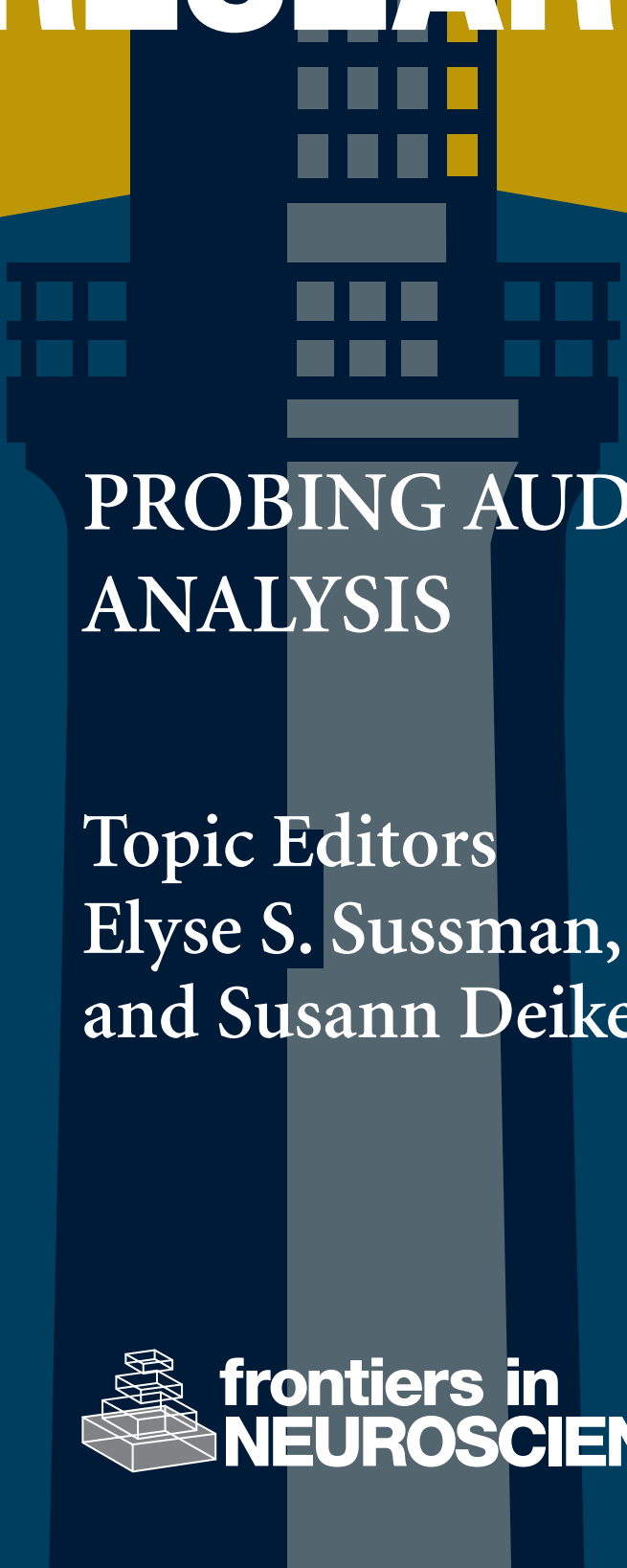
frontiers
RESEARCH TOPICS

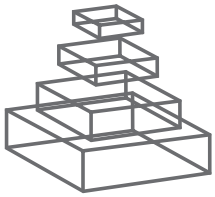


PROBING AUDITORY SCENE ANALYSIS

Topic Editors
Elyse S. Sussman, Susan Denham
and Susann Deike

frontiers in NEUROSCIENCE

frontiers in PSYCHOLOGY



frontiers

Cover image provided by Ibbl sarl, Lausanne CH

ISSN 1664-8714
ISBN 978-2-88919-371-4
DOI 10.3389/978-2-88919-371-4

ABOUT FRONTIERS

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

FRONTIERS JOURNAL SERIES

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

DEDICATION TO QUALITY

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

WHAT ARE FRONTIERS RESEARCH TOPICS?

