random chance levels. Thus, regularity extraction is the foundation of statistical learning in music (Temperley, 2007), and if these relations span multiple time scales, a hierarchical structure of relations can occur (Rohrmeier and Koelsch, 2012; Rohrmeier et al., 2015).

Learned regularities regarding the temporal and spectral relations between events enable probabilistic expectations about which events are likely to occur when (Temperley, 2007). Musical pleasure has been shown to be highest when either prospective uncertainty is low and retrospective surprise is high, or vice versa (Cheung et al., 2019). Since, in music, no one level of the signal is primary *per se* (no single meaning-bearing level of the signal must be unambiguously inferred by a receiver), elements at different levels (e.g., tone frequencies, durational patterns, motifs, etc.) are not constrained to support any one primary level. Thus, both uncertainty and surprise can vary independently of each other at multiple levels, and fulfilment of predictions and surprise can occur concurrently (Rohrmeier and Koelsch, 2012)—think of a certain melodic motif where the expected last tone occurs at the expected time, but within a different harmonic context. This less constrained design allows music to exploit the human reward system very effectively, supporting predictability at some levels and pleasant surprises at others (Zatorre, 2018).

How then can pleasure be gained from repetitive encounters with the same musical piece? Salimpoor et al. (2011) found that for familiar, liked musical pieces, dopamine is released in the striatum both in response to expectations of peak-pleasure events, and to the peak-pleasure events themselves, but in different striatal subregions. This partly explains why, even under low surprise conditions, pleasure can be gained from musical expectations being fulfilled. Representations of musical features might be sparse and decline over time, such that upon repeated listenings new predictions and prediction errors can be generated (Salimpoor et al., 2015). Furthermore, familiar music may remain rewarding upon repeated hearings if its structure is surprising in relation to other pieces of the same genre, that is when it deviates from schema-like representations (Zatorre, personal communication; Salimpoor et al., 2015). Similarly, liking familiar music can even go as far as disliking variant versions of the same song. Repeated listening to a musical piece can also allow listeners to redirect attention to levels not previously attended to and thus to discover new relations between events, again supporting novelty and surprise even in a highly familiar context (Margulis, 2014). Such attentional shifts allow music to occupy a highly rewarding sweet spot between fulfilling the prediction entirely and a total mismatch (i.e., too much information/surprise, see Zatorre, 2018).

In summary, music prototypically enables fulfilment of aesthetic goals while maintaining predictability by preserving the independence of multiple levels of the sound sequence, allowing concurrent surprise and fulfilment of predictions, as well as independent variation of prospective uncertainty and retrospective surprise. Thus, musical design solutions effectively exploit the basic mammalian dopaminergic reward system (Blood and Zatorre, 2001; Ferreri et al., 2019). Hierarchical relations between sounds in a sequence generate expectations in both music and language, but the aesthetic goal alone does not fully explain why particular design features of music arise. This becomes clearer when looking at atypical examples of language and music.

Aesthetics in Language and Propositionality in Music

Unless lyrics are present (implying a meaning-bearing linguistic layer), music rarely conveys propositional meaning. Exceptions include melodies that themselves stand for messages (e.g., whistling 'Happy Birthday' could convey the message of pleasant birthday wishes), 'songlines' that encode pathways across landscapes, connected to mythological stories (e.g., by Australian native peoples, Chatwin, 1987), or music that imitates natural sounds (e.g., birdsong). Whistled speech or 'drum languages' (cf. Busnel and Classe, 1976) encode propositional meaning in a superficial form, for example using pitch as a replacement of formants or phonemic tone from spoken language.

In such cases, propositional content is woven into the musical structure, and we would expect that exact repeatability plays a crucial role, because surprises would increase the uncertainty of the conveyed meaning. Altering the rhythm of 'Happy Birthday' substantially will make it unrecognisable, and keeping the melodic contour but changing the intervals will make it disconcerting or irritating. Imagine someone playing 'Happy Birthday' to you in a minor key—would you perceive this as sarcastic or ironic? It seems that in cases of propositionality in music, the acceptable variability of the musical structures is reduced, even more than in speech acts, because here the propositional message is encoded in several levels of the whole musical structure (e.g., pitch and rhythm), not primarily at a single semantic level. Such propositional musical pieces are thus more similar to words than sentences. On the other hand, adding a surprising context could make the piece aesthetically more interesting, thus shifting the goal toward the aesthetic pole.

What is predicted when language is deployed in a mainly aesthetic context? Language can also exploit the human reward system *via* generation of expectations, *via* its hierarchical structure of elements. When the goal is propositional, variations in semantic predictability are balanced by changes of elements within non-propositional levels to maintain a roughly uniform information density (see above). Thus, prospective prediction and retrospective information are tightly linked. But with an aesthetic goal, this constraint can be released, with levels of the signal becoming more independent. Enhancing the predictability of content words is no longer necessary, more variability in predictive uncertainty and surprise become possible, and attention can be focussed on other levels of the sequence. For example, in poetry intonation, phonology (rhyme), durations, stress patterns, etc., appear to vary more independently of propositional content. Propositional content is often not straightforward in poetry, and ambiguity and multiple possible interpretations are frequent. Indeed, some poetry in art movements like Dada, such as Kurt Schwitters 'Ur-Sonata' (see Schwitters, 1973), focusses on sound quality rather than propositional content (despite, in historical context, 'conveying the message' of ignoring artistic bourgeois conventions).

Re-reading or re-hearing a poem can also yield new ways of interpretation similar to re-listening to a musical piece (but see Margulis, 2014). Increased independence of hierarchical levels might allow greater embodiment and/or a more musical perception, for example in a Shakespearean sonnet versus rap.

Infant-directed speech is another example of speech moving toward the aesthetic pole (e.g., Thiessen et al., 2005), although distress in young children is reduced more in response to infant-directed song than infant-directed speech (Corbeil et al., 2016), even for unfamiliar songs (Cirelli and Trehub, 2020). This might be related to the discreteness (high predictability) of pitch and especially duration in music. Our conception of flexibility along the propositional-aesthetic dimension could readily be applied to theatre and opera, both of which have to fulfil both propositional and aesthetic goals concurrently. We predict that predictability is traded off such that passages perceived as highly aesthetic are lower in information content, and vice versa.

To sum up, both language and music can be deployed in atypical propositional and aesthetic contexts, and similar responses follow: with more propositional goals, the multiple levels of the speech or musical sequence are more interdependent, and vary their information density to support successful inference of propositional content. For aesthetic goals, independent variation across levels enables more unconstrained variation in uncertainty and surprise, effectively exploiting the human reward system. However, given that music and speech can both be deployed in the nontypical context, aesthetic versus propositional goals alone cannot explain why certain design features characterize most music (e.g., discrete pitches or isochronous meter) but not speech (e.g., gliding intonation and variable syllable durations). This implies that further dimensions are necessary to explain these design differences.

THE NOVELTY AND REPETITION DIMENSION

The novelty-repetition dimension is closely linked to the propositional-aesthetic dimension. This dimension involves the repeatability of elements and their relations at different scales (from single elements to entire pieces) and at multiple levels of musical or linguistic sequences, and their balance in use with novel elements and relations. Generally, repetition enhances predictability, whereas novelty is unpredictable and thus high in information.

Repetition in Music

One of the design features distinguishing prototypical music from language cross-culturally is that music is characterised by repetition at multiple levels (Fitch, 2006; Savage et al., 2015). Repetition can involve single notes, melodic motifs, chord progressions, rhythmic patterns, and the entire musical piece. Repetitiveness in music seems to be also a foundational perceptual principle: the speech-to-song illusion is a striking phenomenon in psychological research on music and language, whereby repetition of speech phrases leads to them being perceived as sung speech (Deutsch et al., 2011). Certain speech phrases, especially when characterised by relatively flat within-syllable pitch contours and less variability in tempo, are more prone to be judged as musical by Western listeners (Tierney et al., 2018). The repetition effect has recently been generalised to repetitions of random tone sequences (Margulis and Simchy-Gross, 2016) and of environmental sounds. These were judged as more musical by Western listeners (Rowland et al., 2019), suggesting that repetition leads to the inference of structural relationships between repeated sounds (cf. Winkler et al., 2009), which are then cognitively interpreted as 'musical'.

What specific features of music allow or select for repeatability? Prior to recording technology, repetition entailed that a sound sequence be remembered and reproduced. To be remembered a sequence must be distinguishable from other, similar sequences (e.g., related melodies or rhythmic patterns), and learnable by establishing relationships between the constituent events. The existence of sound categories and hierarchical rules to combine them (Herff et al., 2021; see also Rohrmeier and Pearce, 2018a,b) enables this. The musical design solutions in this respect are discrete tones in scales (in a hierarchical relation), and durations related in a simple fashion. From an information theoretical perspective, this means that the possible uncertainty about forthcoming musical events is reduced from the outset by adopting a smaller 'alphabet'. This allows a lower number of plausible continuations of a sound sequence than if frequency and temporal dimensions were unconstrained. Because hierarchical relations exist between tones this factor also constrains plausible continuations among distant elements. Reduced alphabet size also supports statistical learning and the application of Gestalt principles, both relevant for prediction in music (Snyder, 2000; Morgan et al., 2019).

Repeatability in music seems to be particularly related to the fact that the temporal dimension in music is also hierarchically structured—durational patterns are related to an underlying meter. First, meter supports embodiment *via* beat extraction and entrainment (Kotz et al., 2018), adding a strong motoric component that may increase the memorability of musical sequences (Brown and Palmer, 2012). Second, meter can also function as a kind of glue between multiple levels of a musical sequence by enforcing relations among them, including higher-order levels like chord progressions, motifs etc. The auditory system is able to make predictions and track deviations at multiple levels at the same time (Vuust et al., 2011). High uncertainty in memory at one level of the musical signal (e.g., in melodic arrangement of pitches) can be countered by low uncertainty in another (e.g., rhythm), reducing the joint uncertainty of both levels and enhancing the confidence in the prediction of the ongoing musical sequence ('I remember that this particular pitch followed with this rhythm').

Is repeatability sufficient to explain the occurrence of discrete pitches on scales and meter in music? Rapid learning of auditory events is even possible for arbitrary sounds that are repeated within a stream of random sounds (Agus et al., 2010), suggesting that the auditory system is capable of finding repetition in the auditory stream irrespective of discreteness. This observation is consistent with our claim that specific design solutions for repeatability in music are not strictly necessary for perception, but relate to (re-)production. However, humans are easily capable of reproducing sound sequences that are not characterised by a reduced alphabet in the frequency and/or temporal domain. This suggests that repeatability is not a sufficient explanation for these design solutions of music. What seems to be crucial, we will argue below, is the interactivity between individuals in a group, when making music together in a choric context.

To summarise, repeatability in musical performances involves a reduction in the alphabet in multiple dimensions. This enables higher predictability and structural relations in a hierarchical manner between elements. In music, meter allows strong temporal predictions, enforcing predictive relations in higher-order levels and enabling a strong link to motoric processing. Scales in melody allow equally strong frequency predictions, since the pitch of possible following notes is strictly circumscribed.

Novelty in Language

As emphasised above, language is mainly concerned with the primary goal of transmitting propositional meaning. These messages conveyed should be relevant and informative, and thus (typically) novel (Grice, 1975; Sperber and Wilson, 1986). The novelty typifying language acts is therefore closely linked to propositionality. What design features enable novelty in language? Crucially, language is characterised by duality of patterning (Hockett, 1960), and can be analysed both as an arrangement of meaning-bearing units (morphemes and words) supported by a lower-level arrangement of meaningless phonemes. Meaning-bearing units can be rearranged to convey new messages, which is termed productivity (Ladd, 2014). This productive layer is the main one that realises novelty (although neologisms can also enable novelty at the phonological level). Even repetition of propositional content is typically realised by a different arrangement of words or morphemes.

In language, repetition as a structural relationship of (relatively) categorical sound elements does occur at the phonemic level, where learned structural relationships between phonemes hold within a particular language. This is comparable to reduction of sound categories in music: a finite set of phonemes and specific restrictions on their combinations reduces the uncertainty of which phoneme could follow in a sequence. Words are also repeated (although the size of the lexicon is vast). Indeed, long-term memory for melodies has been proposed to be comparable to the word lexicon (Peretz et al., 2009). Language therefore can be interpreted as balancing novelty and repetition, prototypically by differentially deploying them at different levels of the linguistic stream—phonological repeatability enables morphosyntactic and semantic novelty. Thus, in prototypical conversational language, novelty is realized at the morphosyntactic and semantic levels, with phonology and the rote-memory lexicon playing a supporting role.

Repetition in Language and Novelty in Music

What happens when repetition in language occurs at the productive level, that is with morphemes, words, and sentences? Some instances of repetition are relevant in a propositional sense: repetition of the same word or morpheme (reduplication) can be used for emphasis, or serve grammatical functions like plural marking (Hurch and Mattes, 2005). Repetition might also encourage the receiver to seek different interpretations of the phrase that are not apparent at the first glance, to resolve ambiguity (Knox, 1994).

Some situations however require the repetition of entire speech phrases, for example in ritualised contexts. When memorability needs to be enhanced, this is achieved by emphasising structural relationships in other levels of the speech phrase like intonation, stress, using rhyme or specific repeated syllabic patterns (e.g., poetic forms). This can also be observed in infant-directed speech which is very repetitive (Margulis, 2014). A link to memory might be that attention allocation seems to be related to surprising events (Forth et al., 2016; Koelsch et al., 2019). In the event-related potential, a mismatch negativity, indexing unpredicted and thus surprising events, is usually followed by a P3a component, associated with attention allocation (Schröger et al., 2015). More independence of levels of the speech signal would enable more surprising events due to possible unexpected interactions between levels, emphasising the structural relationships between them. On the other hand, predictive cues can also guide attention to a specific stimulus or stimulus feature (Gazzaley and Nobre, 2012), enhancing memory encoding. Our framework predicts that actions with an aesthetic goal, where we expect a greater independence of representational levels of the sequence and more variety in predictability, should be remembered better. In line with this, Margulis (2014) proposes that memory for music, poetry or utterances with schematic form, like jokes, is based more on acoustic surface structure than in conversational speech: speech involves attention allocation towards propositional content. Note that this enables paraphrasing the same propositional content with different words, which is more difficult for musical structure with notes or chords.

Turning to novelty, because attention is drawn to surprising events (Forth et al., 2016; Koelsch et al., 2019), listening to music that is highly predictable and unsurprising could lead to attentional shift and boredom. Thus, an additional pressure for music is to include a degree of novelty. One design solution to balance both novelty and repetition is meter (hierarchical relation of durational patterns relative to a beat). Meter provides a predictive framework within which novelty—unexpected and surprising events—is well defined (e.g., syncopation). Because multiple levels of the signal allow for predictability within and across levels by means of probabilistic relationships between their elements (tones, intervals, chord progressions, etc.), each level also allows for surprise. In repeated performances novelty can be provided by slight shifts in performance style, tempo, expression, etc., making the interpretation of familiar pieces a common focus of Western classical music concerts or opera. Concerning recordings, the possibility of attentional allocation to different levels of the piece with each repeated listening could be interpreted as listener-generated 'novelty', since new, potentially surprising, relations might be perceived. Thus, music also balances novelty and repetition in multiple ways, but they are quite distinct from prototypical conversational language.

In summary, both music and language balance repetition and novelty, but in different ways. While language usually allocates repeatability to the phonological level and novelty occurs at the morphosyntactic and semantic levels (related to propositionality), music typically allows both novelty and repeatability across all levels of the musical sequence, and meter seems to be especially crucial as a predictive layer throughout, enabling both prediction and surprise. However, language can also be repeatable at the word and sentence level. Thus, despite the clear validity and value of the two traditional dimensions on which music and language are differentiated—goal and novelty—certain design features are still not fully explained. We therefore suggest that understanding the design differences of language and music require a further explanatory dimension, to which we now turn.

INTERACTIVITY: THE CHORIC-DIALOGIC DIMENSION

Our proposed interactivity dimension concerns the temporal coordination of linguistic or musical productions of multiple participants. Both concurrent production, in choric mode, and dialogic turn-taking involve joint actions that are causally coupled, but they pose different constraints on predictability in sequences.

Dialogic Contexts in Speech

Two speakers talking at the same time constitute noise for each others' speech signals: overlapping signals make propositional content harder to decode (Fargier and Laganaro, 2019). Therefore (among possible roots in cooperative social interaction, see Levinson, 2016; Pika et al., 2018; Pougnault et al., 2022; but see Ravignani et al., 2019), dialogic contexts favour the avoidance of overlap and the coordination of turn-taking behaviour (Levinson and Torreira, 2015; Levinson, 2016), and signals should be designed such that receivers can predict the ending of an utterance (see, e.g., Castellucci et al., in press). Information should therefore be low (and predictability high) at the end of signal sequences. Given the requirement to reduce uncertainty in conveyed propositional messages, this should lead to high information density during most of the signal sequence (to optimally exploit the speech channel capacity, see above), with a decrease in information towards its end.

In dialogue, turn completions seem to be predicted based on both prosodic cues (Bögels and Torreira, 2015) and lexicosyntactic content (de Ruiter et al., 2006; Torreira et al., 2015), whereby semantic content seems to be more important in predicting the end of a speaker's phrase than syntax (Riest et al., 2015; see Jongman, 2021, for an overview). On the other hand, content prediction might be used to enable early response planning in parallel with comprehension of the current turn (Levinson and Torreira, 2015; Corps et al., 2018), which helps to avoid large gaps between turns that could themselves be interpreted as meaningful (e.g., Pomerantz and Heritage, 2013). Accordingly, Castellucci et al. (in press) proposed separate pathways for turn-timing and response planning. Since response planning requires neural resources that might compete with those utilised for comprehension (Bögels et al., 2015, 2018, see also Knudsen et al., 2020 for the role of backchannels, fillers and particles in this regard), it should start once the semantic uncertainty is low enough, and preferentially happen in places along the sequence that are low in information. Once the remainder of the sequence is highly predictable, interlocutors can exchange the roles of sender and receiver: taking turns. The next utterance will again start high in uncertainty, requiring informative events, until it nears its end.

The information density trajectory must be perceivable if the current receiver is to be able to predict when uncertainty is low enough to take the floor. This requires an ongoing monitoring of the information density in the received signal and a continuous prediction of the amount of uncertainty reduction that will follow from later events. That is, listeners need to predict when new events do not reduce uncertainty much further, probably based on the semantic uncertainty of the conveyed message and taking several past events into account to capture the general information trajectory. If the time point of turn-taking is marked by low information density, then this should be unambiguously distinguishable from local information density minima that may occur in the signal before. In the case of language, some words are more informative than others even when the speech utterance is not finished yet (Venhuizen et al., 2019). Because a speech act consists of multiple interacting and integrated levels—phoneme level, morpheme and word level, prosodic intonation, stress, lexical tone, etc., as well as paralinguistic information like facial or body expressions (see e.g., Holler et al., 2018; Wohltjen and Wheatley, 2021) and contextual cues in the environment (Ladd, 2014; Kuperberg and Jaeger, 2016; Holler and Levinson, 2019), each level can play its own part in reducing uncertainty about the propositional semantic content that should be conveyed. In line with this, phrase-final lengthening (a prosodic cue that can signal turn-ending, see Wightman et al., 1992) would decrease information per unit time, while in turn speakers accelerate their speech rate and thus information per unit time when they want to continue their utterance (Walker, 2010). We would predict that in order to mark the ending of a speech phrase and to take turns, the end of a speech phrase should be highly predictable across all levels of the sequence, even if one feature (like falling intonation) might preferentially mark the ending of the speech phrase at a prosodic level (see de Ruiter et al., 2006). Evidence seems to confirm that prosodic, lexical and syntactic levels interact to mark turn-taking (reviewed in Forth et al., 2016). This makes sense: if only one single level, such as falling intonation or lengthening, predicted the end of the current speaker's utterance, it would be highly surprising if an unpredicted, highly informative event occurred at another level (for example a highly surprising word). Information density would locally peak and the receiver would likely stop their preparation to take turns and re-allocate attention. Precisely this cross-level effect is used in investigations of speech processing by event-related potentials such as the N400, an evoked potential component which deflects negatively when target words in sentences are semantically unexpected (e.g., Grisoni et al., 2017; but see Maess et al., 2016 for a differential effect for verbs and nouns), even if there is no surprise at all other levels of the speech stream (e.g., intonation, syntax, phonology, etc.). Syntactic violations in contrast are indexed by earlier evoked components like the ELAN (Hahne and Friederici, 1999), illustrating that the brain uses prediction at multiple levels of the speech stream in parallel.

The example mentioned above also illustrates an interaction with the propositional-aesthetic dimension. If there is propositional content, the representational levels of the sequence that carry this content are of primary importance, while other levels support the semantic predictability and are thus less free to vary in their information trajectories than when propositional content is absent (as in nonsense speech or music). An event high enough in information at a supporting level might both alter the semantic understanding and disturb the turn-taking process. Thus, robust semantic understanding under the constraint of efficiency (Gibson et al., 2019) might facilitate successful turn-taking as well. In contrast, the less focus is on the propositional content, the less a need for hierarchy among levels exists, which means information density among levels should be freer to vary, possibly converging only towards the end of the signal sequence to enable turn-taking.

Interestingly, if there are more than two participants in the conversation, information density alone cannot be used to coordinate who will start the next speech act. In order to achieve this, paralinguistic information like pointing gestures, naming or rules about who speaks next need to apply (e.g., Mondada, 2013). The empirical prediction would be that the larger the group, the higher the danger of overlap between former receivers' initiation of speech acts, and the more the requirement for paralinguistic coordination (or an individual designated to choose the next speaker, e.g., the chair of a meeting). However, such overlap should occur only after one interlocutor ends their speech phrase, since all receivers can predict the end of the current speaker's turn.

To sum up, dialogic contexts require that endings of sequences are perceivable by a decrease in information density (which means an increase in predictability) across levels of the sequence, such that later events have on average lower information than former events. Both language and music are designed to fulfil this requirement in dialogic contexts. The propositional focus in most spoken dialogues adds an additional constraint that non-propositional levels should be subordinate to the levels conveying propositional content.