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

PROBING AUDITORY SCENE ANALYSIS

Topic Editors:
Elyse S. Sussman, Albert Einstein College of Medicine, USA
Susan Denham, University of Plymouth, United Kingdom
Susann Deike, Leibniz Institute for Neurobiology, Germany

In natural environments, the auditory system is typically confronted with a mixture of sounds originating from different sound sources. As sounds spread over time, the auditory system has to continuously decompose competing sounds into distinct meaningful auditory objects or "auditory streams" referring to certain sound sources. This decomposition work, which was termed by Albert Bregman as "Auditory scene analysis" (ASA), involves two kinds of grouping to be done. Grouping based on simultaneous cues, such as harmonicity and on sequential cues, such as similarity in acoustic features over time. Understanding how the brain solves these tasks is a fundamental challenge facing auditory scientist. In recent years, the topic of ASA was broadly investigated in different fields of auditory research, including a wide range of methods, studies in different species, and modeling. Despite the advance in understanding ASA, it still proves to be a major challenge for auditory research. This includes verifying whether experimental findings are transferable to more realistic auditory scenes.

A central approach in understanding ASA is the use of certain stimulus parameters that produce an ambiguous percept. The advantage of such an approach is that different perceptual organizations can be studied without varying physical stimulus parameters. Additionally, the perception of ambiguous stimuli can be volitionally controlled by intention or task. By using this one can mirror real hearing situations where listeners intent to identify and to localize auditory sources. Recently it was also found that in classical auditory streaming sequences perceptual ambiguity was not restricted to but was observed over a broad range of stimulus parameters.

The proposed Research Topic pursues to bring together scientist in the different fields of auditory research whose work addresses the issue of perceptual ambiguity. Researchers were welcome to contribute experimental reports, computational modeling, and reviews that consider auditory ambiguity in its modality specific characteristics as well as in comparison to visual ambiguous figures. The overall goal of contributions was to consider the experimental findings from the perspective of real auditory scenes. In a broader sense, the Research Topic was open for contributions which are related to the issue of active listening in complex scenes.

Table of Contents

frontiers in
NEUROSCIENCE

EDITORIAL
published: 12 September 2014
doi: 10.3389/fnins.2014.00293

Probing auditory scene analysis

Susann Deike[1]*, Susan L. Denham[2,3] and Elyse Sussman[4,5]

[1] *Special Lab Non-Invasive Brain Imaging, Leibniz Institute for Neurobiology, Magdeburg, Germany*
[2] *Cognition Institute, University of Plymouth, Plymouth, UK*
[3] *School of Psychology, University of Plymouth, Plymouth, UK*
[4] *Department of Neuroscience, Albert Einstein College of Medicine of Yeshiva University, Bronx, NY, USA*
[5] *Department of Otorhinolaryngology-Head and Neck Surgery, Albert Einstein College of Medicine of Yeshiva University, Bronx, NY, USA*
**Correspondence: sdeike@lin-magdeburg.de*

Edited and reviewed by:
Isabelle Peretz, Université de Montréal, Canada

Keywords: auditory scene analysis, multistable perception, ambiguity, realistic auditory scenes, stream segregation

In natural environments, the auditory system is typically confronted with a mixture of sounds originating from different sound sources. The sounds emanating from different sources can overlap each other in time and feature space. Thus, the auditory system has to continuously decompose competing sounds into distinct meaningful auditory objects or "auditory streams" associated with the possible sound sources. This decomposition of the sounds, termed "Auditory scene analysis" (ASA) by Bregman (1990), involves two kinds of grouping. Grouping based on simultaneous cues (e.g., harmonicity) and on sequential cues (e.g., similarity of acoustic features over time). Understanding how the brain solves these tasks is a fundamental challenge facing auditory scientists. In recent years, the topic of ASA was broadly investigated in different fields of auditory research using a wide range of methods, including studies in different species (Hulse et al., 1997; Fay, 2000; Fishman et al., 2001; Moss and Surlykke, 2001), and computer modeling of ASA (for recent reviews see, Winkler et al., 2012; Gutschalk and Dykstra, 2014). Despite advances in understanding ASA, it still proves to be a major challenge for auditory research, especially in verifying whether experimental findings are transferable to more realistic auditory scenes. This special issue is a collection of 10 research papers and one review paper providing a snapshot of current ASA research. The research paper on visual perception provides a comparative view of modality specific as well as general characteristics of perception.

One approach for understanding ASA in real auditory scenes is the use of stimulus parameters that produce an ambiguous percept (cf. Pressnitzer et al., 2011). The advantage of such an approach is that different perceptual organizations can be studied without varying physical stimulus parameters. Using a visual ambiguous stimulus and combining real-time functional magnetic resonance imaging and machine learning techniques, Reichert et al. (2014) showed that it is possible to determine the momentary state of a subject's conscious percept from time resolved BOLD-activity. The high classification accuracy of this data-driven classification approach may be particularly useful for auditory research investigating perception in continuous, ecologically-relevant sound scenes.

A second advantage in using ambiguous stimuli in experiments on ASA is that perception of them can be influenced by intention or task (Moore and Gockel, 2002). By manipulating task requirements one can mirror real hearing situations where listeners often need to identify and localize sound sources. The studies by Shestopalova et al. (2014) and Kondo et al. (2014) examined the influence of motion on stream segregation. In general, and corresponding to earlier findings, both of these studies found that sound source separation in space promoted segregation. Surprisingly, however, the effect of spatial separation on stream segregation was found to be temporally limited and affected by volitional head motion (Kondo et al., 2014), but unaffected by movement of sound sources or by the presentation of movement-congruent visual cues (Shestopalova et al., 2014). Another study, by Sussman-Fort and Sussman (2014), investigated the influence of stimulus context on the buildup of stream segregation. They found that the build-up of stream segregation was context-dependent, occurring faster under constant than varying stimulus conditions. Based on these findings the authors suggested that the auditory system maintains a representation of the environment that is only updated when new information indicates that reanalyzing the scene is necessary. Two further studies examined the influence of attention on stream segregation. Nie et al. (2014) found that in conditions of weak spectral contrast, attention facilitated stream segregation. Shuai and Elhilali (2014) found that different forms of attention, both stimulus-driven and top-down attentional processes, modulated the response to a salient event detected within a sound stream.

The special issue also includes two research papers that extend current views on multistability and perceptual ambiguity. The psychophysical study by Denham et al. (2014) showed that streaming sequences could be perceived in many more ways than in the traditionally assumed (Integrated vs. Segregated organizations) and that the different interpretations continuously compete for dominance. Moreover, despite being highly stochastic, the switching patterns of individual participants could be distinguished from those of others. Hence, perceptual multistability can be used to characterize both general mechanisms and individual differences in human perception. By comparing stimulus conditions that promote one perceptual organization with those causing an ambiguous percept Dollezal et al. (2014) found specific BOLD responses for the

ambiguous condition in higher cognitive areas (i.e., posterior medial prefrontal cortex and posterior cingulate cortex). Both of these regions were associated with cognitive functions, monitoring decision uncertainty (Ridderinkhof et al., 2004) and being involved when higher task demands were imposed (Raichle et al., 2001; Dosenbach et al., 2007), respectively. This suggests that perceptual ambiguity may be characterized by uncertainty regarding the appropriate perceptual organization, and by higher cognitive load due to this uncertainty.

A second group of research papers within this special issue focused on understanding hearing deficits in older listeners and cochlear implant (CI) users. Gallun et al. (2013) demonstrated that listeners could be categorized in terms of their ability to use spatial and spectrotemporal cues to separate competing speech streams. They showed that the factor of age substantially reduced spatial release from masking, supporting the hypothesis that aging, independent of an individual's hearing threshold, can result in changes in the cortical and/or subcortical structures essential for spatial hearing. Divenyi (2014) compared the signal to noise (S/N) ratio at which normal hearing young and elderly listeners were able to discriminate single formant dynamics in vowel-analog streams and found that elderly listeners required a 15 and 20 dB larger S/N ratio than younger listeners. Since formant transitions represent potent cues for speech intelligibility, this result may at least partially explain the well-documented intelligibility loss of speech in babble noise by the elderly. Böckmann-Barthel et al. (2014) pursued the question whether the time course of auditory streaming differs between normal-hearing listeners and CI users and found that the perception of streaming sequences was similar in quality between both groups. This similarity may suggest that stream segregation is not solely determined by frequency discrimination, and that CI users do not simply respond to differences between A and B sounds but actually experience the phenomenon of stream segregation.

The review by Bendixen (2014) suggests predictability as a cue for sound source decomposition. Bendixen collected empirical evidence spanning issues of predictive auditory processing, predictive processing in ASA, and methodological aspects of measuring ASA. As a result, and as a theoretical framework, an analogy with the old-plus-new heuristic for grouping simultaneous acoustic signals was proposed.

Taken together, this special issue provides a comprehensive summary of current research in ASA, relating the approaches and experimental findings to natural listening conditions. It would be highly desirable in future research on ASA to use more natural stimuli and to test the ecological validity of these findings. With this special issue we hope to raise awareness of this issue.

ACKNOWLEDGMENT

We thank the authors for their contributions and the reviewers for their useful comments. Susann Deike was funded by the National Institutes of Health (DC004263).

REFERENCES

Bendixen, A. (2014). Predictability effects in auditory scene analysis: a review. *Front. Neurosci.* 8:60. doi: 10.3389/fnins.2014.00060

Böckmann-Barthel, M., Deike, S., Brechmann, A., Ziese, M., and Verhey, J. L. (2014). Time course of auditory streaming: do CI users differ from normal-hearing listeners? *Front. Psychol.* 5:775. doi: 10.3389/fpsyg.2014.00775

Bregman, A. S. (1990). *Auditory Scene Analysis. The Perceptual Organization of Sound*. Cambridge: MIT Press.