Choric Contexts in Music

Turning to the concurrent, choric pole, successful concurrent performance requires that signals do not disrupt processing when they overlap. One design feature that avoids masking by concurrent sounds (which is more effective when frequencies are more similar, see Moore, 2014) is to make them discrete and related by small integer ratios, as are the tones on musical scales. One example is the octave, whose existence across cultures is a statistical universal, and which enables all members of a group to sing in unison even when males' vocal range lowers after pubertal vocal change (Harries et al., 1998). In line with this, octave equivalence (perceiving two pitches as categorically the same when they are an octave apart) seems not to be perceived in a culture where individuals rarely sing together (Jacoby et al., 2019). This 'simple ratios' constraint interacts with another melodic design feature: reduction of the set of possible tones by limiting them to a small set (a 'scale'). Again, in information theoretic terms, establishing pitches on scales with strict tonal relations involves a reduction of the alphabet of allowed symbols along the fundamental frequency dimension (for a proposal for the roots of tonality in the physiology of hearing see Trainor, 2018). This limited set of possible tones allows individuals to join a music making chorus, match the produced sound sequences and/or complement them (for example in BaAka polyphonic singing, see Lewis, 2021), thus contributing to a unified sound entity in a coherent performance (a joint action with the deliberate coordination of actions, see Sebanz and Knoblich, 2009, 2021; Tomasello, 2010; Keller et al., 2014) rather than generating a set of sounds that are not causally coupled (cf. Ravignani et al., 2014). If scale tones are hierarchically related, the continuation of a melody can be predicted with a limited uncertainty by the participating individuals, allowing them to contribute in an ongoing manner, as well as allowing for variation and thus individuality in their contribution (cf. Savage et al., 2021).

The choric context also requires that events in separate sound streams should be tightly coordinated in time. Uncertainty in timing would lead to disintegration of concurrency and coordinated joint action. Therefore, the signal should be designed to enable high-precision temporal predictability throughout. One key design feature that enables such predictability is isochrony. Tight coordination however requires participants to attend to the other participants' actions, since ongoing coordination means prediction and monitoring on a moment-to-moment basis (see, e.g., Keller et al., 2014). If the next element can be precisely predicted in time, then the temporal information gain from each event is low, which lowers attentional demands (Koelsch et al., 2019). On the other hand, unpredicted events capture attention. How can these two requirements—high predictability and ongoing attention allocation—be aligned? If isochrony happens not at the level of each individual event, but on a meta-level, providing a scaffolding which still enables novelty (see the section about the novelty dimension), then both requirements can be fulfilled. The design solution satisfying these constraints is the concept of hierarchical meter, which allows certain placements in time and forbids others, but which also gives room for variability to create novelty since not each possible slot needs to be filled by an event. Again, meter represents a reduction in the alphabet, in this case a small set of possible onset and duration patterns relative to the beat.

Unlike in spoken language, there is much less noise and thus much less uncertainty when different participants in a choric performance contribute with different events to each of the levels of the musical sequence. We would therefore expect more degrees of freedom in terms of what event—which tone or chord—occurs than in speech deployed in a choric context (see below). However, the timing, that is when events occur in choric performances, is crucial and needs to be coordinated in a precise manner. Note that with a meter, events do not need to be played all at the same time (synchrony) nor be evenly spaced in time (isochrony). Rather, the burden is to keep events within the metrical scaffolding, so that the contributions of the participants relate to each other in the moment, or there would no longer be one coherent performance. Even in cases of notated music where it is clear which note must be played when, there is still the need for coordination among musicians, and in the absence of strict isochrony, a coherent performance needs to be otherwise synchronised, for example by using participants' body motion.

In summary, we argue that meter is a crucial design feature that develops in music deployed in a choric context, with the goal of balancing coordination and attention, while still permitting variation or improvisation. Meter provides a predictive fabric throughout an ongoing performance. This is complemented by a reduction of the alphabet in the tonal domain: a limited set of hierarchically related pitches allows accurate predictions of possible continuations of an ongoing acoustic performance, and of multiple different complementary event streams, without compromising the coherence of the overall performance as a joint action.

Dialogic Contexts in Music and Choric Contexts in Speech

Dialogic contexts also can occur in music, for example when several musicians take turns soloing in jazz, or in call-and-response singing. Here the same constraints must apply as for spoken dialogue, and we expect musical phrases to show lower information density towards the end. Thus, there must be a means to increase predictability at phrase endings. This aligns well with the notion of musical closure in harmony and melody, or the tendency of melodies to be shaped like an arch (Huron, 2006). Again, in music we expect that phrase endings tend to have low information density across all levels of the sequence on average. There are conventions which time or mark musical phrase endings, like the number of beats a soloist has available to perform their solo, or certain rhythmical or musical motifs, but we predict that even in these cases there should be decreasing information density towards musical phrase endings. An example illustrating how the prediction of a phrase ending can be disturbed if one level is high in information, at the level of harmony, are deceptive cadences in Western classical music, where a surprise chord replaces the highly predicted tonic as final chord, disrupting an expected sense of closure.

Again, we would not expect that one signal level is primary over another, at least if there is no propositional content (as typical in music). Rather, we expect that the levels of the sequence have more degrees of freedom to vary in their information content, as long as the phrase ending remains predictable. For example, a soloist in jazz might introduce a harmonic modulation (with high information content at the harmonic signal level) at the end of a phrase, while keeping melodic and rhythmic levels highly predictable and thus inviting a turn-taking event after which the next solo now occurs in a new key. That musical phrases in a more general sense exist, for example phrases in a solo Lied (song), could thus be an abstraction of deployment in a dialogic context.

What design features are predicted for language in a choric context? The particular tendency of overlapping speech stimuli to act as each other's noise (thus increasing uncertainty) means that simultaneity can only occur if precisely the same words or syllables are uttered at the same time, as happens for example in simultaneous speaking in religious or theatre performances (chanting). The prediction here is that attention is much more focussed on coordination than in dialogic speech acts, and that an isochronic and/or metrical scaffolding should develop (cf. Bowling et al., 2013). Since word order in such a sequence must be pregiven, and thus the information density of the word and phoneme level would be constant and very low, we would expect suprasegmental or paralinguistic levels to vary more in information density and to be used to reduce uncertainty regarding timing. That means body motion, facial expressions or prosodic intonation should be more pronounced in a spoken choric context.

COMPARATIVE PERSPECTIVE

We suggest that the perspective of deployment of sound sequences in a three-dimensional quality space (goal: propositional-aesthetic, novelty: repetition-novelty, interactivity: dialogic-choric) along with the information theoretic concept of reducing uncertainty can also be used in bioacoustics research. We are aware that transferring a concept directly from human language and music might not work for animal communication, but we think that, especially for complex vocal displays in birds or whales, our framework may provide some insight, since these 'animal songs' bear some structural similarities to language and music (Rohrmeier et al., 2015). Information theory has been employed in animal communication research, although the term 'information' has often been used in a colloquial sense or inconsistently (Stegmann, 2013; Fitch, 2014). We hope that applying our framework provides some useful insights regarding complex vocal displays, along with call combinations and sequences (Engesser and Townsend, 2019). The information theoretic framework also encourages us to consider non-vocal levels like body motion that might be especially relevant for mating displays (Mitoyen et al., 2019).

When talking about 'goals' in animal communication, it is necessary to consider that goals other than propositional or aesthetic ones (e.g., social bonding by vocal convergence, providing information about sex or status, etc.) might use the vocal domain independently of the dimensions we derived for human language and music. Often it is unclear what animals communicate in their vocalisations, and some researchers question the existence of communicative content at all, proposing instead that animals manipulate others by means of their signals (Owren et al., 2010; Stegmann, 2013). On the other hand, there is evidence in some cases that animal calls can be functionally referential, reliably co-occurring with external entities (Seyfarth et al., 1980; Price et al., 2015), but little evidence that complex vocalisations like bird song or whale song have functional referential meaning (Engesser and Townsend, 2019). Analysing information trajectories across multiple levels of the sequence might give additional insight into this important question, but this requires that several such levels can be disentangled in the first place, which might not be easily the case in animal vocalisations.

Duality of patterning appears to exist only in human language (Bowling and Fitch, 2015). When vocalisations have a propositional goal (in the sense of referring to external entities and eliciting reliable behavioural responses), we would therefore expect this content to be encoded in a whole structure across levels of the communicative signal, similar to music when deployed in a propositional context. We would also expect high predictability within the signal (since surprising elements could potentially add uncertainty to the inferred content) and probably a relatively uniform information density to enhance transmission. In turn, given that reward systems in other animals (Connell and Hofmann, 2011) would also be related to prediction and surprise (Schultz, 2016), we would expect that sound sequences with independence in information variation between different levels are more likely to fulfil an aesthetic rather than a propositional goal. We speculate that, in mating displays, constrained surprise rather than complexity is what makes displaying individuals attractive for mating, accounting for the common occurrence of individually distinctive songs and song repertoires.

Animal vocalisations like bird or whale song consist of subunits that can be structured in a hierarchical manner, and thus bear some structural similarities to human language and music (Payne and McVay, 1971; Rohrmeier et al., 2015). Repetition of subunits and their recombination often characterise such complex vocal displays. We would expect that subunits used in repetition are categorical to reduce the alphabet and enhance predictability. We would also expect a tendency towards hierarchical organisation in the temporal domain for longer vocalisations relative to shorter ones within comparable species.

What about novelty? Novelty might be realised by using new sounds with high surprisal. For example, the best documented example of 'vocal' learning in chimpanzees involved the addition of a lip buzz or 'raspberry' at the end of pant-hoot sequences (Marshall et al., 1999). However, since such new sounds would not reduce uncertainty in decoding a message for the receiver, unless an association to some external entity is learned, they would probably not have 'content' in the sense of functional referentiality. Novelty can also be realised by rearranging subunits that are structurally highly predictable, as is often characteristic of bird song (Kroodsma, 1978) and somewhat structurally similar to human song (Lomax, 1968).

Repetition can also have referential relevance in call combinations or sequences (see Engesser and Townsend, 2019), for example in chickadee call repeats (e.g., Hailman and Ficken, 1987). Some of these calls appear to be categorical (in the sense of being discriminable) while others are higher in variability. Moreover, some of these combinations are rather short. It seems that such call combinations may not be straightforwardly accounted for by our three-dimensional framework. However, we expect that for stand-alone calls, uncertainty should be higher than when another call (same or different) is appended, as evaluated by changes in attention of receivers. Calls that occur only in fixed combinations might however not induce surprise since there is little uncertainty in referential content when encountering them.

Using the predictions derived from deployment of human language and music in choric or dialogic contexts could reveal whether animal vocalisations show design features based on differential temporal coordination of signals (Ravignani et al., 2019). Coordinated vocal displays, both concurrent and turn-taking, are widespread in animal communication. Duetting for example is widely observed in bird species. Investigating the information trajectories could reveal whether and how individuals relate to each other in their vocalisations. Interlocking vocalisations between two pair-bonded or courting birds for example could be investigated to see whether it is based on decreasing information density after one phrase. This would indicate a dialogic deployment. Competitive vocal displays, for example between two males in territorial contexts, might show dialogic design features with decreasing information density at the end of an individual's vocal phrase. If the display involves masking the competitor, we would predict that overlap is done in moments of high information. On the other hand, duets might be based on a coordination of each event with predictability enabled by some isochronous scaffolding, similar to a musical piece where performers do not play each note concurrently but nonetheless contribute to a unified musical piece. Such 'hocketting' would be indicative of a choric deployment. Bird species differ in their preference for overlap or overlap avoidance and in their flexibility depending on social and environmental context (Pika et al., 2018). Starlings seem to be a particularly interesting model species, showing both turn-taking and overlapping vocalisations depending on social context, and varying in their proportion of either by sociality of subspecies (Henry et al., 2015). We would predict that in all these cases information trajectories are perceived by receivers and used to coordinate their own vocalisation in relation to that of the other individual(s).

Other vocal displays that might be interesting to investigate in this context are antiphonal calling in elephants (Soltis et al., 2005) or bats (Carter et al., 2008) or duets in gibbons (Geissmann, 2002) or indris (Gamba et al., 2014), as well as group calling in meerkats (Demartsev et al., 2018). Castellucci et al. (in press) suggest singing mice as a model species for coordinated vocal timing. Bottlenose dolphins can switch between alternating vocalisations and simultaneous duetting (Lilly and Miller, 1961). We would predict that their alternating vocalisations are coordinated by decrease in information density at the end of phrases, while their duetting might be coordinated on an event-by-event basis as in choric deployment. A particularly interesting case is the group territorial display of plain-tailed wrens (*Thryothorus euophrys*) where group choruses including multiple males and multiple females are used to jointly defend a territory. In this species simultaneous performance occurs within sexes, with turn-taking between sexes (Mann et al., 2006) within a chorus. We expect within-sex performance to be characterised by a tendency toward an isochronous scaffolding, comparable to meter, and turn-taking between sexes to be coordinated by terminal decreases in information density.

Animal group communicative (choric) displays often appear uncoordinated, for example in howler monkeys (Sekulic, 1982) or, at the other extreme, highly synchronised, as in some fireflies (see Ravignani et al., 2014). Group coordination can either be based on predictive or reactive behaviour, and we suggest that information trajectories could be examined to address this issue. If individuals in a group (for example howling wolves) vocalise in a coordinated manner, we expect each individual's contribution to reduce uncertainty about event timing for the other individuals. This would be indicative that the performance aims at creating a coherent entity, implying choric design features.

CONCLUSION

In this paper we have proposed a framework based on information theory, adopting a design stance to investigate differences in language and music. We suggested that some key design features of music and language can be explained as responses to their deployment between dialogic and choric poles of a continuum rooted in interactive performative constraints. This interactivity (choric-dialogic) dimension complements the widely recognised goal (propositional-aesthetic) and novelty (repetition-novelty) dimensions, forming a three-dimensional framework within which different forms of music and language can be placed, and their design differences understood.

We argued that the goal and novelty dimensions alone are not sufficient to explain differences in design features between music and language: the interactivity between individuals is crucial. For dialogic contexts, the only coordinative constraint concerns the timing of the turn-taking between individuals, which should be indexed by a lower information density towards the end of phrases, across all levels of the sonic stream. When there is also a propositional goal, non-propositional levels should be constrained in their variability to support the decoding of propositional content. Information rate should be high—realised mainly by novelty at the propositional level—until turn-taking is indicated. Conversational speech acts and turn-taking are the prototypical features that fulfil these requirements.

In contrast, choric performance requires tight temporal coordination of all contributing individuals, as well as avoidance of masking by simultaneous sound events, enabled by high predictability in timing and frequency of sonic events. When there is also a pressure for novelty, isochronous meter and discrete pitches in scales are design solutions that enable a group of participants to join in making a coherent sound sequence, allowing both novelty and repeatability. This contributes to the independence of multiple levels in the sonic stream with regard to surprise and uncertainty, making these independent levels well suited to exploit the human reward system. The prototypical form of choric performance is joint music making, but our framework also encompasses non-canonical forms of music and language like chant, poetry, or exchange of musical solos, thereby avoiding an overly simplistic dichotomy between language and music. Furthermore, our framework supports comparisons of different forms of communication across distinct modalities and species and can help to generate new hypotheses about optimal design of signals satisfying multiple different requirements. We hope that this framework will also be fruitfully employed in animal communication research, broadening the scope of comparisons with music and/or language.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This article was supported by the Austrian Science Fund (FWF) number W1262-B29 (DK Grant Cognition & Communication). FH was also supported by a VDS CoBeNe final fellowship. The funders had no role in decision on the article content, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

The authors thank Tudor Popescu, Sabrina Engesser, and Gregg Castellucci for helpful discussions and comments, and two reviewers for helpful comments on the manuscript.

REFERENCES

Agus, T. R., Thorpe, S. J., and Pressnitzer, D. (2010). Rapid formation of robust auditory memories: insights from noise. *Neuron* 66, 610–618. doi: 10.1016/j.neuron.2010.04.014

Aylett, M., and Turk, A. (2004). The smooth signal redundancy hypothesis : A functional explanation for relationships between hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech*. *Lang. Speech* 47, 31–56. doi: 10.1177/00238309040470010201

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, 11818–11823. doi: 10.1073/pnas.191355898

Bögels, S., Casillas, M., and Levinson, S. C. (2018). Planning versus comprehension in turn-taking: Fast responders show reduced anticipatory processing of the question. *Neuropsychologia* 109, 295–310. doi: 10.1016/j.neuropsychologia.2017.12.028

Bögels, S., Magyari, L., and Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Sci. Rep.* 5:12881. doi: 10.1038/srep12881

Bögels, S., and Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *J. Phon.* 52, 46–57. doi: 10.1016/j.wocn.2015.04.004

Bowling, D. L., and Fitch, W. T. (2015). Do animal communication systems have phonemes? *Trends Cogn. Sci.* 19, 555–557. doi: 10.1016/j.tics.2015.08.011

Bowling, D. L., Herbst, C. T., and Fitch, W. T. (2013). Social origins of rhythm? Synchrony and temporal regularity in human vocalization. *PLoS One* 8:e80402. doi: 10.1371/journal.pone.0080402

Brown, S. (2007). Contagious heterophony: A new theory about the origins of music. *Music. Sci.* 11, 3–26. doi: 10.1177/102986490701100101

Brown, R. M., and Palmer, C. (2012). Auditory—motor learning influences auditory memory for music. *Mem. Cogn.* 40, 567–578. doi: 10.3758/s13421-011-0177-x

Busnel, R.-G., and Classe, A. (1976). *Whistled Languages*. Berlin, Heidelberg, Germany: Springer.

Carter, G. G., Skowronski, M. D., Faure, P. A., and Fenton, B. (2008). Antiphonal calling allows individual discrimination in white-winged vampire bats. *Anim. Behav.* 76, 1343–1355. doi: 10.1016/j.anbehav.2008.04.023

Castellucci, G. A., Guenther, F. H., and Long, M. A. (in press). A theoretical framework for human and nonhuman vocal interaction. *Annu. Rev. Neurosci.* 45.

Chatwin, B. (1987). *The Songlines*. New York, NY: Viking.

Cheung, V. K. M., Harrison, P. M. C., Meyer, L., Pearce, M. T., Haynes, J., Cheung, V. K. M., et al. (2019). Uncertainty and surprise jointly predict musical pleasure and amygdala, hippocampus, and auditory cortex activity. *Curr. Biol.* 29, 4084.e4–4092.e4. doi: 10.1016/j.cub.2019.09.067

Cirelli, L. K., and Trehub, S. E. (2020). Familiar songs reduce infant distress. *Dev. Psychobiol.* 56, 861–868. doi: 10.1037/dev0000917

Colombo, M., and Wright, C. (2017). Brain and cognition explanatory pluralism: An unrewarding prediction error for free energy theorists. *Brain Cogn.* 112, 3–12. doi: 10.1016/j.bandc.2016.02.003

Connell, L. A. O., and Hofmann, H. A. (2011). The vertebrate mesolimbic reward system and social behavior network: A comparative synthesis. *J. Comp. Neurol.* 519, 3599–3639. doi: 10.1002/cne.22735

Corbeil, M., Trehub, S. E., and Peretz, I. (2016). Singing delays the onset of infant distress. *Infancy* 21, 373–391. doi: 10.1111/infa.12114

Corps, R. E., Crossley, A., Gambi, C., and Pickering, M. J. (2018). Early preparation during turn-taking: listeners use content predictions to determine what to say but not when to say it. *Cognition* 175, 77–95. doi: 10.1016/j.cognition.2018.01.015

Cross, I., Fitch, W. T., Aboitiz, F., Iriki, A., Jarvis, E. D., Lewis, J., et al. (2013). "Culture and evolution," in *Language, Music, and the Brain*. ed. M. A. Arbib (Cambridge, Massachusetts; London, England: The MIT Press), 541–562.

Crupi, V., Nelson, J. D., Meder, B., Cevolani, G., and Tentori, K. (2018). Generalized information theory meets human cognition: introducing a unified framework to model uncertainty and information search. *Cogn. Sci.* 42, 1410–1456. doi: 10.1111/cogs.12613

Csete, M. E., and Doyle, J. C. (2002). Reverse engineering of biological complexity. *Science* 295, 1664–1669. doi: 10.1126/science.1069981

Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*. London, UK: J. Murray.

de Fleurian, R., Harrison, P. M. C., Pearce, M. T., and Quiroga-Martinez, D. R. (2019). Reward. *Acad. Sci.* 116, 20813–20814. doi: 10.1073/pnas.1913244116

Demartsev, V., Strandburg-Peshkin, A., Ruffner, M., and Manser, M. (2018). Vocal turn-taking in Meerkat group calling sessions. *Curr. Biol.* 28, 3661–3666.e3. doi: 10.1016/j.cub.2018.09.065

de Ruiter, J. P., Mitterer, H., and Enfield, N. J. (2006). Projecting the end of a Speaker's turn: A cognitive cornerstone of conversation. *Language* 82, 515–535. doi: 10.1353/lan.2006.0130

Deutsch, D., Henthorn, T., and Lapidis, R. (2011). Illusory transformation from speech to song. *J. Acoust. Soc. Am.* 129, 2245–2252. doi: 10.1121/1.3562174

Engesser, S., and Townsend, S. W. (2019). Combinatoriality in the vocal systems of nonhuman animals. *Wiley Interdiscip. Rev. Cogn. Sci.* 10:e1493. doi: 10.1002/wcs.1493

Fargier, R., and Laganaro, M. (2019). Interference in speaking while hearing and vice versa. *Sci. Rep*. 9:5375. doi: 10.1038/s41598-019-41752-7

Ferreri, L., Mas-Herrero, E., Zatorre, R. J., Ripollés, P., Gomez-Andres, A., Alicart, H., et al. (2019). Dopamine modulates the reward experiences elicited by music. *Proc. Natl. Acad. Sci. USA*. 116, 3793–3798. doi: 10.1073/pnas.1811878116

Fitch, W. T. (2006). The biology and evolution of music: A comparative perspective. *Cognition* 100, 173–215. doi: 10.1016/j.cognition.2005.11.009

Fitch, W. T. (2014). Information considered harmful in animal communication. *Curr. Biol.* 24, R8–R10. doi: 10.1016/j.cub.2013.11.020

Forth, J., Agres, K., Purver, M., and Wiggins, G. A. (2016). Entraining IDyOT: timing in the information dynamics of thinking. *Front. Psychol.* 7:1575. doi: 10.3389/fpsyg.2016.01575

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Gamba, M., Torti, V., Bonadonna, G., Guzzo, G., and Giacoma, C. (2014). "Overlapping and synchronization in the song of the Indris (Indri indri)." in *The Evolution of Language: Proceedings of the 10th International Conference*. eds. E. A. Cartmill, S. Roberts, H. Lyn and H. Cornish; April 14-17, 2014; Vienna, Austria (Singapore: World Scientific Publishing Co.), 90–97.

Gazzaley, A., and Nobre, A. C. (2012). Top-down modulation: bridging selective attention and working memory. *Trends Cogn. Sci.* 16, 129–135. doi: 10.1016/j.tics.2011.11.014

Geissmann, T. (2002). Duet-splitting and the evolution of gibbon songs. *Biol. Rev.* 77, 57–76. doi: 10.1017/S1464793101005826

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., et al. (2019). How efficiency shapes human language. *Trends Cogn. Sci.* 23, 389–407. doi: 10.1016/j.tics.2019.02.003

Glowinski, D., Coletta, P., Volpe, G., Camurri, A., Chiorri, C., and Schenone, A. (2010). "Multi-scale entropy analysis of dominance in social creative activities." in *MM'10- Proceedings of the ACM Multimedia 2010 International Conference*. eds. A. del Bimbo, S.-F. Chang and A. Smeulders; October 25-29, 2010 (New York, NY: Association for Computing Machinery), 1035–1038.