Denham, S., Bohm, T. M., Bendixen, A., Szalardy, O., Kocsis, Z., Mill, R., et al. (2014). Stable individual characteristics in the perception of multiple embedded patterns in multistable auditory stimuli. *Front. Neurosci.* 8:25. doi: 10.3389/fnins.2014.00025

Divenyi, P. (2014). Decreased ability in the segregation of dynamically changing vowel-analog streams: a factor in the age-related cocktail-party deficit? *Front. Neurosci.* 8:144. doi: 10.3389/fnins.2014.00144

Dollezal, L. V., Brechmann, A., Klump, G. M., and Deike, S. (2014). Evaluating auditory stream segregation of SAM tone sequences by subjective and objective psychoacoustical tasks, and brain activity. *Front. Neurosci.* 8:119. doi: 10.3389/fnins.2014.00119

Dosenbach, N. U., Fair, D. A., Miezin, F. M., Cohen, A. L., Wenger, K. K., Dosenbach, R. A., et al. (2007). Distinct brain networks for adaptive and stable task control in humans. *Proc. Natl. Acad. Sci. U.S.A.* 104, 11073–11078. doi: 10.1073/pnas.0704320104

Fay, R. R. (2000). Spectral contrasts underlying auditory stream segregation in goldfish (*Carassius auratus*). *J. Assoc. Res. Otolaryngol.* 1, 120–128. doi: 10.1007/s101620010015

Fishman, Y. I., Reser, D. H., Arezzo, J. C., and Steinschneider, M. (2001). Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. *Hear. Res.* 151, 167–187. doi: 10.1016/S0378-5955(00)00224-0

Gallun, F. J., Diedesch, A. C., Kampel, S. D., and Jakien, K. M. (2013). Independent impacts of age and hearing loss on spatial release in a complex auditory environment. *Front. Neurosci.* 7:252. doi: 10.3389/fnins.2013.00252

Gutschalk, A., and Dykstra, A. R. (2014). Functional imaging of auditory scene analysis. *Hear. Res.* 307, 98–110. doi: 10.1016/j.heares.2013.08.003

Hulse, S. H., MacDougall-Shackleton, S. A., and Wisniewski, A. B. (1997). Auditory scene analysis by songbirds: stream segregation of birdsong by European starlings (*Sturnus vulgaris*). *J. Comp. Psychol.* 111, 3–13. doi: 10.1037/0735-7036.111.1.3

Kondo, H. M., Toshima, I., Pressnitzer, D., and Kashino, M. (2014). Probing the time course of head-motion cues integration during auditory scene analysis. *Front. Neurosci.* 8:170. doi: 10.3389/fnins.2014.00170

Moore, B. C., and Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acustica United with Acustica* 88, 320–332.

Moss, C. F., and Surlykke, A. (2001). Auditory scene analysis by echolocation in bats. *J. Acoust. Soc. Am.* 110, 2207–2226. doi: 10.1121/1.1398051

Nie, Y., Zhang, Y., and Nelson, P. B. (2014). Auditory stream segregation using bandpass noises: evidence from event-related potentials. *Front. Neurosci.* 8:277. doi: 10.3389/fnins.2014.00277

Pressnitzer, D., Suied, C., and Shamma, S. A. (2011). Auditory scene analysis: the sweet music of ambiguity. *Front. Hum. Neurosci.* 5:158. doi: 10.3389/fnhum.2011.00158

Raichle, M. E., Macleod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proc. Natl. Acad. Sci. U.S.A.* 98, 676–682. doi: 10.1073/pnas.98.2.676

Reichert, C., Fendrich, R., Bernarding, J., Tempelmann, C., Hinrichs, H., and Rieger, J. W. (2014). Online tracking of the contents of conscious perception using real-time fMRI. *Front. Neurosci.* 8:116. doi: 10.3389/fnins.2014.00116

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., and Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science* 306, 443–447. doi: 10.1126/science.1100301

Shestopalova, L., Bohm, T. M., Bendixen, A., Andreou, A. G., Georgiou, J., Garreau, G., et al. (2014). Do audio-visual motion cues promote segregation of auditory streams? *Front. Neurosci.* 8:64. doi: 10.3389/fnins.2014.00064

Shuai, L., and Elhilali, M. (2014). Task-dependent neural representations of salient events in dynamic auditory scenes. *Front. Neurosci.* 8:203. doi: 10.3389/fnins.2014.00203

Sussman-Fort, J., and Sussman, E. (2014). The effect of stimulus context on the buildup to stream segregation. *Front. Neurosci.* 8:93. doi: 10.3389/fnins.2014.00093

Winkler, I., Denham, S., Mill, R., Bohm, T. M., and Bendixen, A. (2012). Multistability in auditory stream segregation: a predictive coding view. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 1001–1012. doi: 10.1098/rstb.2011.0359

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

Received: 18 August 2014; accepted: 27 August 2014; published online: 12 September 2014.

Citation: Deike S, Denham SL and Sussman E (2014) Probing auditory scene analysis. Front. Neurosci. ***8****:293. doi: 10.3389/fnins.2014.00293*

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

frontiers in
NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 23 December 2013
doi: 10.3389/fnins.2013.00252

Independent impacts of age and hearing loss on spatial release in a complex auditory environment

Frederick J. Gallun[1,2]*, Anna C. Diedesch[3], Sean D. Kampel[1] and Kasey M. Jakien[2]

[1] *Department of Veterans Affairs, National Center for Rehabilitative Auditory Research, Portland VA Medical Center, Portland, OR, USA*
[2] *Otolaryngology/Head and Neck Surgery, Oregon Health and Science University, Portland, OR, USA*
[3] *Hearing and Speech Sciences, Vanderbilt University, Nashville, TN, USA*

Edited by:
Elyse S. Sussman, Albert Einstein College of Medicine, USA

Reviewed by:
Olli K. Aaltonen, University of Helsinki, Finland
Piers Dawes, University of Manchester, UK

****Correspondence:***
Frederick J. Gallun, Department of Veterans Affairs, National Center for Rehabilitative Auditory Research, Portland VA Medical Center, 3710 SW US Veterans Hospital Road, Portland, OR 97239, USA
e-mail: Frederick.Gallun@va.gov

Listeners in complex auditory environments can benefit from the ability to use a variety of spatial and spectrotemporal cues for sound source segregation. Probing these abilities is an essential part of gaining a more complete understanding of why listeners differ in navigating the auditory environment. Two fundamental processes that can impact the auditory systems of individual listeners are aging and hearing loss. One difficulty with uncovering the independent effects of age and hearing loss on spatial release is the commonly observed phenomenon of age-related hearing loss. In order to reveal the effects of aging on spatial hearing, it is essential to develop testing methods that reduce the influence of hearing loss on the outcomes. The statistical power needed for such testing generally requires a larger number of participants than can easily be tested using traditional behavioral methods. This work describes the development and validation of a rapid method by which listeners can be categorized in terms of their ability to use spatial and spectrotemporal cues to separate competing speech streams. Results show that when age and audibility are not covarying, age alone can be shown to substantially reduce spatial release from masking. These data support the hypothesis that aging, independent of an individual's hearing threshold, can result in changes in the cortical and/or subcortical structures essential for spatial hearing.

Keywords: spatial release, aging, hearing loss, virtual spatial array

INTRODUCTION

Because a listener relies on multiple spatial and spectrotemporal cues to accurately segregate sound sources, there are many neural processes that support auditory scene analysis. Two of the most important acoustical cues, differences in the pitches and spatial locations of the sound sources, depend on accurate neural timing for these cues to be completely encoded by the central auditory system (Moore, 2007). However, reductions in the accuracy of neural timing may have little impact on an individual's ability to detect energy at a particular frequency. This may explain why an audiogram conducted in quiet over headphones does not correlate well with an individual's ability to use temporal cues (Durlach et al., 1981; Buus et al., 1984; Smoski and Trahiotis, 1986; Gabriel et al., 1992; Koehnke et al., 1995; Lacher-Fougere and Demany, 2005; Ross et al., 2007; Strelcyk and Dau, 2009; Grose and Mamo, 2010; Ruggles et al., 2011). An additional issue that makes prediction of suprathreshold hearing abilities difficult is the presence of age-related hearing loss in many of the older participants. While it is possible that hearing loss and aging both reduce neural timing, animal studies (e.g., Caspary et al., 1995, 2008; Helfert et al., 1999; Wang et al., 2009; Scheidt et al., 2010; Walton, 2010) suggest that the two may act at different levels of the auditory system. For example, the impacts of hearing loss may be more related to reductions in the response of the auditory nerve, while the impacts of aging may be more related to impaired temporal coding in the auditory brainstem. This implies that it is very important to develop methods by which the impacts of aging can be revealed even in the presence of hearing loss.