Glowinski, D., and Mancini, M. (2011). "Towards real-time affect detection based on sample entropy analysis of expressive gesture," in *Affective Computing and Intelligent Interaction: LNCS. Vol. 6974.* eds. S. D'Mello, A. Graesser, B. Schuller and J.-C. Martin (Berlin, Heidelberg, Germany: Springer), 527–537.

Glowinski, D., Mancini, M., Cowie, R., Camurri, A., Chiorri, C., and Doherty, C. (2013). The movements made by performers in a skilled quartet: A distinctive pattern, and the function that it serves. *Front. Psychol.* 4:841. doi: 10.3389/fpsyg.2013.00841

Grice, H. P. (1975). "Logic and conversation," in *The Logic of Grammar*. eds. D. Davidson and G. Harman (Encino, CA: Dickenson), 64–153.

Grisoni, L., Miller, T. M., and Pulvermüller, F. (2017). Neural correlates of semantic prediction and resolution in sentence processing. *J. Neurosci* 37, 4848–4858. doi: 10.1523/JNEUROSCI.2800-16.2017

Hahne, A., and Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis: early automatic and late controlled processes. *J. Cogn. Neurosci.* 11, 194–205. doi: 10.1162/089892999563328

Hailman, J. P., and Ficken, M. S. (1987). Combinatorial animal communication with computable syntax: Chick-a-dee calling qualifies as 'language' by structural linguistics. *Anim. Behav*. 34, 1899–1901. doi: 10.1016/S0003-3472(86)80279-2

Hansen, N. C., Dietz, M. J., and Vuust, P. (2017). Commentary: predictions and the brain: how musical sounds become rewarding. *Front. Hum. Neurosci.* 11:168. doi: 10.3389/fnhum.2017.00168

Hansen, N. C., Kragness, H. E., Vuust, P., Trainor, L., and Pearce, M. T. (2021). Predictive uncertainty underlies auditory boundary perception. *Psychol. Sci.* 32, 1416–1425. doi: 10.1177/0956797621997349

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

Harries, M., Hawkins, S., Hacking, J., and Hughes, I. (1998). Changes in the male voice at puberty: vocal fold length and its relationship to the fundamental frequency of the voice. *J. Laryngol. Otol.* 112, 451–454. doi: 10.1017/s0022215100140757

Henry, L., Craig, A., Lemasson, A., and Hausberger, M. (2015). Social coordination in animal vocal interactions. Is there any evidence of turn-taking? The starling as an animal model. *Front. Psychol.* 6:1416. doi: 10.3389/fpsyg.2015.01416

Herff, S. A., Harasim, D., Cecchetti, G., Finkensiep, C., and Rohrmeier, M. A. (2021). "Hierarchical syntactic structure predicts listeners' sequence completion in music." in *Proceedings of the Annual Meeting of the Cognitive Science Society*. *Vol. 43.* July 26–29, 2021; 276–281.

Hlaváčková-Schindler, K., Paluš, M., Vejmelka, M., and Bhattacharya, J. (2007). Causality detection based on information-theoretic approaches in time series analysis. *Phys. Rep.* 441, 1–46. doi: 10.1016/j.physrep.2006.12.004

Hockett, C. (1960). The origin of speech. *Sci. Am.* 203, 88–96. doi: 10.1038/scientificamerican0960-88

Holler, J., Kendrick, K. H., and Levinson, S. C. (2018). Processing language in face-to-face conversation: Questions with gestures get faster responses. *Psychon. Bull. Rev*. 25, 1900–1908. doi: 10.3758/s13423-017-1363-z

Holler, J., and Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends Cogn. Sci.* 23, 639–652. doi: 10.1016/j.tics.2019.05.006

Hollerman, J. R., and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* 1, 304–309. doi: 10.1038/1124

Honing, H. (2018). On the biological basis of musicality. *Ann. N. Y. Acad. Sci.* 1423, 51–56. doi: 10.1111/nyas.13638

Huettig, F., and Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Lang. Cogn. Neurosci.* 31, 19–31. doi: 10.1080/23273798.2015.1072223

Hurch, B., and Mattes, V. (2005). *Studies on Reduplication*. Berlin, Germany: de Gruyter Mouton.

Huron, D. (2006). *Sweet Anticipation*. Cambridge, Massachusetts: MIT Press.

Huron, D. (2016). "Aesthetics," in *The Oxford Handbook of Music Psychology. 2nd Edn.* eds. S. Hallam, I. Cross and M. Thaut (Oxford, UK: Oxford University Press), 233–245.

Jacoby, N., Undurraga, E. A., McPherson, M. J., and McDermott, J. H. (2019). Universal and non-universal features of musical pitch perception revealed by singing. *Curr. Biol.* 29, 3229.e12–3243.e12. doi: 10.1016/j.cub.2019.08.020

Jarvis, E. D. (2019). Evolution of vocal learning and spoken language. *Science* 366, 50–54. doi: 10.1126/science.aax0287

Jongman, S. R. (2021). "The attentional demands of combining comprehension and production in conversation," in *Psychology of Learning and Motivation - Advances in Research and Theory. 1st Edn. Vol. 74*. ed. K. D. Federmeie (Cambridge, Massachusetts: Elsevier Inc.), 95–140.

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

Knox, L. (1994). "Repetition and relevance: self-repetition as a strategy for initiating cooperation in nonnative/native speaker conversations," in *Repetition in Discourse: Interdisciplinary Perspectives*. *Vol. 1*. ed. B. Johnstone (Norwood, NJ: Ablax Publishing Corporation), 195–206.

Knudsen, B., Creemers, A., and Meyer, A. S. (2020). Forgotten little words: How backchannels and particles may facilitate speech planning in conversation? *Front. Psychol*. 11:593671. doi: 10.3389/fpsyg.2020.593671

Koelsch, S., Vuust, P., and Friston, K. (2019). Predictive processes and the peculiar case of music. *Trends Cogn. Sci.* 23, 63–77. doi: 10.1016/j.tics.2018.10.006

Kotz, S. A., Ravignani, A., and Fitch, W. T. (2018). The evolution of rhythm processing. *Trends Cogn. Sci.* 22, 896–910. doi: 10.1016/j.tics.2018.08.002

Krebs, J. R., and Davies, N. B. (eds.) (1997). *Behavioural Ecology: An Evolutionary Approach. 4th Edn.* Hoboken, New Jersey, USA: Wiley-Blackwell.

Kroodsma, D. E. (1978). Continuity and versatility in bird song: support for the monotony–threshold hypothesis. *Nature* 274, 681–683. doi: 10.1038/274681a0

Kuperberg, G. R., and Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* 31, 32–59. doi: 10.1080/23273798.2015.1102299

Ladd, D. R. (2014). *Simultaneous Structure in Phonology*. Oxford, UK/New York: Oxford University Press.

Levinson, S. C. (2016). Turn-taking in human communication - origins and implications for language processing. *Trends Cogn. Sci.* 20, 6–14. doi: 10.1016/j.tics.2015.10.010

Lerdahl, F., and Krumhansl, C. L. (2007). Modeling tonal tension. *Music Percept.* 24, 329–366. doi: 10.1525/mp.2007.24.4.329

Levinson, S. C., and Holler, J. (2014). The origin of human multi-modal communication. *Philos. Trans. R. Soc. B: Biol. Sci.* 369:20130302. doi: 10.1098/rstb.2013.0302

Levinson, S. C., and Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Front. Psychol.* 6:20130302. doi: 10.3389/fpsyg.2015.00731

Lewis, J. (2021). "Why music matters: social aesthetics and cultural transmission," in *Music, Dance, Anthropology*. ed. S. Cottrell (Oxford, UK: Sean Kingston Publishing).

Lilly, J. C., and Miller, A. M. (1961). Sounds emitted by the bottlenose dolphin. *Science* 133, 1689–1693. doi: 10.1126/science.133.3465.1689

Lomax, A. (1968). *Folk Song Style and Culture*. Washington, DC: American Association for the Advancement of Science.

Maess, B., Mamashli, F., Obleser, J., Helle, L., and Friederici, A. D. (2016). Prediction signatures in the brain: semantic pre-activation during language comprehension. *Front. Hum. Neurosci.* 10:591. doi: 10.3389/fnhum.2016.00591

Mahowald, K., Dautriche, I., Braginsky, M., and Gibson, E. (2020). Efficient communication and the organization of the lexicon. PsyArXiv [Preprint], 1–46. doi:10.31234/osf.io/4an6v

Mahowald, K., Fedorenko, E., Piantadosi, S. T., and Gibson, E. (2013). Info/information theory: speakers choose shorter words in predictive contexts. *Cognition* 126, 313–318. doi: 10.1016/j.cognition.2012.09.010

Mann, N. I., Dingess, K. A., and Slater, P. J. B. (2006). Antiphonal four-part synchronized chorusing in a Neotropical wren. *Biol. Lett.* 2, 1–4. doi: 10.1098/rsbl.2005.0373

Margulis, E. H. (2014). *On Repeat: How Music Plays the Mind*. Oxford, UK/New York: Oxford University Press.

Margulis, E. H., and Simchy-Gross, R. (2016). Repetition enhances the musicality of randomly generated tone sequences. *Music. Percept.* 33, 509–514. doi: 10.1525/mp.2016.33.4.509

Marshall, A. J., Wrangham, R. W., and Arcadi, A. C. (1999). Does learning affect the structure of vocalizations in chimpanzees? *Anim. Behav.* 58, 825–830. doi: 10.1006/anbe.1999.1219

Maynard Smith, J. (2000). The concept of information in biology. *Philos. Sci.* 67, 177–194. doi: 10.1017/CBO9780511778759.007

McDonnell, M. D., Ikeda, S., and Manton, J. H. (2011). An introductory review of information theory in the context of computational neuroscience. *Biol. Cybern.* 105, 55–70. doi: 10.1007/s00422-011-0451-9

Mehr, S. A., Krasnow, M. M., Bryant, G. A., and Hagen, E. H. (2021). Origins of music in credible signaling. *Behav. Brain Sci.* 44:E60. doi: 10.1017/S0140525X20000345

Mehr, S. A., Singh, M., York, H., Glowacki, L., and Krasnow, M. M. (2018). Form and function in human song. *Curr. Biol.* 28, 356.e5–368.e5. doi: 10.1016/j.cub.2017.12.042

Meyer, L. P. (1956). *Emotion and Meaning in Music.* Chicago: University of Chicago Press.

Mitoyen, C., Quigley, C., and Fusani, L. (2019). Evolution and function of multimodal courtship displays. *Ethology* 125, 503–515. doi: 10.1111/eth.12882

Moore, B. C. J. (2014). "Psychoacoustics," in *Springer Handbook of Acoustics*. ed. T. D. Rossing (Berlin, Heidelberg, Germany: Springer), 475–517.

Mondada, L. (2013). Embodied and spatial resources for turn-taking in institutional multi-party interactions: Participatory democracy debates. *J. Pragmat*. 46, 39–68. doi: 10.1016/j.pragma.2012.03.010

Morgan, E., Fogel, A., Nair, A., and Patel, A. D. (2019). Statistical learning and gestalt-like principles predict melodic expectations. *Cognition* 189, 23–34. doi: 10.1016/j.cognition.2018.12.015

Narmour, E. (1990). *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model*. Chicago: University of Chicago Press.

Owren, M. J., Rendall, D., and Ryan, M. J. (2010). Redefining animal signaling: influence versus information in communication. *Biol. Philos.* 25, 755–780. doi: 10.1007/s10539-010-9224-4

Payne, R. S., and McVay, S. (1971). Songs of humpback whales. *Science* 173, 585–597. doi: 10.1126/science.173.3997.585

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

Peretz, I., Gosselin, N., Belin, P., Zatorre, R. J., Plailly, J., and Tillmann, B. (2009). Music lexical networks: The cortical organization of music recognition. *Ann. N. Y. Acad. Sci.* 1169, 256–265. doi: 10.1111/j.1749-6632.2009.04557.x

Piantadosi, S. T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proc. Natl. Acad. Sci.* 108, 3526–3529. doi: 10.1073/pnas.1012551108

Piantadosi, S. T., Tily, H., and Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition* 122, 280–291. doi: 10.1016/j.cognition.2011.10.004

Pika, S., Wilkinson, R., Kendrick, K. H., and Vernes, S. C. (2018). Taking turns: bridging the gap between human and animal communication. *Proc. R. Soc. B Biol. Sci.* 285:20180598. doi: 10.1098/rspb.2018.0598

Pomerantz, A., and Heritage, J. (2013). "Preference," in *Handbook of Conversation Analysis. 1st Edn*. eds. J. Sidnell and T. Stivers (Chichester, West Sussex, UK: Wiley-Blackwell), 210–228.

Pougnault, L., Levréro, F., Leroux, M., Paulet, J., Bombani, P., Dentressangle, F., et al. (2022). Social pressure drives "conversational rules" in great apes. *Biol. Rev*. 97, 749–765.

Price, T., Wadewitz, P., Cheney, D., Seyfarth, R., Hammerschmidt, K., and Fischer, J. (2015). Vervets revisited: a quantitative analysis of alarm call structure and context specificity. *Sci. Rep.* 5, 1–11. doi: 10.1038/srep13220

Ravignani, A., Bowling, D., and Fitch, W. T. (2014). Chorusing, synchrony and the evolutionary functions of rhythm. *Front. Psychol.* 5:1118. doi:10.3389/fpsyg.2014.01118

Ravignani, A., Verga, L., and Greenfield, M. D. (2019). Interactive rhythms across species : The evolutionary biology of animal chorusing and turn-taking. *Ann. N. Y. Acad. Sci*. 1453, 12–21. doi: 10.1111/nyas.14230

Richardson, R. C. (2003). Engineering design and adaptation. *Philos. Sci.* 70, 1277–1288. doi: 10.1086/377407

Riest, C., Jorschick, A. B., and de Ruiter, J. P. (2015). Anticipation in turn-taking: mechanisms and information sources. *Front. Psychol.* 6:89. doi: 10.3389/fpsyg.2015.00089

Rohrmeier, M. A., and Koelsch, S. (2012). Predictive information processing in music cognition. A critical review. *Int. J. Psychophysiol.* 83, 164–175. doi: 10.1016/j.ijpsycho.2011.12.010

Rohrmeier, M., and Pearce, M. (2018a). "Musical syntax I: theoretical perspectives," in *Springer Handbook of Systematic Musicology*. ed. R. Bader (Berlin, Heidelberg, Germany: Springer Verlag), 473–486.

Rohrmeier, M., and Pearce, M. (2018b). "Musical syntax II: empirical perspectives," in *Springer Handbook of Systematic Musicology*. ed. R. Bader (Berlin, Heidelberg, Germany: Springer Verlag), 473–486.

Rohrmeier, M., Zuidema, W., Wiggins, G. A., and Scharff, C. (2015). Principles of structure building in music, language and animal song. *Philos. Trans. R Soc. Lond. B Biol. Sci.* 370:20140097. doi: 10.1098/rstb.2014.0097

Rowland, J., Kasdan, A., and Poeppel, D. (2019). There is music in repetition: looped segments of speech and nonspeech induce the perception of music in a time-dependent manner. *Psychon. Bull. Rev.* 26, 583–590. doi: 10.3758/s13423-018-1527-5

Salimpoor, V. N., Benovoy, M., Larcher, K., Dagher, A., and Zatorre, R. J. (2011). Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nat. Publ. Group* 14, 257–262. doi: 10.1038/nn.2726

Salimpoor, V. N., Zald, D. H., Zatorre, R. J., Dagher, A., and McIntosh, A. R. (2015). Predictions and the brain: how musical sounds become rewarding. *Trends Cogn. Sci.* 19, 86–91. doi: 10.1016/j.tics.2014.12.001

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.* 112, 8987–8992. doi: 10.1073/pnas.1414495112

Savage, P. E., Loui, P., Tarr, B., Schachner, A., Glowacki, L., Mithen, S., et al. (2021). Music as a coevolved system for social bonding. *Behav. Brain Sci.* 44:E59. doi: 10.1017/S0140525X20000333

Schröger, E., Marzecová, A., and Sanmiguel, I. (2015). Attention and prediction in human audition: a lesson from cognitive psychophysiology. *Eur. J. Neurosci.* 41, 641–664. doi: 10.1111/ejn.12816

Schultz, W. (2016). Dopamine reward prediction error coding. *Dialogues Clin. Neurosci.* 18, 23–32. doi: 10.31887/DCNS.2016.18.1/wschultz

Schultz, W. (2017). Reward prediction error. *Curr. Biol.* 27, R369–R371. doi: 10.1016/j.cub.2017.02.064

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599. doi: 10.1126/science.275.5306.1593

Schwitters, K. (1973). *Das literarische Werk, 5: Manifeste und kritische Prosa*. eds. K. Schwitters and F. Lach (Köln, Germany: DuMont Schauberg).

Sebanz, N., and Knoblich, G. (2009). Prediction in joint action: what, when, and where. *Top. Cogn. Sci.* 1, 353–367. doi: 10.1111/j.1756-8765.2009.01024.x

Sebanz, N., and Knoblich, G. (2021). Progress in joint action research. *Curr. Dir. Psychol. Sci.* 30, 138–143. doi: 10.1177/0963721420984425

Seifert, U., Verschure, P. F. M. J., Arbib, M. A., Cohen, A. J., Fogassi, L., Fritz, T. H., et al. (2013). "Semantics of internal and external worlds," in *Language, Music, and the Brain: A Mysterious Relationship*. ed. M. A. Arbib (Cambridge, Massachusetts; London, England: The MIT Press), 203–232.

Sekulic, R. (1982). The function of howling in red howler monkeys (*Alouatta seniculus*). *Behaviour* 81, 38–54. doi: 10.1163/156853982X00517

Seyfarth, B. Y. R. M., Cheney, D. L., and Marler, P. (1980). Vervet monkey alarm calls: semantic communication in a free-ranging primate. *Anim. Behav.* 28, 1070–1094. doi: 10.1016/S0003-3472(80)80097-2

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Snyder, B. (2000). *Music and Memory. An Introduction*. Cambridge, Massachusetts: MIT Press.

Soltis, J., Leon, K., and Savage, A. (2005). African elephant vocal communication I: antiphonal calling behaviour among affiliated females. *Anim. Behav.* 70, 579–587. doi: 10.1016/j.anbehav.2004.11.015

Sperber, D., and Wilson, D. (1986). *Relevance: Communication and Cognition*. Oxford, UK & Cambridge, MA: Blackwell.

Stegmann, U. (2013). *Animal Communication Theory: Information and Influence*. Cambridge, UK: Cambridge University Press.

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

Thiessen, E. D., Hill, E. A., and Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy* 7, 53–71. doi: 10.1207/s15327078in0701_5

Thompson, W. F., Sun, Y., and Fritz, T. (2019). "Music across cultures," in *Foundations in Music Psychology*. eds. P. J. Rentfrow and D. J. Levitin (Cambridge, Massachusetts: MIT Press).

Tierney, A., Patel, A. D., and Breen, M. (2018). Acoustic foundations of the speech-to-song illusion. *J. Exp. Psychol. Gen.* 147, 888–904. doi: 10.1037/xge0000455

Tomasello, M. (2010). *Origins of Human Communication*. Cambridge, Massachusetts: MIT Press.

Tooby, J., and Cosmides, L. (2005). "Conceptual foundations of evolutionary psychology," in *The Handbook of Evolutionary Psychology. 1st Edn*. ed. D. M. Buss (Hoboken, New Jersey: John Wiley & Sons, Inc.).

Torreira, F., Bögels, S., and Levinson, S. C. (2015). Breathing for answering: the time course of response planning in conversation. *Front. Psychol.* 6:284. doi: 10.3389/fpsyg.2015.00284

Trainor, L. J. (2018). "The origins of music: auditory scene analysis, evolution, and culture in musical creation," in *The Origins of Musicality*. ed. H. Honing (Cambridge, Massachusetts: MIT Press), 81–112.

Trehub, S. E., Becker, J., and Morley, I. (2015). Cross-cultural perspectives on music and musicality. *Philos. Trans. R Soc. Lond. B Biol. Sci.* 370:20140096. doi: 10.1098/rstb.2014.0096

Venhuizen, N. J., Crocker, M. W., and Brouwer, H. (2019). Semantic entropy in language comprehension. *Entropy* 21, 1–21. doi: 10.3390/e21121159

Vuust, P., Brattico, E., Glerean, E., Seppänen, M., Pakarinen, S., Tervaniemi, M., et al. (2011). New fast mismatch negativity paradigm for determining the neural prerequisites for musical ability. *Cortex* 47, 1091–1098. doi: 10.1016/j.cortex.2011.04.026

Wacewicz, S., and Żywiczyński, P. (2015). Language evolution: why Hockett's design features are a non-starter. *Biosemiotics* 8, 29–46. doi: 10.1007/s12304-014-9203-2

Walker, G. (2010). "The phonetic constitution of a turn-holding practice: rush-throughs in English talk-in-interaction," in *Prosody in Interaction. Studies in Discourse and Grammar. Vol. 23*. eds. D. Barth-Weingarten, E. Reber and M. Selting (Amsterdam, Philadelphia: John Benjamins Publishing Company), 51–72.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *J. Acoust. Soc. Am.* 91, 1707–1717. doi: 10.1121/1.402450

Williams, P. L., and Beer, R. D. (2010). Nonnegative decomposition of multivariate information. arxiv [Preprint], 1–14.

Williams, P. L., and Beer, R. D. (2011). Generalized measures of information transfer. arxiv [Preprint], 1–6.

Winkler, I., Denham, S. L., and Nelken, I. (2009). Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends Cogn. Sci.* 13, 532–540. doi: 10.1016/j.tics.2009.09.003

Wohltjen, S., and Wheatley, T. (2021). Eye contact marks the rise and fall of shared attention in conversation. *PNAS* 118:e2106645118. doi: 10.1073/pnas.2106645118

Zatorre, R. J. (2018). Why Do We Love Music ? *Cerebrum: The Dana Forum on Brain Science.* 1–12. Available at: http://www.dana.org/Cerebrum/2018/Why_Do_We_Love_Music (Accessed March 01, 2022).

Zbili, M., and Rama, S. (2021). A quick and easy way to estimate entropy and mutual information for neuroscience. *Front. Neuroinform.* 15:596443. doi: 10.3389/fninf.2021.596443

Zuidema, W., Hupkes, D., Scharff, C., and Rohrmeier, M. (2018). "Formal models of structure building in music, language, and animal song," in *The Origins of Musicality*. ed. H. Honing (Cambridge, Massachusetts; London, England: The MIT Press), 253–286.