One of the reasons for limited evidence in humans of the distinction between the neural processes associated with aging and those associated with hearing loss (beyond the limited ability of researchers to make direct measurements of the auditory nerve and brainstem in humans) is that the age of the listener has often been allowed to covary with hearing loss. Studies that have controlled for age effects (e.g., Hawkins and Wightman, 1980; Dubno et al., 2002, 2008) have found that spatial hearing can be substantially reduced in listeners with hearing loss, but to date no studies have effectively controlled for hearing loss, primarily due to issues of sample size.

One of the key abilities associated with binaural processing is spatial release from masking, which occurs when the ability to detect or identify a target sound (often speech) in the presence of one or more masking sounds is improved by spatial differences between target and masker(s) (Carhart et al., 1969; Hawley et al., 2004; Best et al., 2005; Gallun et al., 2005; Brungart and

Portions of this research were presented at the 2011 and 2013 Midwinter Meetings of the Association for Research in Otolaryngology, Baltimore, MD, as well as at the 2013 International Conference of Acoustics in Montreal, Quebec, CA and preliminary analyses of the first experiment were published in the Proceedings of Meetings in Acoustics (Gallun and Diedesch, 2013).

Simpson, 2007). While it has been fairly well established than the benefit obtained by hearing-impaired listeners in a multitalker segregation task is less than that obtained by normally-hearing listeners (Gelfand et al., 1988; Arbogast et al., 2005; Best et al., 2010; Hopkins and Moore, 2010), the impacts of age on spatial release from masking are less understood. In one of the best controlled studies with the largest sample of subjects, Marrone et al. (2008) were unable to clearly distinguish the effects of age from the effects of hearing loss due to a lack of statistical power. The experiments described here were designed to overcome the statistical lack of power common in previous reports in order to provide the strongest test possible of the hypothesis that aging, independent of hearing loss, can result in significant reductions in spatial release from masking.

EXPERIMENT ONE: ADAPTIVE TRACKS; ANECHOIC CHAMBER

The goal of the first experiment was to document the ways in which the ages and hearing levels of the participants interacted with the spatial and pitch cues available in the stimuli to produce a particular level of speech identification performance.

MATERIALS AND METHODS

Spatial release from masking was tested using stimuli drawn from the Coordinate Response Measure (CRM; Bolia et al., 2000), which consists of a simple corpus of sentences with the form "Ready (CALL SIGN) go to (COLOR) (NUMBER) now." There are eight possible call signs (Arrow, Baron, Charlie, Eagle, Hopper, Laker, Ringo, Tiger), and 12 keywords: four colors (red, green, white, and blue) and the numbers 1–8. All possible combinations of the call signs, colors, and numbers are spoken by four male and four female talkers. The CRM corpus consists of high-quality recordings of each of the eight talkers saying all 256 possible combinations of call signs and keywords. The intelligibility of each of the keywords as spoken by each of the talkers was examined by Brungart (2001), who showed that there is no significant advantage to the listener of being asked to identify any one of the potential keywords.

Each listener was presented with a set of three simultaneous utterances from the CRM corpus and the goal was to attend to one of the sentences, identified by the callsign "Charlie." Each sentence was presented from one of seven loudspeakers arranged at 15° separation in front of the listener in an anechoic chamber containing a ring of 24 loudspeakers surrounding the listener, each at approximately head height and at a distance of 1 m from the listener's head. Four spatial configurations were used: colocated (all three sentences presented from 0° azimuth), 15° separation (target at 0°, one masker at + 15° and another masker at −15°), 30° separation (target at 0°, maskers at ±30°), and 45° separation (target at 0°, maskers at ±45°). In addition, each spatial condition was tested in four talker-gender combination conditions: male/male (male target, male maskers), male/female (male target, female maskers), female/female (female target, female maskers), and female/male (female target, male maskers).

Identification thresholds, in terms of target-to-masker ratio (TMR), were estimated based on the average of four adaptive tracks for each of the 16 combinations of four spatial and four talker-gender conditions. In the masked conditions the target level was fixed at 50 dB SPL and the masker levels were adaptively varied, using a one-up/one-down procedure (estimating 50% correct; Levitt, 1971). Masker sentences each started at a level of 40 dB and were changed in level by 5 dB until three reversals were obtained, at which point the step size was changed to 1 dB and six more reversals were measured. The threshold was estimated to be the average of the last six reversals. TMR is expressed as the difference in level between the target and each masker. Thus, if the target level is 50 dB SPL and each masker is also 50 dB SPL, the TMR is 0 dB[1].

To ensure audibility of the target sentence and familiarize listeners with the testing procedure, the first test session started with a single threshold estimate of "quiet threshold," which was defined as the level supporting 50% identification of a male target presented alone at 0° azimuth. Thresholds were obtained using one-up/one-down tracking and a three-stage tracking method. Initially, the level was set at 50 dB SPL and reduced in level by 10 dB until a single error was made, at which point the level was increased in 5dB steps until the sentence was correctly identified, after which six reversals were measured using a 1dB step size, and the threshold was the average of these last six reversal values. No listeners were tested for whom this initial threshold exceeded 45 dB SPL.

Responses were obtained using a touch screen monitor located on a stand within arm's reach of the listener seated in the middle of the anechoic chamber. The monitor was directly in front of the listener but below the plane of the loudspeakers. The listener initiated each adaptive track and indicated the color and number keywords associated with the target callsign "Charlie" by selecting a virtual button with the appropriate colored number from among a grid of the 32 possible color/number combinations. The names of the colors were written to the side of the rows of colored buttons to ensure that colorblind participants could use the interface. Feedback was given after each presentation in the form of "Correct" or "Incorrect." Approximately one second of silence followed the response being registered, prior to the next stimulus presentation.

All participants completed the testing in at least three sessions of no more than 2 h each. Participants were encouraged to take breaks and initiate each run only when they felt ready, which resulted in some variability in the total test time, especially between the oldest and youngest listeners. All procedures were approved by the Portland VA Medical Center Institutional Review Board and all participants gave written consent and were monetarily compensated for their time.

[1]Note that this is different from the signal-to-noise ratio, which is the ratio of the target to the combined level of the two maskers, which in this case would be −3 dB, due to the uncorrelated signals in the two maskers. Acoustic measurements of the combined maskers confirmed that the summed signals were, on average, 3 dB higher than either alone. Due to the multiple summation combinations possible at each ear with the various spatial configurations, it is generally preferred to refer to TMR in studies of speech on speech masking when spatial separations are employed.

PARTICIPANTS

Thirty-four listeners participated, varying in age from 25 to 74 years (mean of 50.3 years, standard deviation of 15.6 years). All had hearing thresholds of 50 dB HL or better at octave frequencies of 2 kHz and below in both ears (mean values varied from 11.3 to 18.6 dB HL, with standard deviations of no more than 12.5 dB). At 4 and 8 kHz, greater losses were present (thresholds of up to 95 dB HL in a few listeners), but all listeners had pure-tone averages (0.5, 1, 2, and 4 kHz) that were all below 32 dB HL, and all had fairly symmetrical hearing at 2 kHz and below (no differences exceeding 10 dB at more than one frequency, and no differences exceeding 20 dB at any frequency).

RESULTS

Quiet thresholds were found to vary between 16 and 41.5 dB SPL and were not significantly correlated with age ($R^2 = 0.05$, $p > 0.05$), although quiet speech thresholds were significantly correlated with the average pure-tone thresholds at the two ears. Correlations were stronger for the average of three low- to mid-frequency thresholds (5, 1, 2 kHz: $R^2 = 0.62$, $p < 0.0001$) than for three mid- to high-frequency thresholds (1, 2, 4 kHz: $R^2 = 0.52$, $p < 0.0001$).

Data were analyzed with a within-subjects analysis of variance, (ANOVA; SPSS v.20) in which between-subject variables were entered as covariates. Target (male, female), masker (same gender, different gender), and spatial separation (0, 15, 30, 45°) were entered as within-subjects factors and the between-subjects factors of age and identification threshold in quiet ("quiet threshold") were entered as covariates. Quiet threshold was chosen as the most influential measure of hearing sensitivity based on exploratory stepwise regression with a variety of potential predictors, including individual frequency thresholds and various pure tone averages. Quiet threshold was found to be significantly correlated with target-to-masker threshold in each condition (values ranging from 0.32 to 0.68, $p < 0.05$). Statistically significant effects are shown in **Table 1**. Mean values as a function of target, masker, and spatial separation are shown in **Figure 1**. As shown in **Table 1**, the main effects of each were statistically significant and the variance accounted for (as measured by partial eta-squared) was greater than 35%.

The between-subject factors of age and quiet threshold were both statistically significant and accounted for more than 32% of the variance in TMR. Interactions among the between-subject factors (age and quiet threshold) and the within-subjects factors were non-significant for all two-way combinations with the exception of spatial separation and quiet threshold. The only significant interaction of within-subject factors was the two-way interaction of masker and spatial separation. Both between-subjects factors formed three-way interactions with masker and spatial separation. None of the other three-way interactions or any four-way interactions were significant.