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 10 June 2022
doi: 10.3389/fnins.2022.860127

The Effect of Vocal Intonation Therapy on Vocal Dysfunction in Patients With Cervical Spinal Cord Injury: A Randomized Control Trial

Xiaoying Zhang[1,2,3,4,5†], Yi-Chuan Song[1,5], De-Gang Yang[1,7], Hong-Wei Liu[1,7], Song-Huai Liu[1,5], Xiao-Bing Li[6] and Jian-Jun Li[1,2,3,4*†]*

[1] School of Rehabilitation Medicine, Capital Medical University, Beijing, China, [2] China Rehabilitation Science Institute, Beijing, China, [3] Beijing Key Laboratory of Neural Injury and Rehabilitation, Beijing, China, [4] Center of Neural Injury and Repair, Beijing Institute for Brain Disorders, Beijing, China, [5] Music Therapy Center, China Rehabilitation Research Center, Beijing, China, [6] Laboratory of Music Artificial Intelligence, Central Conservatory of Music, Beijing, China, [7] Department of Spinal and Neural Functional Reconstruction, China Rehabilitation Research Center, Beijing, China

OPEN ACCESS

Edited by:
*Mary Miu Yee Waye,
The Chinese University of Hong Kong,
Hong Kong SAR, China*

Reviewed by:
*Jeanette Tamplin,
University of Melbourne, Australia
Claire Slinger,
Lancashire Teaching Hospitals NHS
Foundation Trust, United Kingdom*

***Correspondence:**
*Xiao-Bing Li
lixb@163.com
Jian-Jun Li
crrclijj@163.com*

†ORCID:
*Xiaoying Zhang
orcid.org/0000-0002-6329-2958
Jian-Jun Li
orcid.org/0000-0002-8441-7537*

Specialty section:
*This article was submitted to
Auditory Cognitive Neuroscience,
a section of the journal
Frontiers in Neuroscience*

Received: *22 January 2022*
Accepted: *10 May 2022*
Published: *10 June 2022*

Citation:
Zhang X, Song Y-C, Yang D-G, Liu H-W, Liu S-H, Li X-B and Li J-J (2022) The Effect of Vocal Intonation Therapy on Vocal Dysfunction in Patients With Cervical Spinal Cord Injury: A Randomized Control Trial. Front. Neurosci. 16:860127. doi: 10.3389/fnins.2022.860127

In this study, the vocal intonation therapy (VIT) was compared with the standard respiratory therapy for people suffering from respiratory dysfunction as a result of cervical spinal cord injury (CSCI) to observe its effect on vocal quality. Thirty patients with vocal dysfunction after CSCI with the injury time of more than 3 months were screened for inclusion in the trial, and 18 patients completed the 12-weeks, each participant had 60 sessions in total in the clinical trial. All patients were allocated to the intervention group or the control group. The intervention group received VIT training and the control group received respiratory phonation therapy. Both groups were trained by professional therapists, and the training time was 30 min/day, 5 days/week, for 60 sessions for each group in a total of 12 weeks. In the Baseline (T0), mid-intervention period (after 6 weeks, T1), and after intervention (after 12 weeks, T2), the vocal quality of the two groups of patients was tested with a computer-aided real-time audio analyzer 2.1.6 (Adobe Systems, United States) for Sing-SPL ($p < 0.0001$), Speech-SPL ($p < 0.0001$), SNL ($p < 0.0001$), and F0 ($p < 0.0001$) of the intervention group were significantly improved compared with the control group. In comparing the spectrometry analysis of vocal quality for the 2 groups of participants, there was a significant difference in the results of Sing-SPL and Speech-SPL acoustic analysis in the intervention group of patients at T2 (after 12 weeks) compared to the control group. Vocal intonation therapy—music therapy can improve the speech sound quality of cervical CSCI patients and provide CSCI patients with a practical, highly operable treatment that has both functional training effects and can bring a pleasant experience that can be promoted in the medical field. This study was approved by the Ethics Committee of China Rehabilitation Research Center (CRRC) (approval No. 2019-83-1) on May 20th, 2019. It was registered with the National Health Security Information Platform, medical research registration, and filing information system (Registration No. MR-11-21-011802) on January 28th, 2021.

Keywords: cervical spinal cord injury, vocal intonation therapy, vocal dysfunction, music therapy, vocal quality

INTRODUCTION

Cervical spinal cord injury (CSCI) is a critical injury that often entails disability. Dyspnea may occur in patients with upper CSCI or quadriplegia if the diaphragm and intercostal muscles are paralyzed, often resulting in severe restrictive ventilatory impairment or medical social burden (Koda et al., 2021). Vocal dysfunction due to respiratory disorders following CSCI often presents difficulties in articulation and vocal endurance. Therefore, CSCI often causes a negative effect on the quality of patients' vocalization (Williams et al., 2020). Presently, the prevalence of respiratory dysfunction caused by CSCI and the positive correlation between it and voice output has been proposed (Mesbah et al., 2021); however, most studies have focused on the dysfunction of the respiratory system caused by CSCI (Lemos et al., 2020), and not many studies have been conducted on the vocalization function and voice quality after CSCI.

After CSCI, the motor function of the phrenic nerve innervation is affected, respiratory function is impaired, and pronunciation quality is also affected to a certain extent. The most important manifestations are patients' decreased volume in daily communication, the difficulty in adhering to the oral expression of long sentences, and the increase of inspiratory time (Wadsworth et al., 2012). In recent years, some studies have explored breathing and voice function after CSCI and found that respiratory dysfunction after CSCI usually results in overcompensation of lung function to deal with decreases in expiratory muscle compliance and increase speech loudness for conversation (Lu et al., 2008; Tamplin et al., 2011; MacBean et al., 2017). Some studies have demonstrated that healthy individuals only require 20% of their respiratory lung capacity for natural volume pronunciation or vocalization (Tamplin et al., 2013). In comparison, patients with respiratory dysfunction after CSCI can use 30–50% of their lung capacity depending on the extent of the injury (Hixon et al., 1973). In conclusion, it is common for CSCI patients to overcompensate for lung function to ensure the normal speech volume. Moreover, during the emergency medical treatment of CSCI patients, laryngeal dysfunction caused by intubation and tracheotomy can also lead to a mild vocal cord movement disorder, and in severe cases, vocal cord polyps, vocal nodules, or complete vocal cord paralysis (Watson and Hixon, 2001).

Some studies conducted auditory perception assessment and analysis on patients with vocalization dysfunction after CSCI and concluded that their speech features were of a reduced volume, mostly short sentences, increased inhalation time (Gadomski et al., 2021; OSCIS investigators et al., 2021), and deviations in articulation, articulation accuracy, and speech quality (Tamplin et al., 2014; Scheuren et al., 2021). Although most CSCI patients can maintain the volume required for a normal conversation in a quiet room, patients may have difficulty raising the volume in the presence of other noises (Berlowitz et al., 2016). These decreases in volume and length are directly caused by impaired respiratory function (Clini et al., 2018). Studies have also concluded that CSCI patients can significantly feel more drastic volume problems than healthy controls (Oraee et al., 2021). Some scholars have used singing to perform breathing training for quadriplegic patients, and found that singing activates accessory respiratory muscles (as measured by surface electromyography) and stimulates greater respiratory function than speaking (Tamplin et al., 2011). A follow up study in this area with 24 patients, found improvements in speech loudness (SPL) and maximum phonation length, supported by improved respiratory function following a therapeutic singing intervention (Tamplin et al., 2013). The Perceptual Voice Profile (PVP) and Voice Handicap Index (VHI), both are subjective rating scale, are used to assess vocal function scores (Crispiatico et al., 2021). The results showed that CSCI patients had low voice quality, insufficient volume in long space, insufficient breath during speech, and common social disorder problems of varying degrees in social situations.

Since singing and speaking share the same neural network, in recent years, an increasing number of studies have focused on the treatment of speech abnormalities caused by functional disorders of the nervous system (Jasmin et al., 2016). Vocal vocalization therapy (VIT) (Thaut and Hoemberg, 2014) uses vocal music training to practice vocal function control problems caused by structural, neurological, physiological, psychological, or functional abnormalities of vocal organs (Strohl et al., 2020). VIT directly stimulates the muscles associated with breathing, vocalization, articulation, and resonance and requires more sound control (Scheuren et al., 2021) and intensity (Rodrigues et al., 2021) than speech. Furthermore, VIT training can enhance respiratory muscle strength (Tamplin et al., 2014; Zhang et al., 2021). VIT training used in clinical practice resembles open voice exercises used by vocal teachers or choral conductors, addressing issues related to vocal function control, such as breath control, tone, pitch, timbre, and strength issues. VIT starts with a melody and uses various vowel and consonant vocalization combinations to provide practical and effective non-invasive treatment for patients with vocal dysfunction after spinal cord injury. Under the guidance of the music therapist, patients gradually adjust the vocal apparatus during the vocalization process and then further combine with the complete singing, conducive to improving the vocalization function and voice quality. In this study, VIT was used to intervene in patients with vocal dysfunction after cervical spinal cord injury and compared patients who received routine breathing pronunciation training. This process gauged whether VIT can achieve faster and more effective functional recovery than physical therapy to find a faster and effective treatment for patients with CSCI.

PARTICIPANTS AND METHODS

This study was approved by the Ethics Committee of China Rehabilitation Research Center (CRRC) (approval No. 2019-83-1) on May 20th, 2019, and informed consent was obtained from the participants, relatives, or guardians before commencing the study. The study trial was registered with the National Health Security Information Platform, medical research registration, and filing information system (Registration No. MR-11-21-011802) on January 28th, 2021.

Participants

Eighteen patients with spinal cord injury, including complete spinal cord injury and incomplete spinal cord injury, who was hospitalized in CRRC from January 2021 to December 2021, were recruited. Inclusion criteria: (1) classified as spinal cord injury class A and class B by the American spinal cord injury association (ASIA) (American Spinal Cord Injury Association [ASIA]); (2) The course of the disease should have been at least 3 months (inclusive); (3) Aged between 18 and 75; (4) Moderate or above respiratory dysfunction and phonation disorder after cervical spinal cord injury, voice volume in normal conversation is lower than 40 decibels; (5) no tracheotomy or tracheotomy has healed; (6) Can tolerate seat training for more than 15 min at a time without postural hypotension; (7) No previous music learning experience; (8) native Mandarin Chinese; (9) Patients and their families were informed and consented to this study. Exclusion criteria: (1) a history of severe speech disorder, a history of mental disorder, or a history of severe respiratory disease before injury; (2) Severe cognitive dysfunction, MMSE score < 17 (illiteracy) or < 20 (primary school); (3) tracheotomy, vocal cord damage, or posterior damage; (4) epilepsy, malignant arrhythmia, or other serious physical diseases. Criteria for withdrawal and termination: The study can be promptly terminated if the patient's condition changes or discharged or voluntarily quit the study. The general information of patients is as follows (**Table 1**).

Research Design

Since this study was a randomized controlled trial, n = Z2·σ2/d2 was calculated according to the sample size formula, where n was the minimum sample size. Z is the confidence interval, usually 90%; Sigma is the standard deviation; D is the sampling error range; usually, 0.5 and the minimum sample size was thus calculated. Therefore, this study met the requirement of a minimum sample size (n $\approx$ 18) and was divided into two groups. Computer-generated sequences (Excel 2013, US, Washington, Seattle, Microsoft Office) randomly divided patients into two groups: vocal intonation therapy group (n = 9) and respiratory phonation training group (n = 9). This study was conducted from January 2021 in CRRC.

TABLE 1 | Participants' characteristics in this study.

	Intervention group	Control group	*T*	*p*
Total number	9	9		>0.05
Gender				
Male	7	8		>0.05
Female	2	1		>0.05
Age	38.60 ± 17.89	34.78 ± 11.13	0.6186	>0.05
Months since injury	4.20 ± 4.0	5.64 ± 4.08	0.0109	>0.05
Height (cm)	172.00 ± 11.00	166.10 ± 9.00	1.25	>0.05
Weight (kg)	63.20 ± 9.90	63.90 ± 17.92	0.9919	>0.05
BMI	21.37 ± 3.03	22.91 ± 5.24	0.0223	>0.05
AISA classification				
ASIA A	6	5		>0.05
ASIA B	3	4		>0.05
Injury level grading				
C4	5	6		>0.05
C5	4	3		>0.05

Data were expressed as the number in total number, gender, age, months since injury (time), height (cm), weight (kg), body mass index (BMI), AISA classification, and injury level grading. Other data were expressed as the mean ± SD and analyzed by paired t-test. Intervention group: vocal intonation therapy (VIT) group; Control group: respiratory phonation training. ASIA, American Spinal Injury Association.

Procedure

After gaining the Scientific Research Foundation of CRRC's approval, participants diagnosed with ASIA (National Spinal Cord Injury Statistical Center [NSCISC], 2017) classification A and B with C4, C5 injuries, combined with vocal disorders, were firstly screened by clinical medical experts such as spinal cord injury specialists, spinal surgeons, and rehabilitation specialists according to the study inclusion criteria, applied for consultation and referred to the music therapy center. The music therapy researchers conducted the initial assessment and test of the vocal function of the referred patients, and the patients meeting the inclusion criteria will be included in the study to be randomly assigned, and those who fail to meet the criteria will be excluded. All patients and their families were informed and consented to this study. Participants included in the study were randomly divided into the intervention group as the vocal intonation therapy group (n = 9) and control group, respiratory phonation group (n = 9) according to the computer-generated sequence based on routine inpatient rehabilitation therapy. Patients in the intervention group were given vocal intonation therapy (VIT) by music therapists. The training process was 30 min of one-on-one VIT conducted by music therapists, 5 times per week for 12 weeks, 60 sessions in total. The control group was given respiratory phonation therapy by respiratory physiotherapists, 30 min a day, 5 times a week, for 12 weeks, 60 sessions in total. The enrollment and allocation of participants are shown in **Figure 1**.

Figure 1 illustrates that 30 participants were enrolled in the study, 12 participants withdrew from the study as they failed to meet the inclusion criteria (n = 10) dropped off the trial (n = 1); private reasons (n = 1). The intervention group was treated with VIT (total n = 9, ASIA A n = 6, ASIA B n = 3; C4 n = 5, C5 = 4); the control group underwent respiratory phonation training (total n = 9, ASIA A n = 5, ASIA B n = 4; C4 n = 6, C5 = 3). Among the 9 participants in the intervention group, all the participants were able to sing the trained items during the entirety of the study period. The data analysis included a sample of 18 CSCI patients. ASIA: American Spinal Injury Association; VIT: vocal intonation training; CSCI: cervical spinal cord injury.

Interventions

Patients in the control group were given respiratory phonation training (RPT). RPT is a vocal training method whereby respiratory physiotherapists use physical therapy techniques to press the abdominal cavity with external force after the patient inhales to help them produce long sounds (Katz et al., 2022). Patients in the intervention group were treated by music therapy professionals using VIT techniques. The biggest difference

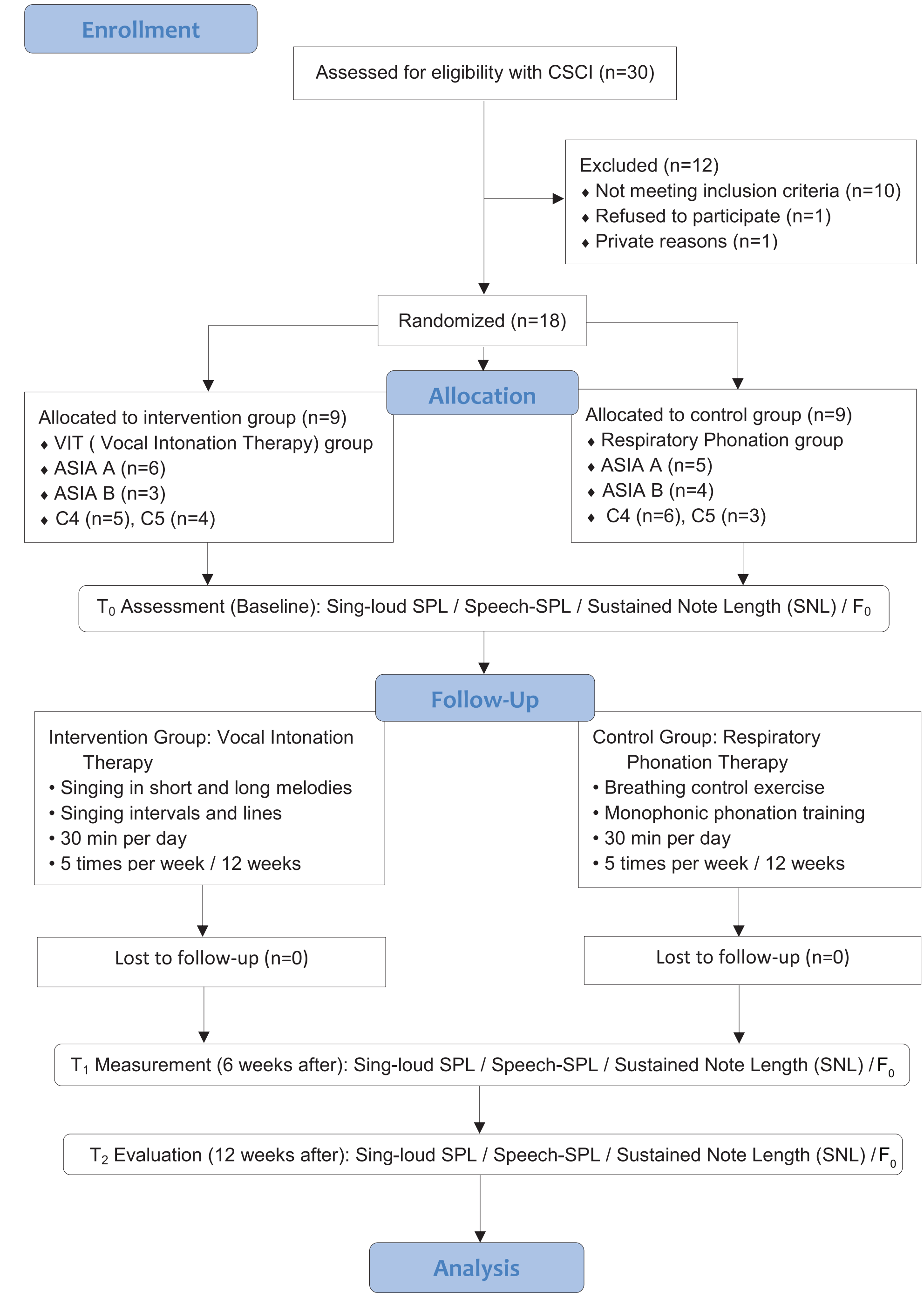

Enrollment
Assessed for eligibility with CSCI (n=30)
Excluded (n=12)
♦ Not meeting inclusion criteria (n=10)
♦ Refused to participate (n=1)
♦ Private reasons (n=1)
Randomized (n=18)
Allocation
Allocated to intervention group (n=9)
♦ VIT (Vocal Intonation Therapy) group
♦ ASIA A (n=6)
♦ ASIA B (n=3)
♦ C4 (n=5), C5 (n=4)
Allocated to control group (n=9)
♦ Respiratory Phonation group
♦ ASIA A (n=5)
♦ ASIA B (n=4)
♦ C4 (n=6), C5 (n=3)
T0 Assessment (Baseline): Sing-loud SPL / Speech-SPL / Sustained Note Length (SNL) / F0
Follow-Up
Intervention Group: Vocal Intonation Therapy
• Singing in short and long melodies
• Singing intervals and lines
• 30 min per day
• 5 times per week / 12 weeks
Control Group: Respiratory Phonation Therapy
• Breathing control exercise
• Monophonic phonation training
• 30 min per day
• 5 times per week / 12 weeks
Lost to follow-up (n=0)
Lost to follow-up (n=0)
T1 Measurement (6 weeks after): Sing-loud SPL / Speech-SPL / Sustained Note Length (SNL) / F0
T2 Evaluation (12 weeks after): Sing-loud SPL / Speech-SPL / Sustained Note Length (SNL) / F0
Analysis

FIGURE 1 | Flow diagram, consort flowchart for participants' recruitment and allocation.

between RPT and VIT is that RPT uses external force to press the patient's abdominal cavity to guide the vocalization; while VIT is a method whereby patients make their voices autonomous under the guidance of music. In the intervention group, as well as the music therapy group, first, a vocalization assessment was performed, whereby the therapist determines the location of the difficulty based on the articulation of the patient's voice. Following this, the preparatory exercises were then started. The specific intervention steps are: (1) Vocal warming. The therapist instructs the patient to practice monophonic vocalations starting at Andante speed (approximately ♩ = 72), using the combined inhalation pattern of chest and abdomen, drawing respiratory to the waist and abdominal diaphragm. (2) Vocal intonation. After repeated repetitions in the previous step, the therapist instructs the patient to practice monophonic vocalizations after a quick inhalation. For example, "mi-ma-mi," "li-lu-li" and other melody lines. The therapist uses 4/4 to complete the training of guiding the patient. During the guidance process, the patient's vocal quality was always scrutinized. This step is repeated 20 times with different combinations of vowels, consonants, and melodies in different tones. As the final step of the intervention, therapist used a song to reinforce the previous VIT practice session (**Figure 2**).

(3) Rhythmic induced intonation. It starts with an intensive rhythm of moderate (♩ = 96), switches between word and tone combinations, exercises with a faster rhythm, and improves diaphragm jumping. Finally, the lyrics are read aloud with the volume of natural speech to normalize the vocalization function of oral communication (**Figure 3**). Patients in the control group were trained by respiratory physiotherapists in the bedside posture.

Measurements

The intervention group was trained with VIT, and the control group was trained with respiratory phonation therapy at doses as previously described. The other clinical indications remained identical in both groups. Before intervention (baseline, T0), during therapy (mid-test, 6 weeks later, T1), and after therapy (evaluation, 12 weeks later, T2), using computer-aided real-time audio analyzer 2.1.6 (Adobe Systems, United States) (Reimers and Stewart, 2016) for Sing-loud Pressure Level (Sing-SPL), Speak Pressure Level (Speech-SPL), Sustained note Length (SNL), and Fundamental Frequency (F0) assessment. The measurement results in the three-time points are analyzed and compared.

Sing-Loud Pressure Level (Sing-SPL, Decibels)

Musical sound refers to the characteristics of hearing that can distinguish the height of musical sound, determined by the frequency of sound wave vibration, with high and low frequencies. The ratio of the effective value p(e) of the sound pressure to be measured to the reference sound pressure p(ref) is commonly used logarithmic and calculated in decibels (dBA) (Schloneger and Hunter, 2017).

Speech-Loud Pressure Level (Speech-SPL, Decibels)

Speech-SPL (Behrman et al., 2020) measures the effective volume of sound relative to a benchmark value with a standard sound pressure of p0 ms. Speech is the sound flow formed by the human voice language, generally expressed in decibels (dBA).

Sustained Note Length (SNL, s)

Musical terms are time values, and in music are the relative duration between musical tones in seconds (s) or milliseconds (ms) (Bruschettini et al., 2020).

Sound Frequency (F0, Hz)

In the category of the human voice, it is called the fundamental frequency, and the fundamental frequency (F0, Hz) refers to the basic sound source produced by the vibration of the vocal cords (Witold, 2020). The resulting sound unit is hertz (Hz), which is also commonly used as kilowatt-hertz (kHz) and megahertz (MHz).

Statistical Analysis

According to the International Standard Organization (ISO), the U.S. Environmental Protection of America (EPA) (Geneid et al., 2020), the Chinese National Acoustic Environmental Quality Standard (GB 3096-2008) (Molina et al., 2021), and the Social Life Environmental Noise Standard (GB 22337-2008), office ambient noise for sound acquisition are controlled to a control standard of less than 30 dBA (Molina et al., 2021). Data was collected from both sets using a computer voice test system at three-time points before the intervention (T0, baseline), 6 weeks after (T1), and 12 weeks (T2). Formulas were used to calculate the standard deviation of the mean and normal distribution for each group, and two-way ANOVA was used to analyze differences between groups, time effects, and differences in time interactions between groups. Data from 18 patients with vocal dysfunction following CSCI were analyzed using SPSS 22.0 (SPSS Statistical Software Inc., Chicago, Illinois, United States) who completed this part of the study. Before analysis, discrete data intervals were analyzed to screen for missing and outlier values, ensure the accuracy of data entry, and determine the specific effects of interventions. Sensitive and private information was blurred out, and all statistical data were kept confidential.

RESULTS

Comparison of the Overall Results of the Vocal Function Test in Two Groups

Data on vocal function were collected at three-time points, before the intervention (T0, baseline), 6 weeks after (T1), and after 12 weeks (T2). These include Sing-SPL (dBA), Speech-SPL (dBA), SNL (s), and sound frequency (F0). Multivariate ANOVA was used to analyze two groups of patients T0, T1, and T2 data. The results showed that in the vocal function test, the Sing-SPL of the intervention group at the T2 time point (T2 = 54.33 ± 11.30, $p < 0.0001$), Speech-SPL (T2 = 47.83 ± 11.30, $p = 0.0029$), SNL (T2 = 11.19 ± 3.25, $p < 0.0001$), sound frequency (F0, T2 = 305.89 ± 80.39, $p < 0.0001$) possessed significant differences between groups compared with the control group. (In the analysis and comparison of the Sing-SPL, Speech-SPL, SNL, F0, the intervention group patients had a very significant inter-group difference compared with the control group and had a significant difference in time effect). The results of the statistical analysis of the two sets of data are as follows (see **Table 2** and **Figure 4**).

The Results Analysis of Sing-Loud Pressure Level in Two Groups

In the analysis of the results of the vocal pressure level (Sing-SPL, dBA) of patients in the intervention group and the control group, no significant differences were noted in the Sing-SPL values of the two groups of patients in the baseline test at the T0 point in time, which showed that the distribution of the two groups of patients was even and met the requirements of random control. In the T1 point-in-time (after 6 weeks) test, the Sing-SPL values of the two groups of patients began to differ, and by the T2 point-in-time (after 12 weeks), significant differences have been noted in

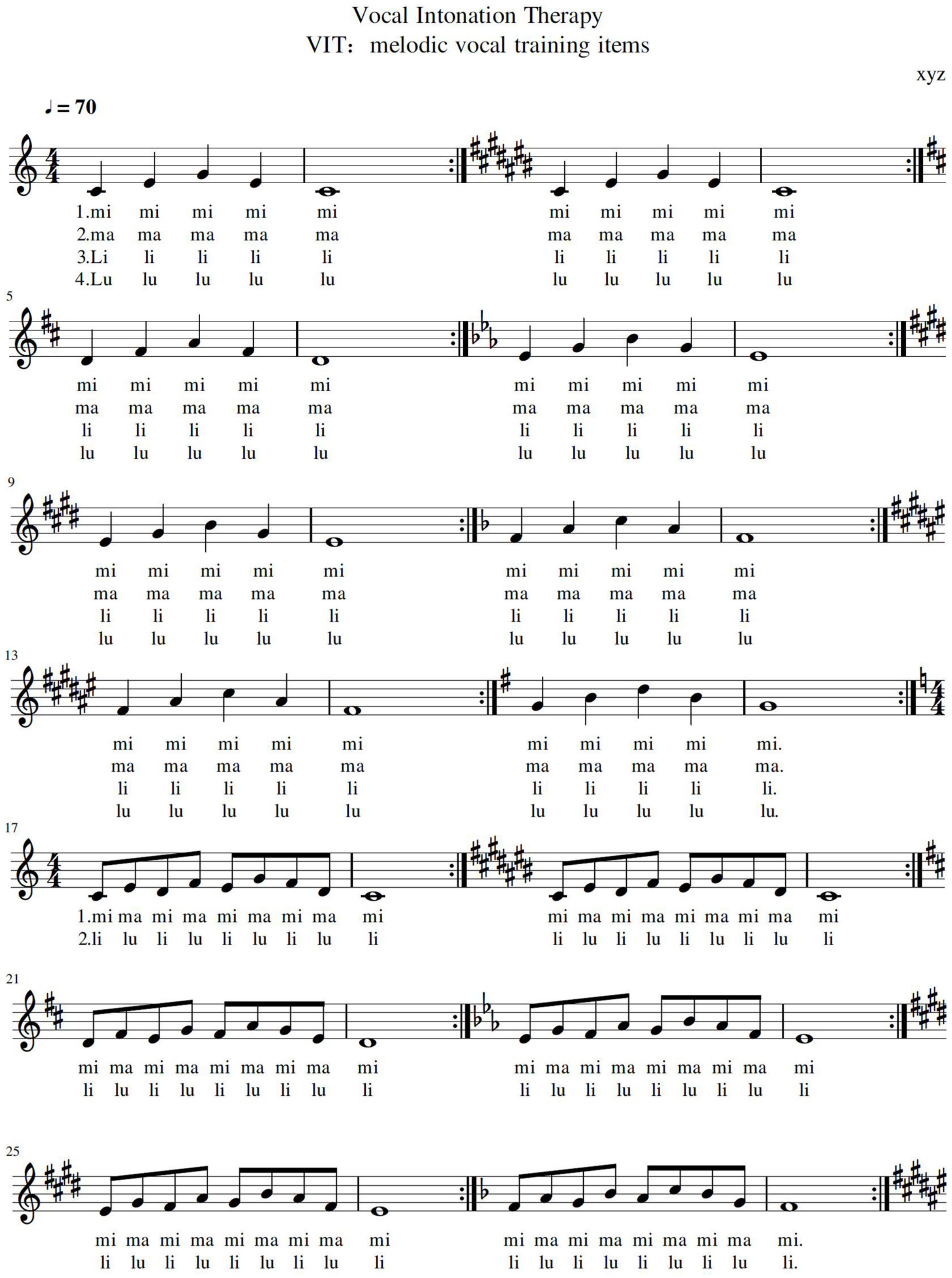
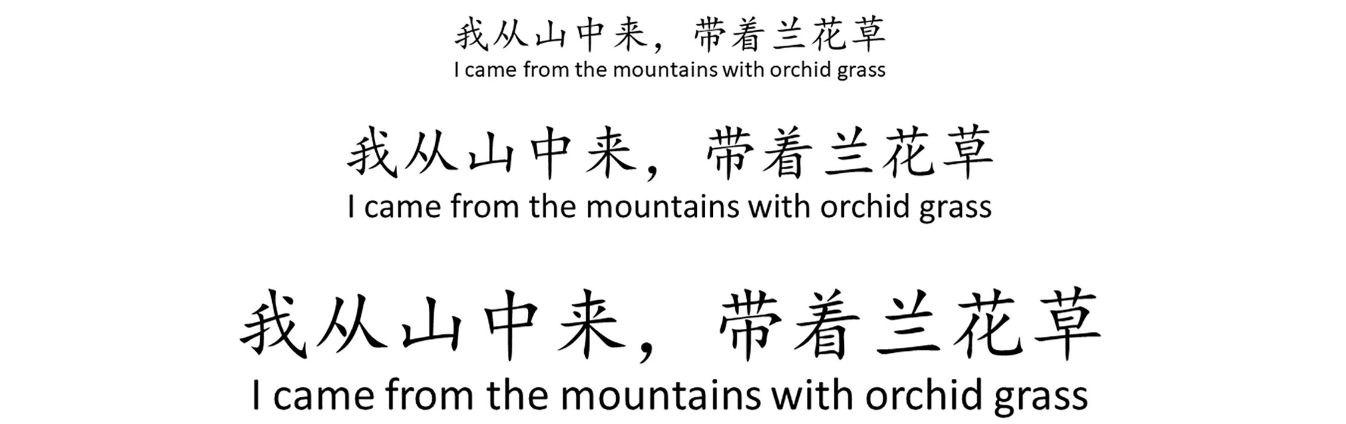

Vocal Intonation Therapy
VIT：melodic vocal training items
xyz
♩= 70
1.mi mi mi mi mi
2.ma ma ma ma ma
3.Li li li li li
4.Lu lu lu lu lu
5
9
13
mi. ma. li. lu.
17
1.mi ma mi ma mi ma mi ma mi
2.li lu li lu li lu li lu li
21
25
mi ma mi ma mi ma mi ma mi.
li lu li lu li lu li lu li.

FIGURE 2 | VIT items with different keys, from C to G.



我从山中来，带着兰花草
I came from the mountains with orchid grass

我从山中来，带着兰花草
I came from the mountains with orchid grass

我从山中来，带着兰花草
I came from the mountains with orchid grass

FIGURE 3 | Lyric reading voice volume prompt diagram. The larger the font size is, the larger the volume required. Line 1, font number 3, low volume, about 20–30 dBA; Line 2, small 2 font, moderate volume, about 31–40 dBA; Line 3, font number 2, large volume, about 41–55 dBA.

TABLE 2 | Vocal quality results in CSCI patients across the study period for the intervention and control group.

		Intervention group (*n* = 9)	Control group (*n* = 9)	t	P
		Mean ± SD	Mean ± SD		
Sing-SPL (dBA)	t_0	21.50 ± 5.11	17.00 ± 2.40	1.226	0.0697
	t_1	41.44 ± 6.48	27.00 ± 6.20	3.935	0.0001^{**a}
	t_2	54.33 ± 11.30	38.56 ± 11.12	4.297	0.0001^{**b}
Speech-SPL (dBA)	t_0	20.61 ± 5.92	17.11 ± 3.90	0.8819	0.2426
	t_1	37.11 ± 11.54	27.44 ± 8.02	2.437	0.0004^{**a}
	t_2	47.83 ± 11.30	34.89 ± 7.06	3.26	0.0001^{**b}
SNL (s)	t_0	5.03 ± 1.43	4.29 ± 0.78	0.94	0.1429
	t_1	8.16 ± 1.40	6.39 ± 0.41	1.486	0.0001^{**a}
	t_2	11.19 ± 3.25	8.33 ± 1.19	3.633	0.0001^{**b}
F0 (Hz)	t_0	82.33 ± 20.22	58.44 ± 15.56	1.077	0.05
	t_1	155.89 ± 44.51	121.89 ± 42.14	1.531	0.0001^{**a}
	t_2	305.89 ± 80.39	208.89 ± 49.09	4.369	0.0002^{**b}

*Intervention group: vocal intonation therapy group; Control group: respiratory phonation group. Data were expressed as mean ± SD (n = 9), and analyzed by repeated measures analysis of variance. **P < 0.01. Superscript a represents difference factor at the same time between groups, and superscript b represents difference effect of time factor in inter-group. Sing-SPL, sing-loud sound pressure level; Speech-SPL, speech-loud sound pressure level; SNL, sustained note length; F0, sound frequency.*

the Sing-SPL results between the two groups. The distribution of individual values of Sing-SPL values in the two groups of patients is compared below (**Figure 5**).

The Results Analysis of Speech-SPL in Two Groups

In the analysis of the Speech-SPL (dBA) results of the intervention group and the control group, no significant difference was noted in the Speech-SPL value of the two groups at the baseline test at T0 time point, indicating that the distribution of the two groups of patients was uniform and met the requirement of randomized control. At the T1 time point (6 weeks later), sing-SPL values began to differ between the two groups, and at the T2 time point (12 weeks later), Speech-SPL results showed significant differences between the two groups. The distribution of individual values of Speech-SPL values in the two groups was compared as follows (**Figure 6**).

Analysis of Sustained Note Length (s) Results

In the analysis of the results of Sustained Note Length (SNL) of the intervention group and the control group, no significant difference was noted in the SNL value of the two groups in the baseline test at T0 time point, indicating that the distribution of the two groups of patients is uniform and meets the requirements of randomized control. At T1 (6 weeks later), SNL values began to differ between the two groups, and at T2 (12 weeks later), SNL results showed significant differences between the two groups. The individual value distribution of SNL values in the two groups was compared as follows (**Figure 7**).

Fundamental Frequency (F0, Hz) Result Analysis

In the analysis of the results of the sound frequency (F0, Hz) of the intervention group patients and the control group, no significant difference was noted in the F0 values of the two groups of patients in the baseline test at the T0 time point, indicating that the two groups of patients were evenly distributed and met the requirements of the random control. In the test at the T1 time point (after 6 weeks), the F0 values of the two groups began to differ, and by the T2 time point (after 12 weeks), there was a significant difference in F0 results between the two groups. The individual value distributions of F0 values in the two groups of patients are compared below (**Figure 8**).

Spectrometry Analysis of Sing-Loud Pressure Level (Decibels) Between Two Sets of Vocal Quality

In the analysis of Sing-SPL results in the intervention group patients and control group patients, it was found that there was no significant difference in Sing-SPL values between the two

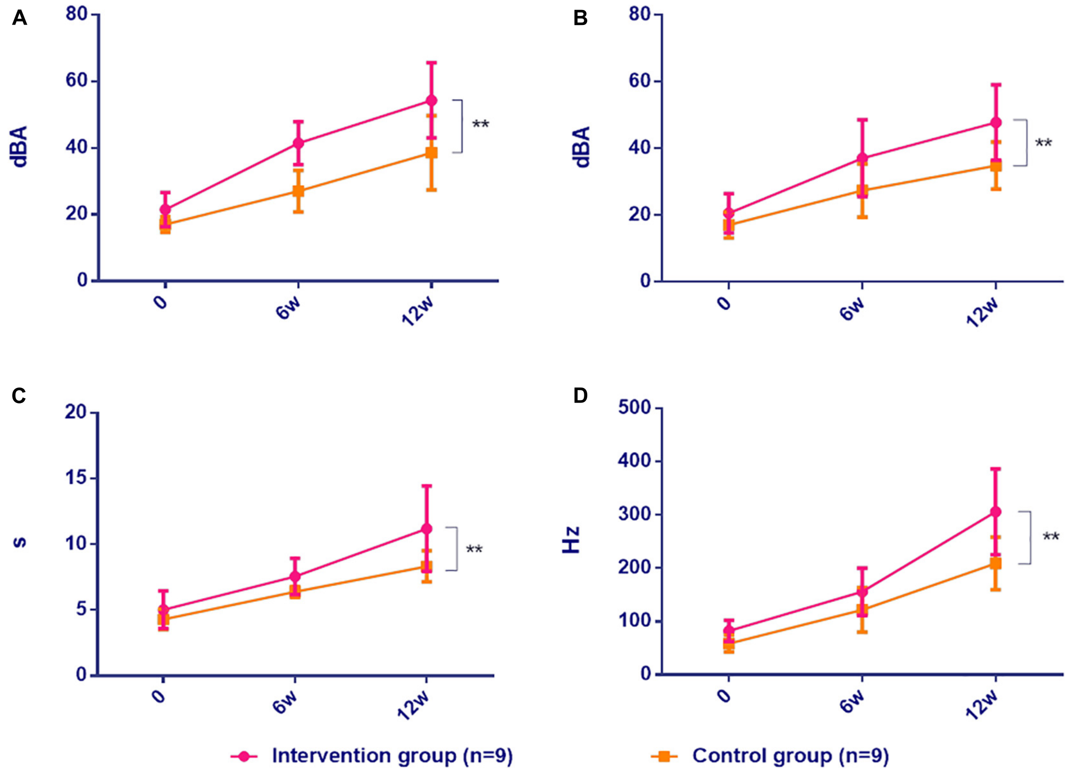


FIGURE 4 | Comparison of vocal function in CSCI patients between two groups. Intervention group: vocal intonation therapy group; Control group: respiratory photation group. **(A)** Sing-SPL, **(B)** Speech-SPL, **(C)** SNL, **(D)** F_0. Data were expressed as mean ± SD (n = 9) and analyzed by repeated-measures analysis of variance. $^*P < 0.05$, $^{**}P < 0.01$. 0, baseline; 6w, after 6 weeks; 12w, after 12 weeks. Sing-SPL, sing pressure level; Speech-SPL speaks pressure level; SNL sustained note level; F0, Fundamental Frequency.

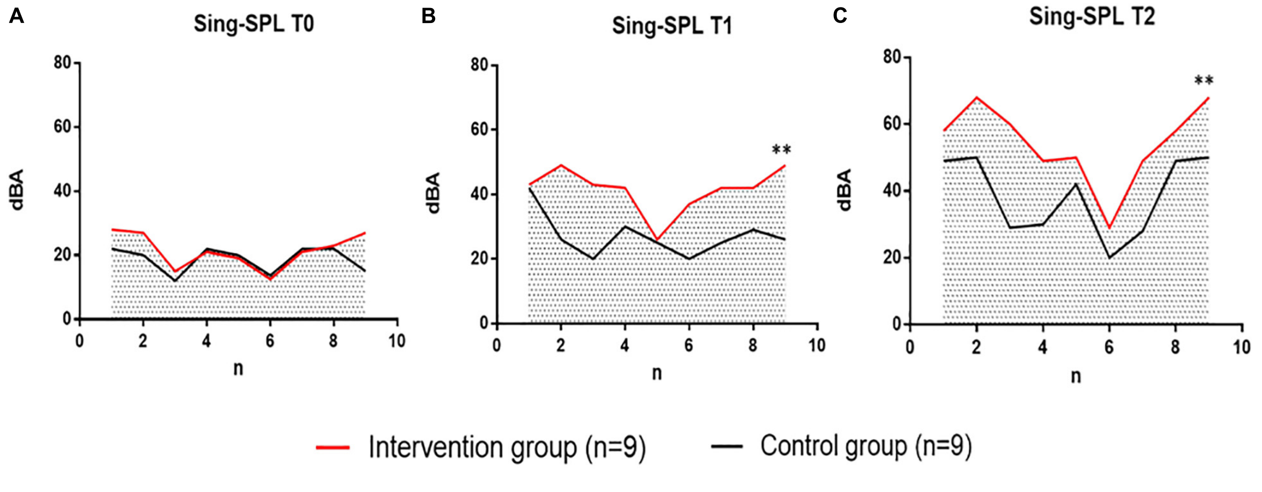


FIGURE 5 | Comparison of Sing-SPL at three time points: T0, T1, and T2 between the intervention group undergoing VIT and the control group that performed respiratory phonation training. The intervention group is indicated in red. Control group, black indicates. **(A)** Individual difference distribution of Sing-SPL values at T0 time points between patients in the intervention group (n = 9) and patients in the control group (n = 9). **(B)** Individual difference distribution of Sing-SPL values at T1 time points between patients in the intervention group (n = 9) and patients in the control group (n = 9). **(C)** Individual difference distribution of Sing-SPL values at T2 time points between patients in the intervention group (n = 9) and patients in the control group (n = 9). dBA, decibel; n, the number of participants; T0, baseline; T1, after 6 weeks; T2, after 12 weeks. $^{**}p < 0.01$, with significant differences between groups.

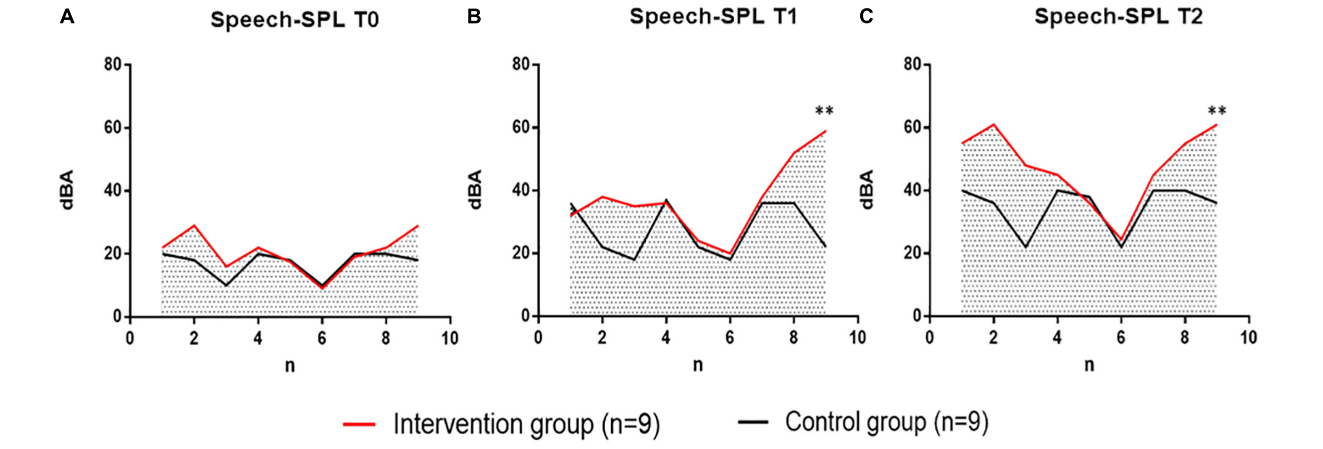


FIGURE 6 | Comparison of voice pressure levels (Speech-SPL, dBA) at T0, T1, and T2 between the intervention group receiving vocal phonation therapy (VIT) and the control group receiving breathing pronunciation training. Intervention Group, intervention group, in red. The Control group is in black. **(A)** The individual difference distribution of Speech-SPL values between patients in the intervention group ($n = 9$) and patients in the control group ($n = 9$) at T0 time point. **(B)** Individual difference distribution of Speech-SPL values at T1 time between patients in the intervention group ($n = 9$) and patients in the control group ($n = 9$). **(C)** Individual difference distribution of Speech-SPL values at T2 time points between patients in the intervention group ($n = 9$) and patients in the control group ($n = 9$). dBA, decibels; N, Number of subjects; T0, baseline; T1, 6 weeks later; T2, 12 weeks later. $^{**}P < 0.01$, the difference between groups was extremely significant.

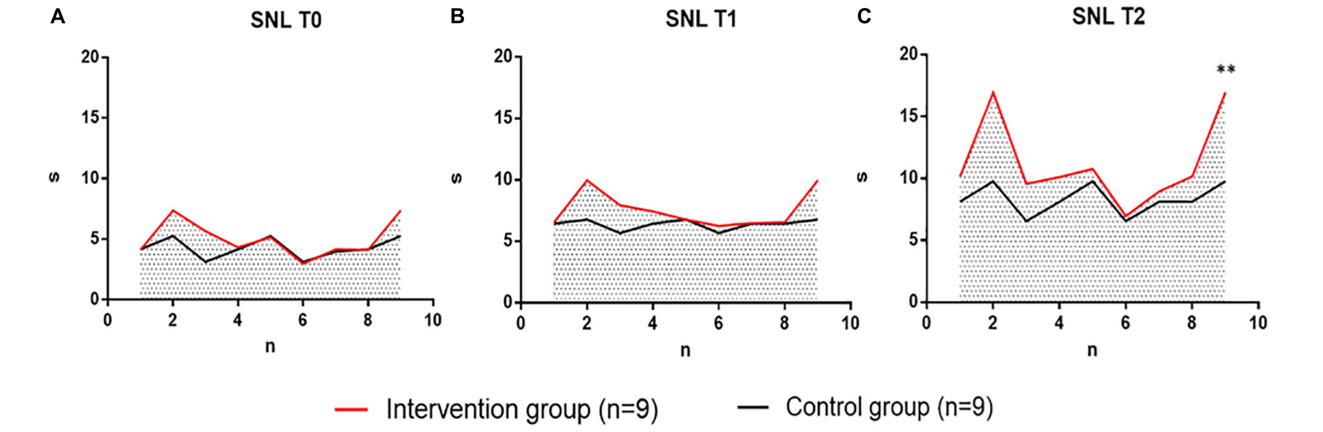


FIGURE 7 | The intervention group performing vocal therapy (VIT) compared with the control group for respiratory pronunciation training at three-time points of time (Sustained Note Length, SNL, s) at T0, T1, T2. Intervention group, intervention group, indicated in red. Control group, control group, black indicates. **(A)** Individual difference distribution of SNL values at T0 time points between patients in the intervention group ($n = 9$) and patients in the control group ($n = 9$). **(B)** Individual difference distribution of SNL values at T1 time points between patients in the intervention group ($n = 9$) and patients in the control group ($n = 9$). **(C)** Individual difference distribution of SNL values at T2 time points between patients in the intervention group ($n = 9$) and patients in the control group ($n = 9$). s, Seconds; n, the number of participants; T0, baseline; T1, after 6 weeks; T2, after 12 weeks. $^{**}p < 0.01$, with significant differences between groups.

groups of patients in the baseline test at the T0 time point, which indicated that the two groups of patients were evenly distributed and met the requirements of randomized controls. However, after the end of all interventions, at the T2 time point (after 12 weeks), there was a significant difference in the results of Sing-SPL acoustic analysis between the two groups (**Figure 9**).

Spectrometry Analysis of Two Sets of Speech-SPL (Decibels)

In the analysis of The Speed-SPL results of patients in the intervention group and the control group, it was found that there was no significant difference in the Speed-SPL values of the two groups of patients in the baseline test at the T0 time point, which indicated that the two groups of patients were evenly distributed and met the requirements of the randomized control. However, after the end of the entire intervention, at the T2 time point (after 12 weeks), there was a significant difference in the acoustic analysis results of Speech-SPL between the two groups. Due to a large amount of data, a total of 6 patients with the most obvious changes in the intervention group and the control group were taken as an example, and the individual differences in Speech-SPL of the two groups were shown (**Figure 10**).

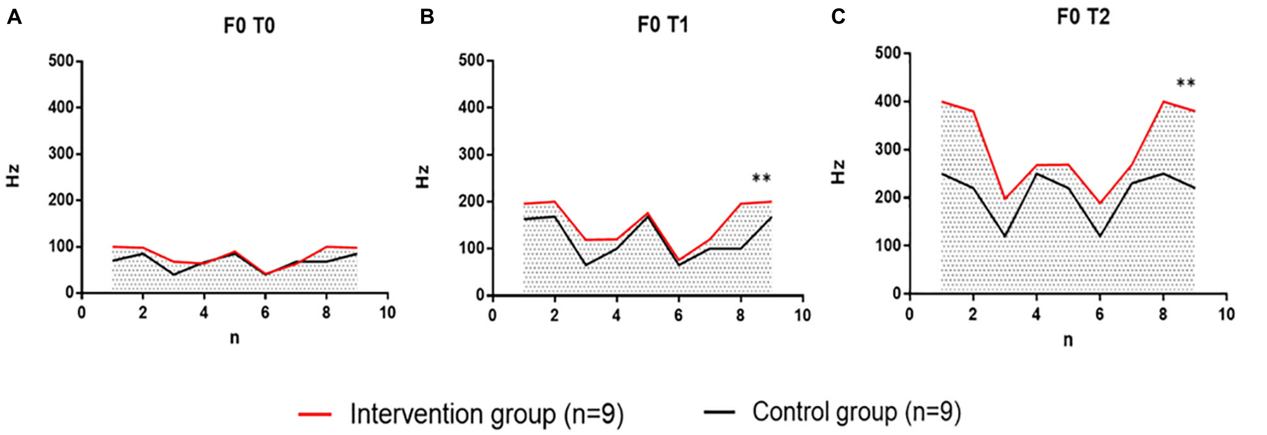


FIGURE 8 | Comparison of sound frequencies (F0, Hz) at three points in time at T0, T1, and T2 in the intervention group performing vocal therapy (VIT) with the control group for respiratory pronunciation training. Intervention group, intervention group, indicated in red. Control group, control group, black indicates. **(A)** Individual difference distribution of F0 values at T0 time point between patients in the intervention group (*n* = 9) and patients in the control group (*n* = 9). **(B)** The individual difference distribution of the F0 values of the patients in the intervention group (*n* = 9) and the control group (*n* = 9) at the T1 time point. **(C)** Individual difference distribution of F0 values at T2 time points between patients in the intervention group (*n* = 9) and patients in the control group (*n* = 9). Hz, Hertz; n, the number of participants; T0, baseline; T1, after 6 weeks; T2, after 12 weeks. ***p* < 0.01, with significant differences between groups.

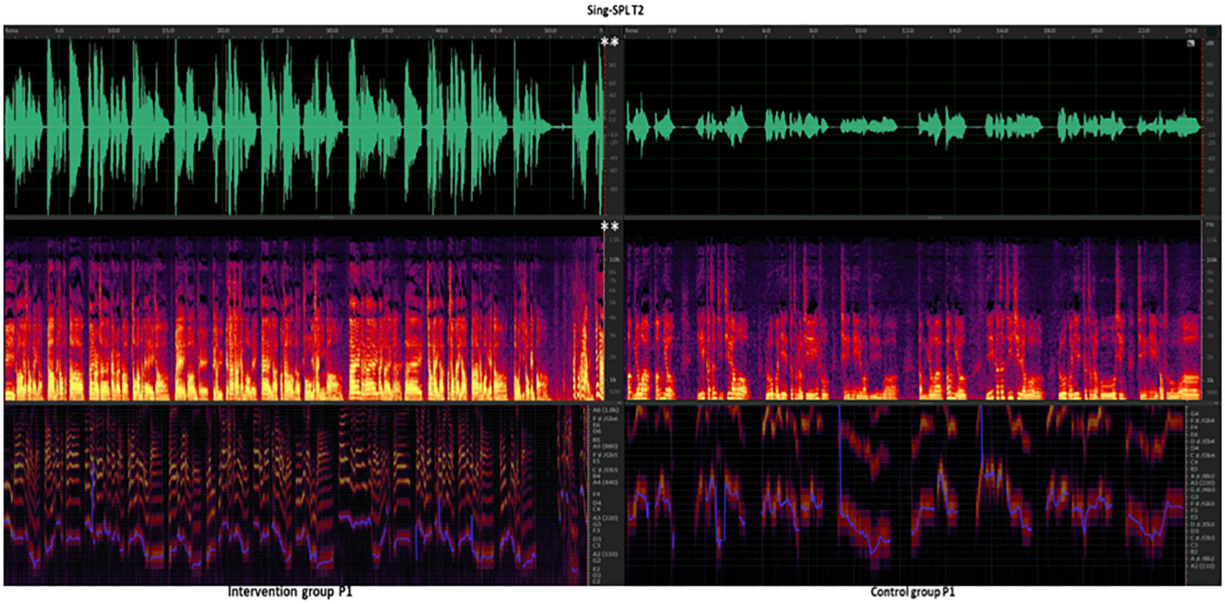


FIGURE 9 | Intervention group vs. Control group at T2 time point Sing-SPL (dBA) after the 12-week intervention. P, patient. Speech-SPL T2, intervention group patients vs. control group of Speak -SPL at T2 time point of spectral analysis. Line 1, green, the waveform spectrum represents the sound decibel (dBA) value; Line 2, purple, red, and yellow, fundamental frequency (F0, Hz) values; Line 3, purple, blue, indicates pitch (reference value). White ** indicates the *p* < 0.01, and the difference between groups is very significant.

DISCUSSION

Impaired vocal quality is a common problem in patients with decreased respiratory function after CSCI (Qu et al., 2019). The improved voice quality related to respiratory function has become the key to enhancing the quality of rehabilitation and improving quality of life and social function in patients with CSCI (Fan et al., 2016). Subglottic respiratory volume reserve to produce vocal function, maintenance of vocal cord function, control of sound intensity (dBA), and control of subglottic pressure (SPL). Moreover, increased sound pressure provides evidence of changes in loudness for the improvement of vocal function. Most of the previous studies have focused on respiration, focusing on surface electromyography of respiratory muscles, respiratory physiological function, quality of life, and mood, while few observations have been made to improve

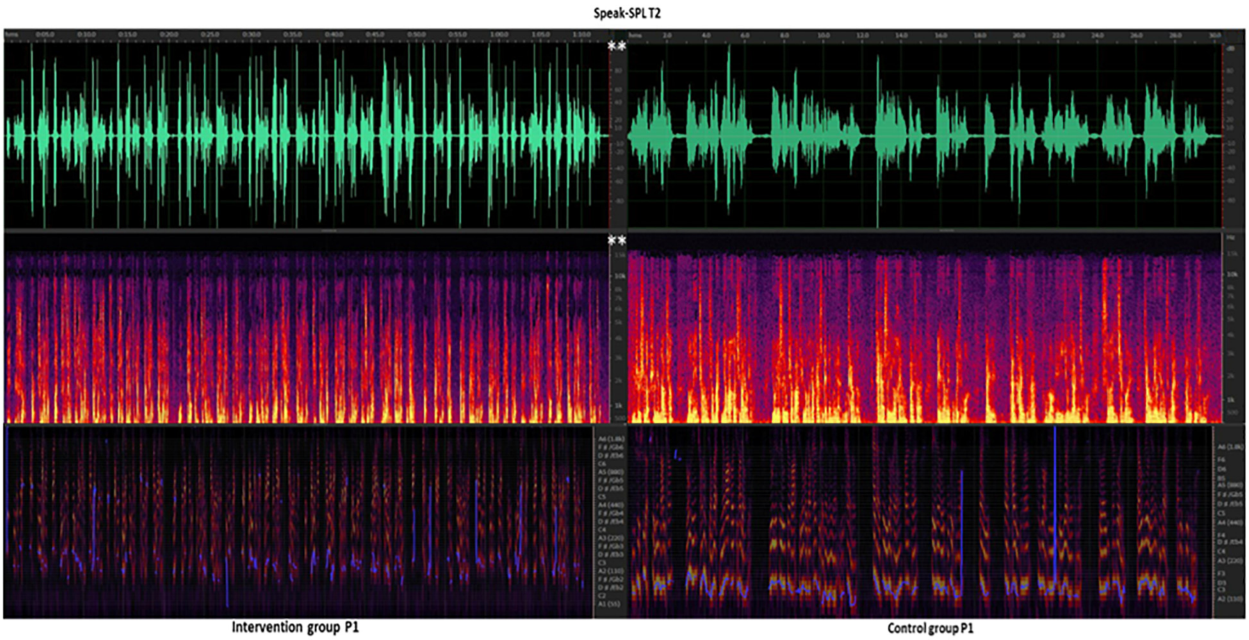


FIGURE 10 | Intervention group vs. Control group at T2 time point Speech-SPL (dBA) after the 12-week intervention. P, patient. Speech-SPL T2, intervention group patients vs. control group of Speak -SPL at T2 time point of spectral analysis. Line 1, green, the waveform spectrum represents the sound decibel (dBA) value; Line 2, purple, red, and yellow, fundamental frequency (F0, Hz) values; Line 3, purple, blue, indicates pitch (reference value). White ** indicates the $p < 0.01$, and the difference between groups is very significant.

sound quality or loudness. This study focused on the use of computer analysis-based sound spectrum analysis to observe the improvement of sound quality in CSCI patients with music therapy (Tamplin et al., 2013; Katz et al., 2022). In this study, the researchers used a meticulous and professional music intervention method to improve the sound quality of CSCI patients. 18 patients completed the 12-weeks, each participant had 60 sessions in total in the clinical trial. The improvement of musical vocal function, speech vocal function, duration, and audio frequency was observed.

Vocal Intonation Therapy Enhances Sing-Loud Pressure Level of Cervical Spinal Cord Injury

In this study, patients with CSCI were trained to sing short melodies to long melodic vowels and consonants using VIT step by step (vowel pronunciation function and consonant pronunciation function), focusing on the output of vocal function during singing. This was manifested by an increase in Sing-SPL (T1) time point after 6 weeks compared to the baseline (T0) time point results (**Figure 5**), at which point there was a statistical difference between the two groups of patients receiving different training methods. Among them, the vocal pressure level of the intervention group with CSCI who received VIT was significantly higher than that of the control group who received respiratory phonation training. Their vocal function was significantly improved in the accumulation of 6 weeks. In the second course of training that followed, the vocal function of the patients in the intervention group continued to improve with the guidance of effective training methods and the increase of training time. Their vocal function improved more obviously compared to the control group after 12 weeks (**Figure 9**). This shows that in the cumulative time effect of the same amount of training, the vocal function of patients with CSCI in different groups has obvious differences in intervention methods: the sound pressure value of the vocal function of the patients trained in VIT has been significantly increased, and the vocal function has been significantly improved.

Vocal Intonation Therapy Enhances Speech-SPL of Cervical Spinal Cord Injury

In this study, patients used vocalization methods to adjust vocal habits when practicing speech function. After adapting to 4 weeks of 30 min/day of training, most patients showed good behavioral changes that resulted in significant changes in speech function, which was manifested by an increase in Speech-SPL after 6 weeks (T1) compared to the baseline (T0) results. There was a statistical difference between the two groups of patients who received different training methods (**Figure 6**). Among them, the Speech-SPL of the intervention group CSCI patients who received VIT was significantly higher than that of the control group, which showed that after cervical spinal cord injury, patients with vocal dysfunction were trained in VIT, and their vocal function was significantly improved in the accumulation of 6 weeks. There was a more obvious improvement after 12 weeks, and the difference was more obvious (**Figure 10**). This shows that in the cumulative time effect of the same amount of training, the vocal function of patients with CSCI in different groups possesses obvious differences in intervention methods: the sound pressure value of a vocal function of patients trained in music therapy has been significantly increased, and the vocal function has been significantly improved.

Vocal Intonation Therapy Significantly Increased Sustained Note Length in Patients With Cervical Spinal Cord Injury

In the SNL test, the patients in the intervention group significantly increased the duration of the continuous tone after 12 weeks, indicating that when practicing singing short melodies (up and down the third chord, **Figure 2**) and long melodies (three-degree melodic intervals overlapping up and down, **Figure 2**), long-term sound output could be performed under the support of the thoracic and abdominal resonance cavity. Patients undergo VIT therapy in conjunction with adjustments to respiratory function to support the output of vocalization. After 6 weeks of 30 min/day of training, the patients in the intervention group improved SNL in terms of the length of music and the length of speech output. The SNL in the intervention group of patients receiving VIT (T1) was significantly longer than the control group (T1) trained in respiratory pronunciation, which indicates that patients with vocal dysfunction after CSCI were trained in vocal methods, and their SNL improvement was significantly improved after 6 weeks. In the second course of training that followed, the SNL (T2) of the patients in the intervention group continued to improve with the guidance of effective training methods and the increase of training time, and the difference was more conspicuous after 12 weeks (**Figure 7**). This shows that in the cumulative time effect of the same amount of training, there are obvious differences in the intervention methods of SNL in different groups of CSCI patients: patients who received VIT had significantly increased their pronunciation duration and improved significantly.

Vocal Intonation Therapy Significantly Increased F_0 in Patients With Cervical Spinal Cord Injury

In the test of F0, the amplitude of the F0 of the patients in the intervention group increased significantly after 12 weeks, indicating that after VIT exercises and singing exercises, it was conducive to the vocal function output of high-density amplitude under the support of the thoracic and abdominal resonance cavity. Patients undergo VIT and singing training in conjunction with adjusting breathing function, while using more conserving breath to support vocal output. After adapting to 6 weeks of 30 min/day of training, the patients in the intervention group showed a change in the amplitude of sound frequency. Still, there was no significant statistical difference at that time. After continuing the training with the same amount of time accumulation for 12 weeks, the output of the sound frequency of the experimental group increased significantly. Specifically, the intervention group F0 significantly improved the accumulation of time at 12 weeks compared with the control group (**Figure 8**). After completing 12 weeks of therapeutic training, the F0 (T2) of the intervention group patients increased significantly with the guidance of vocal training methods and the increase in training time. This shows that in the cumulative time effect of the same amount of training, the vocal function of different groups of CSCI patients has obvious differences in the intervention method: the sound waveform amplitude of patients who received VIT vocal training and singing training was significantly enhanced, and the width of the sound frequency was significantly increased.

LIMITATIONS

One limitation was the limited sample, as previously detailed. 12 participants dropped out of the study, which may have caused the variance in group allocation. The comparison might have been more accurate if a blank control group had been added to observe self-healing. Patients' quality of life or patient outcome measures, or feedback from family about post-intervention functional improvement would have been a good adjunct to fit alongside the acoustic analysis. The Voice Handicap Index (VHI) may have been a useful measure to include in this study to capture any effects of the intervention on communication-related quality of life. As found in previous research using the VHI (Tamplin et al., 2013), increases in speech volume for people with CSCI make their speech more audible and intelligible to others. Besides, this study only recruited 18 patients. If larger size studies are conducted in the future, the therapeutic outcomes could be more precisely observed.

Clinical Guidance by Vocal Intonation Therapy for the Improvement of Vocal Function

Clinicians can introduce songs gradually in order of difficulty (based on the length and pitch range of phrases). For example, a simpler song supplemental training "call-respond-chakra" singing format allows patients participating in a group practice to learn and practice together. During the first few weeks of the treatment process, the therapist can use songs with short phrases and songs with adequate breathing intervals. Actual clinical experience demonstrates that many patients with CSCI (or the general population with vocal dysfunction) initially have little confidence in their vocal function, so there is little emphasis on the quality of sound or pitch accuracy during actual VIT training. In the initial VIT singing training, the training method is mainly used to improve respiratory function and sound presentation and bring a pleasant emotional experience. Patients can sing with the accompaniment of an instrument or in karaoke style under background music recorded in the soundtrack to improve their vocal skills. According to the experience of clinical VIT plus singing training, the main recommended songs are (1) "Farewell"; (2) "Orchid Grass"; (3) "Hawthorn Tree"; (4) "Kangding Love Song"; (5) "Country Road"; (6) "The Country Road Takes Me Home"; (7) "Friendship lasts for a long time"; (8) "The Wind Blows the Wheat Waves"; (9) "Once Upon a Time"; (10) "Crooked Moon."

CONCLUSION

Vocal intonation therapy—music therapy can improve the loudness of cervical CSCI patients and provide CSCI patients with a practical, highly operable treatment that has both functional training effects and can bring a pleasant experience which can be vigorously promoted in the clinic.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

This study was reviewed and approved by the Medical Ethics Committee of China Rehabilitation Research Center (CRRC) on May 20th, 2019. Approval number 2019-83-1. Written informed consent was obtained from all participants for their participation in this study.

AUTHOR CONTRIBUTIONS

XZ: study design, statistical description, allocation, and statistical analysis. H-WL and D-GY: patient recruitment. Y-CS: music therapy. X-BL: project application. J-JL: study guidance. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Scientific Research Project “Establishment of the Winter Olympics Sports Injury Rehabilitation Diagnosis and Treatment System and Green Channel Demonstration” of Ministry of Science and Technology in China (No. 2018YFF0301104) to J-JL and the National Social Science Foundation of China Major Arts Project “Research on the Development Theory of Collaborative Innovation between Music and Artificial Intelligence” (No. 212D19) to X-BL. This research was supported by the China Rehabilitation Research Center (CRRC).

ACKNOWLEDGMENTS

We thank our colleagues from Music Therapy Center, CRRC; Chinese Institute of Rehabilitation Science; Center of Neural Injury and Repair; Department of Spinal and Neural Functional Reconstruction and Beijing Key Laboratory of Neural Injury and Rehabilitation, CRRC; who provided technical support, modification advice, and statistical recommendations. It is greatly assisted to the research.

REFERENCES

American Spinal Cord Injury Association [ASIA]. *Impairment Scale, Clinical Syndromes, and Standard Neurological Classification of Spinal Cord Injury*. Available online at: https://asia-spinalinjury.org (accessed May 2, 2022).

Behrman, A., Cody, J., Elandary, S., Flom, P., and Chitnis, S. (2020). The effect of speak out and the loud crowd on dysarthria due to parkinson’s disease. *Am. J. Speech Lang. Pathol.* 29, 1448–1465. doi: 10.1044/2020_AJSLP-19-00024

Berlowitz, D. J., Wadsworth, B., and Ross, J. (2016). Respiratory problems and management in people with spinal cord injury. *Breathe* 12, 328–340. doi: 10.1183/20734735.012616

Bruschettini, M., O’Donnell, C. P., Davis, P. G., Morley, C. J., Moja, L., and Calevo, M. G. (2020). Sustained versus standard inflations during neonatal resuscitation to prevent mortality and improve respiratory outcomes. *Cochrane Database Syst. Rev.* 7:CD004953.

Clini, E., Holland, A., Pitta, F., and Troosters, T. (2018). *Textbook of Pulmonary Rehabilitation*. Berlin: Springer International Publishing.

Crispiatico, V., Baldanzi, C., Napoletano, A., Tomasoni, L., Tedeschi, F., Groppo, E., et al. (2021). Effects of voice rehabilitation in people with MS: A double-blinded long-term randomized controlled trial. *Mult. Scler* [Epub ahead of print]. doi: 10.1177/13524585211051059

Fan, L. Z., Li, H., Zhuo, J. J., Zhang, Y., Wang, J. J., Chen, L. F., et al. (2016). The Human Brainne tome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cereb. Cortex* 26, 3508–3526. doi: 10.1093/cercor/bhw157

Gadomski, B. C., Hindman, B. J., Page, M. I., Dexter, F., and Puttlitz, C. M. (2021). Intubation Biomechanics: clinical Implications of Computational Modeling of Intervertebral Motion and Spinal Cord Strain during Tracheal Intubation in an Intact Cervical Spine. *Anesthesiology* 135, 1055–1065. doi: 10.1097/ALN.0000000000004024

Geneid, A., Nawka, T., Schindler, A., Oguz, H., Chrobok, V., Calcinoni, O., et al. (2020). Union of the European Phoniatriciansâ position statement on the exit strategy of phoniatric and laryngological services: staying safe and getting back to normal after the peak of coronavirus disease (issued on 25th May 2020). *J. Laryngol. Otol.* 134, 661–664. doi: 10.1017/S002221512000122X

Hixon, T. J., Goldman, M. D., and Mead, J. (1973). Kinematics of the chest wall during speech production: volume displacements of the rib cage, abdomen, and lung. *J. Speech Hear Res.* 16, 78–115. doi: 10.1044/jshr.1601.78

Jasmin, K. M., McGettigan, C., Agnew, Z. K., Lavan, N., Josephs, O., Cummins, F., et al. (2016). Cohesion and Joint Speech: right Hemisphere Contributions to Synchronized Vocal Production. *J. Neurosci.* 36, 4669–4680. doi: 10.1523/JNEUROSCI.4075-15.2016

Katz, S. L., Mah, J. K., McMillan, H. J., Campbell, C., Bijelić, V., Barrowman, N., et al. (2022). Routine lung volume recruitment in boys with Duchenne muscular dystrophy: a randomized clinical trial. *Thorax* [Epub ahead of print]. doi: 10.1136/thoraxjnl-2021-218196

Koda, M., Hanaoka, H., Fujii, Y., Hanawa, M., Kawasaki, Y., Ozawa, Y., et al. (2021). Randomized trial of granulocyte colony-stimulating factor for spinal cord injury. *Brain* 144, 789–799.

Lemos, J. R., da Cunha, F. A., Lopes, A. J., Guimarães, F. S., do Amaral Vasconcellos, F. V., Dos Santos Vigário, P., et al. (2020). Respiratory muscle training in non-athletes and athletes with spinal cord injury: A systematic review of the effects on pulmonary function, respiratory muscle strength and endurance, and cardiorespiratory fitness based on the FITT principle of exercise prescription. *J. Back Musculoskelet. Rehabil.* 33, 655–667. doi: 10.3233/BMR-181452

Lu, J., Wu, X., Li, Y., and Kong, X. (2008). Surgical results of anterior corpectomy in the aged patients with cervical myelopathy. *Eur. Spine J.* 17, 129–135. doi: 10.1007/s00586-007-0518-4

MacBean, V., Pooranampillai, D., Howard, C., Lunt, A. C., and Greenough, A. (2017). The influence of dilution on the offline measurement of exhaled nitric oxide. *Physiol. Meas.* 39:025004. doi: 10.1088/1361-6579/aaa455

Mesbah, S., Ball, T., Angeli, C., Rejc, E., Dietz, N., Ugiliweneza, B., et al. (2021). Predictors of volitional motor recovery with epidural stimulation in individuals with chronic spinal cord injury. *Brain* 144, 420–433. doi: 10.1093/brain/awaa423

Molina, E. J., Shah, P., Kiernan, M. S., Cornwell, W. K. I. I. I., Copeland, H., Takeda, K., et al. (2021). The Society of Thoracic Surgeons Intermacs 2020 Annual Report. *Ann. Thorac. Surg.* 111, 778–792. doi: 10.1016/j.athoracsur.2020.12.038

National Spinal Cord Injury Statistical Center [NSCISC] (2017). *Spinal Cord Injury Facts and Figures at a Glance*. Birmingham, Alabama: National Spinal Cord Injury Statistical Center.

Oraee, Y. S., Akhlaghpasand, M., Golmohammadi, M., Hafizi, M., Zomorrod, M. S., Kabir, N. M., et al. (2021). Combining cell therapy with human autologous Schwann cell and bone marrow-derived mesenchymal stem cell in patients with subacute complete spinal cord injury: safety considerations and possible outcomes. *Stem Cell Res. Ther.* 12:445. doi: 10.1186/s13287-021-02515-2

OSCIS investigators, Chikuda, H., Koyama, Y., Matsubayashi, Y., Ogata, T., Ohtsu, H., et al. (2021). Effect of Early vs Delayed Surgical Treatment on Motor Recovery in Incomplete Cervical Spinal Cord Injury With Preexisting Cervical Stenosis: A Randomized Clinical Trial. *JAMA Netw. Open* 4:e2133604. doi: 10.1001/jamanetworkopen.2021.33604

Qu, X., Wang, Q., Chen, W., Li, T., Guo, J., Wang, H., et al. (2019). Combined machine learning and diffusion tensor imaging reveals altered anatomic fiber connectivity of the brain in primary open-angle glaucoma. *Brain Res.* 1718, 83–90. doi: 10.1016/j.brainres.2019.05.006

Reimers, S., and Stewart, N. (2016). Auditory presentation and synchronization in Adobe Flash and HTML5/JavaScript Web experiments. *Behav. Res. Methods* 48, 897–908. doi: 10.3758/s13428-016-0758-5

Rodrigues, F. O., Sarmet, M., Maldaner, V., Yamasaki, R., Behlau, M., and Davison, M. L. (2021). Traumatic Spinal Injury: preliminary Results of Respiratory Function, Voice and Quality of Life. *J. Voice* [Epub ahead of print]. doi: 10.1016/j.jvoice.2021.02.009

Scheuren, P. S., David, G., Kipling, K. J. L., Jutzeler, C. R., Hupp, M., Freund, P., et al. (2021). Combined Neurophysiologic and Neuroimaging Approach to Reveal the Structure-Function Paradox in Cervical Myelopathy. *Neurology* [Epub ahead of print]. doi: 10.1212/WNL.0000000000012643

Schloneger, M. J., and Hunter, E. J. (2017). Assessments of Voice Use and Voice Quality Among College/University Singing Students Ages 18-24 Through Ambulatory Monitoring With a Full Accelerometer Signal. *J. Voice* 31, 124.e21–124.e30. doi: 10.1016/j.jvoice.2015.12.018

Strohl, M. P., Choy, W., Clark, A. J., Mummaneni, P. V., Dhall, S. S., Tay, B. K., et al. (2020). Immediate Voice and Swallowing Complaints Following Revision Anterior Cervical Spine Surgery. *Otolaryngol. Head Neck Surg.* 163, 778–784. doi: 10.1177/0194599820926133

Tamplin, J., Baker, F. A., Buttifant, M., and Berlowitz, D. J. (2014). The effect of singing training on voice quality for people with quadriplegia. *J. Voice* 28, 128.e19–128.e26. doi: 10.1016/j.jvoice.2013.08.017

Tamplin, J., Baker, F. A., Grocke, D., Brazzale, D. J., Pretto, J. J., Ruehland, W. R., et al. (2013). Effect of singing on respiratory function, voice, and mood after quadriplegia: a randomized controlled trial. *Arch. Phys. Med. Rehabil.* 94, 426–434. doi: 10.1016/j.apmr.2012.10.006

Tamplin, J., Brazzale, D. J., Pretto, J. J., Ruehland, W. R., Buttifant, M., Brown, D. J., et al. (2011). Assessment of breathing patterns and respiratory muscle recruitment during singing and speech in quadriplegia. *Arch. Phys. Med. Rehabil.* 92, 250–256. doi: 10.1016/j.apmr.2010.10.032

Thaut, M., and Hoemberg, V. (2014). *Handbook of Neurological Music Therapy*. Oxford: Oxford University Press.

Wadsworth, B. M., Haines, T. P., Cornwell, P. L., Rodwell, L. T., and Paratz, J. D. (2012). Abdominal binder improves lung volumes and voice in people with tetraplegic spinal cord injury. *Arch. Phys. Med. Rehabil.* 93, 2189–2197. doi: 10.1016/j.apmr.2012.06.010

Watson, P. J., and Hixon, T. J. (2001). Effects of abdominal trussing on breathing and speech in men with cervical spinal cord injury. *J. Speech Lang. Hear Res.* 44, 751–762. doi: 10.1044/1092-4388(2001/059)

Williams, A. M., Manouchehri, N., Erskine, E., Tauh, K., So, K., Shortt, K., et al. (2020). Cardio-centric hemodynamic management improves spinal cord oxygenation and mitigates hemorrhage in acute spinal cord injury. *Nat. Commun.* 11:5209. doi: 10.1038/s41467-020-18905-8

Witold, M. (2020). Reducing the harmful effects of noise on the human environment. Sound insulation of industrial skeleton enclosures in the 10-40 kHz frequency range. *J. Environ. Health Sci. Eng.* 18, 1451–1463. doi: 10.1007/s40201-020-00560-2

Zhang, X. Y., Song, Y. C., Liu, C. B., Qin, C., Liu, S. H., and Li, J. J. (2021). Effectiveness of oral motor respiratory exercise and vocal intonation therapy on respiratory function and vocal quality in patients with spinal cord injury: a randomized controlled trial. *Neural Regen. Res.* 16, 375–381. doi: 10.4103/1673-5374.290909

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

BRIEF RESEARCH REPORT
published: 05 July 2022
doi: 10.3389/fcomm.2022.887739

Iconic Associations Between Vowel Acoustics and Musical Patterns, and the Musical Protolanguage Hypothesis

*Gertraud Fenk-Oczlon**

University of Klagenfurt, Klagenfurt, Austria

Vowels are the most musical and sonic elements of speech. Previous studies found non-arbitrary associations between vowel intrinsic pitch and musical pitch in senseless syllables. In songs containing strings of senseless syllables, vowels are connected to melodic direction in close correspondence to their *intrinsic pitch* or the frequency of the second formant F2. This paper shows that also *vowel intrinsic duration* is related to musical patterns. It is generally assumed that low vowels like [a ɔ o] have a higher intrinsic duration than high vowels like [i y u] and that there is a positive correlation between the first formant F1 and duration. Analyzing 20 traditional Alpine yodels I found that vowels with longer intrinsic duration tend to align with longer notes, whereas vowels with shorter intrinsic duration with shorter notes. This new result might shed some light on size-sound symbolism in general: Since there is a direct match between vowel intrinsic duration and the "size" of musical notes, there is no need to explain the "size" of musical notes via Ohala's "frequency code" hypothesis. Moreover, I will argue that the iconic associations found between vowel acoustics and musical patterns support the idea of a sound-symbolic musical protolanguage. Such a protolanguage may have started with vowel syllables conveying pitch, timbre, as well as emotional, indexical, and sound-symbolic information.

Keywords: intrinsic vowel duration, size-sound symbolism, iconicity, yodels, musical notes, evolution, musical protolanguage, Ohala's "frequency code" hypothesis

OPEN ACCESS

Edited by:
Caicai Zhang,
The Hong Kong Polytechnic University, Hong Kong SAR, China

Reviewed by:
Oliver Niebuhr,
University of Southern Denmark, Denmark
Julien Meyer,
Centre National de la Recherche Scientifique (CNRS), France

****Correspondence:***
Gertraud Fenk-Oczlon
Gertraud.fenk@aau.at

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *01 March 2022*
Accepted: *09 June 2022*
Published: *05 July 2022*

Citation:
Fenk-Oczlon G (2022) Iconic Associations Between Vowel Acoustics and Musical Patterns, and the Musical Protolanguage Hypothesis.
Front. Commun. 7:887739.
doi: 10.3389/fcomm.2022.887739

INTRODUCTION

Language and music share many commonalties, consistent with a view according to which both have a common evolutionary precursor. The hypothesized common ancestor is often referred to as "musilanguage" (Brown, 2000), "musical protolanguage" (Fitch, 2005), or "prosodic protolanguage" (Fitch, 2006). A growing number of researchers further emphasizes the idea that affective/emotional and iconic vocalizations could have played a significant role in the joint evolution of speech and music (Rousseau, 1781; Darwin, 1871; Fonagy, 1981; Levman, 1992; Scherer, 1995; Thompson et al., 2012; Perlman and Cain, 2014; Brown, 2017; Filippi and Gingras, 2018; Reybrouck and Podlipniak, 2019; Filippi, 2020).

This paper focuses on the role of vowels in the hypothetical construct "musical protolanguage." I will briefly review some literature that has demonstrated tight relationships between vowels and music, and that has revealed the essential role of vowels in speech intelligibility of sentences, in conveying emotional content and talker discrimination, as well as in size-sound symbolism. I then present new results showing an iconic relationship between vowel duration and musical notes in Alpine yodels. The implications for sound symbolism in general, as well as for the idea of a sound-symbolic musical protolanguage will be discussed.

The most obvious commonality between speech and music is sound, and it is the vowels that are the main carriers of sound and prosodic information in speech and singing (e.g., Fenk-Oczlon and Fenk, 2009b). Vowels are produced without obstructing the airflow from the lungs and are relatively continuous or steady-state sounds exhibiting a greater periodicity than consonants (Cutler and Mehler, 1993). According to Halle et al. (1957, p. 116) vowels can be matched easily in pitch to pure tones, whereas determinations of pitch of consonants "usually refer to the terminal stage of the second formant in the adjacent vowel." Vowels are distinguished by their timbre, which depends on their harmonics or overtones, whereby the formants F1 and F2 are most relevant for their identification (Peterson and Barney, 1952). The main articulatory parameters responsible for vowel timbre are tongue height, front-to back position of the tongue, and lip rounding. The changes in the vowels' resonances are audible in the case of whispering, when the vocal chords do not vibrate, or when speaking in a creaky voice (Ladefoged, 2001). Indeed, when whispering series of words like *heed, hid, head, had*, *hawed* one can hear the descending pitch of F2; and when speaking the series *hawed, had, head, hid, heed* in a creaky voice, the descending pitch of F1 is audible.

Timbre is clearly the primary parameter that allows for discriminating between different vowels, but vowels differ also in intrinsic pitch, intensity and duration. It is known since Meyer (1896) that, all other things being equal, high vowels such as /i/ have a higher intrinsic fundamental frequency IF0 than low vowels such as /a/. Whalen et al. (1995) could observe this effect in a sample of 31 languages and even in babbling. While the mechanism determining IF0 is still a subject of debate, there seems to be general agreement that vowel pitch depends primarily on the frequency of the second formant F2 (Marks, 1975; Traunmüller, 1986). Concerning vowel intrinsic duration it is generally assumed that low vowels have a higher intrinsic duration than high vowels like [i u y]. and that there is a positive correlation between the first formant F1 and duration, i.e., the lower the vowel, the higher F1, and the higher the intrinsic duration of the vowel (House and Fairbanks, 1953; Peterson and Lehiste, 1960; Lehiste, 1970; Sol and Ohala, 2010; Toivonen et al., 2015). According to House and Fairbanks (1953) intrinsic vowel duration differences show in various types of consonant environments (voiced and voiceless stops and fricatives, nasals); for instance, when pooled across all environments the vowel /i/ has a mean duration of 0.199 s and the vowel /a/ of 0.244 s.

Evidently, vowels show all the core properties of music—timbre, intrinsic pitch, intensity and duration—and they are the most musical components of speech. Recent studies revealed tight relationships between vowels and music. For example, in Fenk-Oczlon (2017) I reported correspondences between the number of vowels and the number of pitches in musical scales across cultures: an upper limit of roughly 12 elements, a lower limit of 2, and a frequency peak at 5 to 7 elements. The match between vowels and musical pitches shows even in specific cultures: e.g., cultures with three vowels tend to have tritonic scales. Concerning relationships between vowel acoustics and musical pitch, Fürniss (1991) reported associations between low vowels and the "low yodel register" and closed vowels and the "high yodel register" in the yodeling of Aka Pygmies; Fenk-Oczlon and Fenk (2009a,b) showed non-arbitrary associations between vowel intrinsic pitch and musical pitch in Alpine yodeling and in Austrian songs containing meaningless syllables. The tight bond between vowels and music is supported by experimental findings demonstrating strong interactions in the processing of vowels and melody, but not between consonants and musical information: "Vowels sing but consonants speak" (Kolinsky et al., 2009, p. 1). Similarly, Lidji et al. (2010) revealed a close processing relationship between vowels and pitch even at a pre-attentive level. Moreover, experiments by Zhang et al. (2017) demonstrated that congenital amusics not only show deficits in the perception of pitch but also in the perception of formant frequency in vowels.

Vowels and their acoustic properties are essential in many further aspects of language and speech, such as in speech intelligibility of sentences, in talker identity discrimination and in conveying emotional state, or in sound symbolism. For example, experimental studies revealed that the intelligibility of sentences was significantly better when hearing vowel-only sentences than when hearing consonant-only sentences (Cole et al., 1996; Kewley-Port et al., 2007). Vowels, unlike consonants, also provide rich indexical information about speaker identity and characteristics such as age, biological sex, origin and emotional state (Owren and Cardillo, 2006). Concerning relationships between vowels and emotional state, Rummer et al. (2014) demonstrated that subjects in a positive mood tend to invent words with /i:/, whereas when in a negative mood they tend to invent more words with /o:/.

As to sound symbolism (the non-arbitrary relation between sound and meaning), vowels are the main drivers in "size-sound symbolism" or "magnitude sound symbolism," i.e., the association between size (large/small) and sound. In a classic study, Sapir (1929) demonstrated that participants associate meaningless words containing low and back vowels like /a/ (e.g., as in *mal*) with large concepts and meaningless words containing high and front vowels like /i/ (e.g., as in *mil*) with small concepts. Numerous experimental studies could replicate Sapir's finding showing the postulated association between vowel quality and size (Bentley and Varon, 1933; Peña et al., 2011; Parise and Spence, 2012; Shinohara and Kawahara, 2016; Knoeferle et al., 2017; Vainio, 2021). Likewise, statistical studies in typologically diverse languages found associations between the high front vowel /i/ and the concept of small (Ultan, 1978; Haynie et al., 2014; Blasi et al., 2016; Johansson et al., 2020). Most recently, Winter and Perlman (2021) demonstrated that—in English—size adjectives clearly feature iconicity, and that the high front vowels /i/ and /I/ are associated with "small," while the low back vowel /α/ predicts "large." The only consonant that predicts size symbolism in their English sample was /t/. In general, consonants seem to play a rather marginal role in sound-size associations, whereas their role in sound-shape associations as in the *maluma/takete* effect (Köhler, 1929) or the *bouba–kiki* effect (Ramachandran and Hubbard, 2001) is well-attested (but see Cuskley et al., 2017 on possible influences of orthography.)

Further cross-modal correspondences between vowels and other sensory modalities have been demonstrated between "vowels and quickness" (Jespersen, 1933), "vowels and brightness" (Marks, 1975), "vowels and spatial deixis" (Traunmüller, 1986; Johansson and Zlatev, 2013; Rabaglia et al., 2016; Vainio, 2021), "vowels and color" (Moos et al., 2014; Cuskley et al., 2019), or "vowels and taste" (Simner et al., 2010; Patak and Calvert, 2021).

Here I investigate whether there are iconic associations between the acoustic vowel property "intrinsic duration" (see above) and the length of musical notes. More specifically, I hypothesized that in songs containing meaningless syllables, syllables with low vowels like [a ɔ o] should be favored for long notes and syllables with high vowels like [i u y] for short notes.

MATERIALS AND METHODS

The singing of senseless syllables, where "the pressures of sense are relaxed to those of sound" (Butler 2015, p. 106) provides an ideal material to study relationships between vowels and musical notes. Senseless syllables are used in numerous cultures as complete or partial song texts, for example in Native American songs (Nettl, 1954), in "lilting" or "diddling," in the singing of Scottish or Irish dance melodies, in children's songs and jazz scat singing, or in yodeling. Here, I chose yodels for testing the hypothesized relationship between vowels and musical notes. The yodeling style, although on the whole not very frequent, can be found around the world (Grauer, 2006), for instance in Paleosiberian cultures, in the tropical forest of Africa (Pygmies), in the Kalahari Desert (Bushmen), and in the Alps (Austria, Switzerland). According to Grauer (2006) yodels are characterized across cultures by a continuous flow of sound, no embellishment, relaxed open voices, non-sense vocables, wide intervals and a polyphonic style. These characteristics also apply to traditional Alpine yodels, which are preferably polyphonic and mostly—but not necessarily—sung with frequent alternation between low and high registers (cf. Wey, 2019); they are yodeled straight without vibrato or portamento and with meaningless syllables. The yodel-syllables are predominately codaless, with rather weak or sonorant consonants in the syllabic onset, such as [jɔ, *ha*, *h*ɔ, ji, ri, ho, ha]. Vowel-only syllables and codaless syllables with a liquid in the syllabic nucleus like "dl," occur as well. The transcriptions into musical notation of the previously only orally transmitted Alpine yodels started at the beginning of the 19th century (Wey, 2019). The traditional yodels for the present study are taken from Pommer's (1906) collection of 20 yodels. Most of the yodels of this collection are still yodeled in Austria and are well-known, so that the grapheme—phoneme correspondence of this more than 100 years old transcriptions can be checked. For instance, the grapheme "å" is still used in Bavarian writing to denote an open "o" /ɔ/.

All 20 yodels in the collection were analyzed. I determined all relative note values in the sample: half notes (the longest note values in the sample), quarter notes, eighth notes, sixteenth notes, and thirty-second notes (the shortest notes in the sample). The notes were assigned to the respective syllables containing either high close vowels like [i u y] or low back vowels like [a ɔ o] Furthermore, all dotted notes—the dot increases the duration of the basic note by half of its original value—were identified and matched with the particular syllables.

RESULTS

The total number of notes/syllables in the sample amounts to 1,836. The most frequent note values are eighth notes ($n = 845$), followed by quarter notes ($n = 672$), half notes ($n = 190$), sixteenth notes ($n = 95$), and thirty-second notes ($n = 34$); the number of dotted notes amounts to 348. Syllables with high vowels ($n = 1,203$) are more often used in the yodel sample than syllables with low vowels ($n = 633$); ($X^2 = 176.961$, $p < 0.0001$).

A detailed analysis: Eighth notes are more often aligned with high vowels (590x) than with low vowels (255x), ($X^2 = 132.811$, $p < 0.0001$). Quarter notes are 405 times aligned with high vowels and 267 times with low vowels ($X^2 = 28.339$, $p < 0.0001$). Sixteenth notes are associated with high vowels 45 times and with low vowels 50 times ($X^2 = 0.263$, n.s.). Thirty-second notes are 28 times aligned with high vowels and 6 times with low vowels ($X^2 = 14.235$, $p < 0.001$).

On the contrary half notes, the longest note values in the sample, are more often aligned with low vowels (135x) and less frequently associated with high vowels (55x), ($X^2 = 33.684$, $p < 0.0001$). This also holds for dotted notes which are 265 times associated with low vowels and only 83 times with high vowels ($X^2 = 95.184$, $p < 0.0001$). **Figure 1** shows an example.

DISCUSSION

Our analysis of 20 Alpine yodels demonstrates that short musical notes such as eighth notes, quarter notes and thirty-second notes tend to align with vowels with smaller intrinsic duration, whereas relative long notes such as half notes or dotted notes are associated with vowels with longer intrinsic duration. These results need to be confirmed in further studies that use an extended sample of songs containing meaningless syllables. It would also be interesting to investigate, whether in an artificial music composition game, people will tend to align vowels with longer intrinsic duration to longer notes.

Vowel Intrinsic Pitch and Size-Sound Symbolism

The iconic associations between vowel intrinsic duration and length of musical notes may shed some light on size-sound symbolism in general. Although "duration" of musical notes only metaphorically corresponds to "size" of notes, our data are in line with results by Knoeferle et al. (2017) suggesting F1 and vowel duration are decisive factors in size-sound symbolism; F0 or Ohala (1984, 1994) "frequency code" hypothesis, according

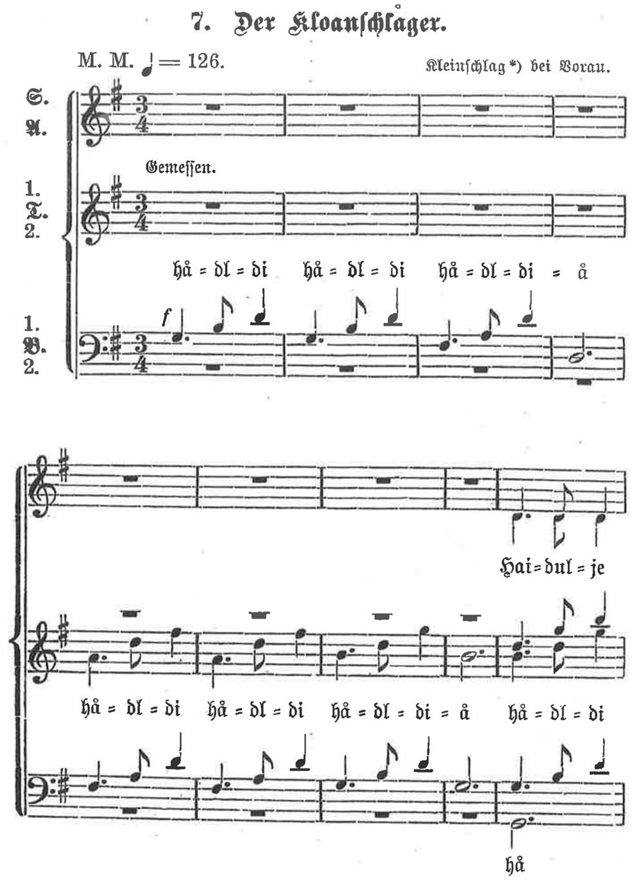

7. Der Kloanſchlager.
M. M. ♩ = 126.
Kleinſchlag *) bei Vorau.
S.
A.
Gemeſſen.
1.
T.
2.
hå = dl = di hå = dl = di hå = dl = di = å
1.
B.
2.
Hai = dul = je
hå = dl = di hå = dl = di hå = dl = di = å hå = dl = di
hå

FIGURE 1 | An example of a yodeler from our sample shows that dotted and half notes tend to be linked with syllables containing the vowel å /ɔ/ that has a longer intrinsic duration.

TABLE 1 | Examples of vowel-only sentences and vowel-only expletives in Japanese, Carinthian and in the language of the Mbendjele Pygmies.

ue o ui, o ooi, ai o ou, ai ue o [worried about hunger, concealing old age, he seeks love, a love- hungry man] *ooo, oooo, oo ooo* [the courageous king conceals his tail when he goes out]	Japanese examples from Tsunoda (1985) cited in Bannan (2008)
a i a? Me too? "a" question particle, "i" ich (I) "a" auch (also) *a e i a*! Me too! "a" interjection (astonishment) *:e(h)* particle, "I" ich (I) "a" auch (also)	Carinthian (South Bavarian dialectal variant)
iiiiiii expletive for surprise or disgust uuuuooooo expletive to accompany a dangerous or outrageous act iiiieeee expletive to indicate pleasure	Mbendjele Pygmies examples from Lewis (2009)

to which size-symbolism mirrors the size of the vocalizers producing either lower or higher frequencies, do not seem to play a role in their experiments on visual size judgements. Similarly, Vainio (2021) reports that F0 values did not show to be relevant in his study on magnitude sound symbolism. Since our results demonstrate a direct match between vowel intrinsic duration and the "size" of musical notes, there is no need to explain the "size" of musical notes via Ohala's "frequency code" hypothesis. Therefore, a possible answer to the question What is, for example, so small about *mil* and large about *mal*? (Vainio 2021, p. 2) might be: Small about *mil*, is the small intrinsic duration of the vowel /i/, and large about *mal* is the large intrinsic duration of the vowel /a/.

Vowels and a Sound-Symbolic Musical Protolanguage

The non-arbitrary associations between vowel *intrinsic duration* and musical notes are consistent with the results of previous studies (Fenk-Oczlon and Fenk, 2009a,b) reporting non-arbitrary associations between vowel *intrinsic pitch* and musical pitch in meaningless syllables: In songs containing strings of meaningless syllables, vowels are connected to melodic direction in close correspondence to their *intrinsic pitch* or the frequency of the second formant F2. The tight relationships between vowel acoustics and musical intervals indicate that in the case of singing senseless syllables, where there is no pressure of text, vowels and melody seem to merge. This might strengthen the idea that both music and speech evolved from a common prosodic precursor.

In Fenk-Oczlon (2017) I speculated that the earliest human vocal communication may have started with vowels or vowel syllables strung together, which were connected by semivowels or glides such as [w], [h], [j] or the glottal stop [ʔ]. The vowel sequences exhibited pitch and timbre modulations which were used to express different social and pragmatic functions, and were probably propositionally meaningless. The main arguments for this speculation were based on findings from language ontogeny, ethnomusicology, and parallels between vowels and musical patterns. In the 2017 paper I did not consider the huge sound symbolic potential of vowels and their disproportionate role in talker identity discrimination, including characteristics such as age, biological sex, origin, or emotional state. Considering all these properties of vowels, it seems plausible that the sequences of vowel syllables were not *bare phonology* in the sense of Fitch (2010), but instead conveyed sound symbolic information about the environment, about emotional states, or speaker identity. The sequences of vowel syllables probably also contained interjections similar to present-day words such as *ah, oh, eh, huh*. In this context it is interesting to note that Dingemanse et al. (2013) reported that all variants of the interjection word *huh* in their cross-linguistic sample consisted either of a vowel-only syllable, a syllable with a glottal stop [ʔ], or a glottal fricative [h] in the onset.

The vowel sequences were likely very polysemous, because of the small number of vowels (present-day languages have on average 5–6 vowels; Maddieson, 2005) which does not allow much variation in a sequence. Only pitch, duration, intonational contour, rhythmic grouping and situational context could help to discriminate the different (sound symbolic) meanings.

Even in present-day languages, vowel-only sentences can be observed. **Table 1** gives some examples from Japanese (Tsunoda, 1985), Carinthian (my own native knowledge) and vowel-only expletives from the Mbendjele Pygmies (Lewis, 2009). I am not able to analyze the Japanese examples, but the Carinthian example shows that the word "*a*"/ a/ is quite polysemous: It can be a question particle, an interjection of astonishment, and also denotes *auch* "also." The expletives from the Mbendjele Pygmies nicely demonstrate the potential of vowels to convey emotional content. Furthermore, Lewis (2009) reports that vowel-only sentences can also be observed in very intimate communication situations between two persons of the Mbendjele Pygmies, who "tend to omit consonants, leaving only tone and vowels" (Lewis 2009, p. 241).

One might speculate that the earliest stage of human vocal communication, where mere vowel syllables connected by semivowels were strung together, best represents the hypothesized common prosodic precursor of speech and music. The vowel syllables exhibited all core elements of music, pitch, timbre, duration, and intensity. They conveyed prosodic information such as intonation, rhythm, tempo, but also (semantic) sound-symbolic or onomatopoetic information about the environment, inner mental states or speaker identity. In a later stage, consonants such as obstruents emerged and were combined with vowels into consonant-vowel syllables. This was likely the emergence of articulated speech (Jordania, 2006), and of utterances which could express propositional meaning.

Grauer (2006) speculated that yodeling might be a vestige of the earliest singing style of humanity. The Alpine yodel syllables investigated in this paper may not be too different from the vowel syllables in the hypothesized earliest stage of human vocal communication.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

I thank the reviewers for their insightful comments and helpful suggestions.

REFERENCES

Bannan, N. (2008). Language out of music: the four dimensions of vocal learning. *Aust. J. Anthropol.* 19, 272–293. doi: 10.1111/j.1835-9310.2008.tb00354.x

Bentley, M., and Varon, E. J. (1933). An accessory study of "phonetic symbolism." *Am. J. Psychol.* 45, 76–86. doi: 10.2307/1414187

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., and Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proc. Natl. Acad. Sci. U.S.A.* 113, 10818–10823. doi: 10.1073/pnas.1605782113

Brown, S. (2000). "The Musilanguage model of music evolution," in: *The Origins of Music*, eds N. L. Wallin, B. Merker, and S. Brown (Cambridge, MA: The MIT Press). doi: 10.7551/mitpress/5190.001.0001

Brown, S. (2017). A joint prosodic origin of language and music. *Front. Psychol.* 8:1894. doi: 10.3389/fpsyg.2017.01894

Butler, S. (2015). *The Ancient Phonograph*. Boston, MA: Zone Books. doi: 10.2307/j.ctv14gpj13

Cole, R., Yan, Y., Mak, B., Fanty, M., and Bailey, T. (1996). "The contribution of consonants versus vowels to word recognition in fluent speech," in: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing ICASSP'96*. Atlanta, GA.

Cuskley, C., Dingemanse, M., Kirby, S., and van Leeuwen, T. M. (2019). Cross-modal associations and synesthesia: categorical perception and structure in vowel–color mappings in a large online sample. *Behav. Res. Methods*.51, 1651–1675. doi: 10.3758/s13428-019-01203-7

Cuskley, C., Simner, J., and Kirby, S. (2017). Phonological and orthographic influences in the bouba–kiki effect. *Psychol. Res*.81, 119–130. doi: 10.1007/s00426-015-0709-2

Cutler, A., and Mehler, J. (1993). The periodicity bias. *J. Phonetics* 21, 103–108. doi: 10.1016/S0095-4470(19)31323-3

Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*. London: J.Murray. doi: 10.5962/bhl.title.24784

Dingemanse, M., Torreira, F., and Enfield, N. J. (2013). Is 'Huh?' a universal word? Conversational infrastructure and the convergent evolution of linguistic items. *PLoS ONE* 8:e78273. doi: 10.1371/journal.pone.0078273

Fenk-Oczlon, G. (2017). What vowels can tell us about the evolution of music. *Front. Psychol.* 8:1581. doi: 10.3389/fpsyg.2017.01581

Fenk-Oczlon, G., and Fenk, A. (2009a). "Musical pitch in nonsense syllables: correlations with the vowel system and evolutionary perspectives," in *Proceedings of 7th Triennial Conference of the Europaean Society for the Cognitive Sciences of Music*, eds J. Louhivuori, T. Eeerola, S. Saarikallio, T. Himberg, and P.-S. Eerola (Jyväskylä: European Society for the Cognitive Sciences of Music).

Fenk-Oczlon, G., and Fenk, A. (2009b). Some parallels between language and music from a cognitive and evolutionary perspective. *Music. Sci.* 13, 201–226. doi: 10.3389/fnins.2016.00274

Filippi, P. (2020). Emotional voice intonation: A communication code at the origins of speech processing and word-meaning associations? *J. Nonverb. Behav.* 44, 395–417. doi: 10.1007/s10919-020-00337-z

Filippi, P., and Gingras, B. (2018). "Emotion communication in animal vocalizations, music and language: An evolutionary perspective," in: *The Talking Species*, eds E. M. Luef and M. M. Marin (Graz: Uni-Press Graz Verlag GmbH).

Fitch, W. T. (2005). The evolution of language: A comparative review. *Biol. Philosophy* 20, 193–230. doi: 10.1007/s10539-005-5597-1

Fitch, W. T. (2006). The biology and evolution of music: a comparative perspective. *Cognition* 100, 173–215 doi: 10.1016/j.cognition.2005.11.009

Fitch, W. T. (2010). *Evolution of Language*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511817779

Fonagy, I. (1981). "Emotions, voice and music," in: *Research Aspects on Singing*, ed J. Sundberg (Stockholm and Paris: Royal Swedish Academy of Music).

Fürniss, S. (1991). *Die Jodeltechnik der Aka-Pygmäen in Zentralafrika.* Berlin: Dieter Reimer.

Grauer, V. A. (2006). Echoes of our forgotten ancestors. *World Music* 48, 5–58.

Halle, M., Hughes, G. W., and Radley, J. -P. A. (1957). Acoustic properties of stop consonants. *J. Acoust. Soc. Am*. 29, 107. doi: 10.1121/1.1908634

Haynie, H., Bowern, C., and La Palombara, H. (2014). Sound symbolism in the languages of Australia. *PLoS ONE* 9:e92852. doi: 10.1371/journal.pone.0092852

House, A. S., and Fairbanks, G. (1953). The influence of consonant envi-ronment upon the secondary acoustical characteristics of vowels. *J. Acoust. Soc. Am*. 25, 105–113. doi: 10.1121/1.1906982

Jespersen, O. (1933). *Symbolic value of the vowel i. In Linguistica; Selected Papers in English, French, and German*. Copenhagen: Levin & Munksgaard.

Johansson, N., and Zlatev, J. (2013). Motivations for sound symbolism in spatial deixis: a typological study of 101 languages. *Public J. Semiot.* 5, 3–20. doi: 10.37693/pjos.2013.5.9668

Johansson, N. E., Anikin, A., Carling, G., and Holmer, A. (2020). The typology of sound symbolism: Defining macro-concepts via their semantic and phonetic features. *Linguist. Typol.* 24, 253–310. doi: 10.1515/lingty-2020-2034

Jordania, J. (2006). *Who Asked the First Question? The Origins of Human Choral Singing, Intelligence, Language and Speech. The Origins of Human Choral Singing, Intelligence*. Tbilisi: Logos.

Kewley-Port, D., Burkle, T. Z., and Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *J. Acoust. Soc. Am.* 122, 2365–2375. doi: 10.1121/1.2773986

Knoeferle, K., Li, J., Maggioni, E., and Spence, C. (2017). What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings. *Sci. Rep.* 7, 1–11. doi: 10.1038/s41598-017-05965-y

Köhler, W. (1929). *Gestalt Psychology*. New York, NY: Liveright.

Kolinsky, R., Pascale Lidji, P., Peretz, I., Besson, M., and Morais, J. (2009). Processing interactions between phonology and melody: Vowels sing but consonants speak. *Cognition* 112, 1–20. doi: 10.1016/j.cognition.2009.02.014

Ladefoged, P. (2001) *Vowels and Consonants: An Introduction to the Sounds of Languages*. Oxford: Blackwell, Blackwell Publications, Malden.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: The MIT Press.

Levman, B. (1992). The genesis of music and language. *Ethnomusicology* 36, 147–117 doi: 10.2307/851912

Lewis, J. (2009). "As well as words: Congo Pygmy hunting, mimicry, and play," in: *The Cradle of Language*, eds R. Botha and C. Knight (Oxford: Oxford University Press).

Lidji, P., Jolicoeur, P., Régine Kolinsky, R., Moreau, P., Connolly, J. F., and Peretz, I. (2010). Early integration of vowel and pitch processing: A mismatch negativity study. *Clin. Neurophysiol.* 121, 533–541. doi: 10.1016/j.clinph.2009.12.018

Maddieson, I. (2005). "Vowel quality inventories," in *The World Atlas of Language Structures*, eds M. Haspelmath, M. S. Dryer, D. Gil, and B. Comrie (Oxford: Oxford University Press).

Marks, L. E. (1975). On colored-hearing synesthesia: Cross-modal trans- lations of sensory dimensions. *Psychol. Bullet.* 82, 303–331. doi: 10.1037/0033-2909.82.3.303

Meyer, E. A. (1896). Zur tonbewegung des vokals im gesprochenen und gesungenen einzelwort. *Phonet. Stud.* 10, 1–21.

Moos, A., Smith, R., Miller, S. R., and Simmons, D. R. (2014). Cross- modal associations in synaesthesia: Vowel colours in the ear of the beholder. *i-Perception,* 5, 132–142. doi: 10.1068/i0626

Nettl, B. (1954). Text-music relationships in Arapaho songs. *Southwestern J. Anthropol.* 10, 192–199. doi: 10.1086/soutjanth.10.2.3628825

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica* 41, 1–16. doi: 10.1159/000261706

Ohala, J. J. (1994). "The frequency code underlies the sound-symbolic use of voice pitch," in: *Sound Symbolism*, eds H. Leanne, N. Johanna and O. John (Cambridge: Cambridge University Press). doi: 10.1017/CBO9780511751806.022

Owren, M. J., and Cardillo, G. C. (2006). The relative roles of vowels and consonants in discriminating talker identity versus word meaning. *J. Acoust. Soc. Am.* 119, 1727–1739. doi: 10.1121/1.2161431

Parise, C., and Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test. *Experi. Brain Res.* 220, 319–333. doi: 10.1007/s00221-012-3140-6

Patak, A., and Calvert, G. A. (2021). Sooo sweeet! Presence of long vowels in brand names lead to expectations of sweetness. *Behav. Sci.*11:12. doi: 10.3390/bs11020012

Peña, M., Mehler, J., and Nespor, M. (2011). The role of audiovisual processing in early conceptual development. *Psychol. Sci.* 22, 1419–1421. doi: 10.1177/0956797611421791

Perlman, M., and Cain, A. A. (2014). Iconicity in vocalization, comparisons with gesture, and implications for theories on the evolution of language. *Gesture* 14, 320–350. doi: 10.1075/gest.14.3.03per

Peterson, G. E., and Barney, H. L. (1952). Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* 24, 175–184. doi: 10.1121/1.1906875

Peterson, G. E., and Lehiste, I. (1960). Duration of syllable nuclei in English. *J. Acoustical Soc. Am.* 32, 693–703. doi: 10.1121/1.1908183

Pommer, J. (1906). *Zwanzig echte alte Jodler*. Wien: Adolf Robitschek.

Rabaglia, C. D., Maglio, S. J., Krehm, M., Seok, J. H., and Trope, Y. (2016). The sound of distance. *Cognition* 152, 141–149. doi: 10.1016/j.cognition.2016.04.001

Ramachandran, V. S., and Hubbard, E. M. (2001). Synaesthesia – a window into perception, thought and language. *J Consciousness Stud.* 8, 3–34.

Reybrouck, M., and Podlipniak, P. (2019). Preconceptual spectral and temporal cues as a source of meaning in speech and music. *Brain Sci.* 9:53. doi: 10.3390/brainsci9030053

Rousseau, J.-J. (1781). *Essay on the Origin of Languages. English Translation by J. H. Moran and A. Gode (1986)*. Chicago, IL: University of Chicago Press.

Rummer, R., Schweppe, J., Schlegelmilch, R., and Grice, M. (2014). Mood is linked to vowel type: The role of articulatory movements. *Emotion* 14, 246–250. doi: 10.1037/a0035752

Sapir, E. (1929). A study in phonetic symbolism. *J. Experi. Psychol.* 12:225. doi: 10.1037/h0070931

Scherer, K. R. (1995: Expression of emotion in voice und music. *J. Voice* 9, 235–248. doi: 10.1016/S0892-1997(05)80231-0

Shinohara, K., and Kawahara, S. (2016). "A cross-linguistic study of sound symbolism: the images of size," in *Proceedings of the Thirty-Sixth Annual Meeting of the Berkeley Linguistics Society*. Berkeley. doi: 10.3765/bls.v36i1.3926

Simner, J., Cuskley, C., and Kirby, S. (2010). What sound does that taste? Cross-modal mappings across gustation and audition. *Perception* 39, 553–569. doi: 10.1068/p6591

Sol,é, M. J., Ohala, J. J. (2010). "What is and what is not under the control of the speaker. Intrinsic vowel duration," in: *Papers in Laboratory Phonology 10*, eds C. Fougeron, B. Kühnert, M. D'Imperio, and N. Vallée (Berlin: de Gruyter).

Thompson, W. F., Marin, M. M., and Stewart, L. (2012). Reduced sensitivity to emotional prosody in congenital amusia rekindles the musical protolanguage hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19027–19032. doi: 10.1073/pnas.1210344109

Toivonen, I., Blumenfeld, L., Gormley, A., Hoiting, L., Lo-gan, J., Ramlakhan, N., and Stone, A. (2015) "Vowel height and duration," *Proceedings of the 32nd West Coast Conference on Formal Linguistics*.

Traunmüller, H. (1986). "Some aspects of the sound of speech sounds," in *The Psychophysics of Speech Perception*, ed M. E. H. Schouten (Dordrecht: Martinus Nijhoff), 293–305. doi: 10.1007/978-94-009-3629-4_24

Tsunoda, T. (1985). *The Japanese Brain.* Tokyo: Taishukan.

Ultan, R. (1978). "Size-sound symbolism," in *Universals of Human Language: Phonology*, eds J. Greenberg (Stanford, CA: Stanford University Press).

Vainio, L. (2021). Magnitude sound symbolism influences vowel production. *J. Memory Lang.* 118:104213. doi: 10.1016/j.jml.2020.104213

Wey, Y. (2019). *Transkription wortloser Gesänge*. Innsbruck: Innsbruck University Press. doi: 10.15203/3187-81-8

Whalen, D. H., Levitt, A. G., Hsiao, P.-L., and Smorodinsky, I. (1995). Intrinsic F0 of vowels in the babbling of 6-, 9- and 12-month-old French-and English-learning infants. *J. Acoustical Soc. Am.* 97, 2533–39. doi: 10.1121/1.411973

Winter, B., and Perlman, M. (2021). Size sound symbolism in the English lexicon. *Glossa J. Gen. Linguist.* 6, 1–13. doi: 10.5334/gjgl.1646

Zhang, C., Shao, J., Huang, X. (2017). Deficits of congenital amusia beyond pitch: Evidence from impaired categorical perception of vowels in Cantonese-speaking congenital amusics. *PLoS ONE* 12:e0183151. doi: 10.1371/journal.pone.0183151

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

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: frontiersin.org/about/contact

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