To further examine the cause of the significant interactions, *post-hoc* trend analysis was conducted. The three-way interaction of masker, spatial separation, and age was found to be based on linear effects of all three factors ($p < 0.01$, partial eta-squared = 0.23), as was the case for the three-way interaction of masker, spatial separation, and quiet threshold ($p < 0.05$, partial eta-squared = 0.18). In an attempt to illustrate the interactions among these factors, as well as the basic main effects, which are obvious across multiple combinations of factors, **Figure 2** shows target-to-masker threshold as a function of age for the two masker types at each of the four spatial separations (one per panel, all panels), while the four panels of **Figure 3** demonstrate the same relationships for masker, quiet threshold, and spatial separation.

Table 1 | Within-subjects ANOVA for Experiment One.

	Degrees of freedom	*F*	*p*-value	Partial eta squared
TESTS OF BETWEEN-SUBJECTS EFFECTS				
Quiet threshold	1,31	21.336	<0.001	0.408
Age	1,31	15.062	0.001	0.327
TESTS OF WITHIN-SUBJECTS EFFECTS				
Masker	1,31	29.149	<0.001	0.485
Spatial separation	3,93	28.281	<0.001	0.477
Target	1,31	18.417	<0.001	0.373
Masker ∗ Spatial separation	3,93	20.035	<0.001	0.393
Spatial separation ∗ Quiet threshold	3,93	8.782	<0.001	0.221
Masker ∗ Spatial separation ∗ Age	3,93	5.166	0.002	0.143
Masker ∗ Spatial separation ∗ Quiet threshold	3,93	3.366	0.022	0.098

Between-subjects factors entered as covariates. Greenhouse-Geisser corrections for violations of the assumption of sphericity were conducted for the spatial separation factor (the only within-subjects factor with more than two levels), but the results were unchanged. Statistical interaction effects that failed to reach significance are not shown.

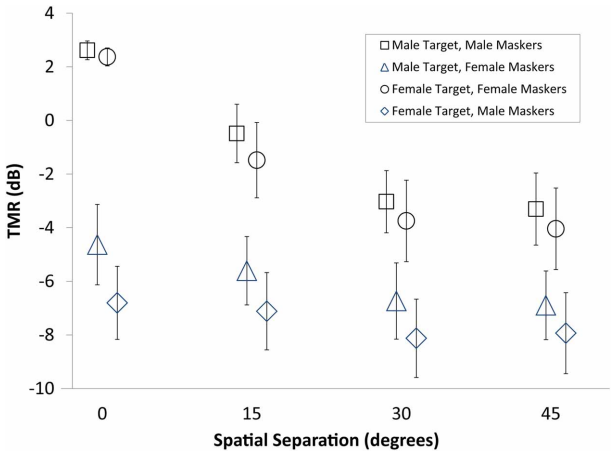


FIGURE 1 | Average thresholds for the subjects in Experiment One. Target-to-masker ratio (TMR) is plotted as a function of spatial separation between targets and maskers for male and female targets identified in the presence of male and female maskers. Spatial separation of 0° corresponds to colocated targets and maskers presented from the same loudspeaker.

DISCUSSION

These results strongly support the hypothesis that age and hearing loss are independently responsible for reduced spatial release from masking. **Figures 2**, **3** demonstrate this by showing that the patterns of results are fairly similar for age and quiet threshold, despite the low correlation the between two factors (less than 5% shared variance). In both cases, colocated TMR values are fairly similar across listeners for the conditions where target and masker are the same gender, but much greater variability is observed when either the genders are different or the spatial separation is greater than 0°. This can be seen most easily by comparing the error bars for the same gender maskers in the colocated condition in **Figure 1** with the error bars for all other conditions.

Both the high variability and the relatively strong relationship with age and/or hearing loss have been observed in recent publications (Marrone et al., 2008; Strelcyk and Dau, 2009; Neher et al., 2011; Glyde et al., 2013), but in each of these previous studies the small sample size, the heterogeneity of the study sample, and/or the complexity of the statistical analyses attempted has resulted in reduced or non-existent support for the hypothesis that aging has an influence on spatial release from masking independent of hearing loss. In each case, multiple regression was conducted and, once the variance accounted for by one factor has been partialled out, no further variance could be explained by the other factor. To examine this issue, multiple stepwise linear regression was conducted on the TMR data from each condition (reported in Gallun and Diedesch, 2013). As in previous work, both factors were significant contributors for many of the conditions. Unlike in previous work, when only one factor accounted for all of the variability, that factor was age. The proportion of variance accounted for by age and hearing loss (alone or in combination) was between 22 and 54%.

EXPERIMENT TWO: ADAPTIVE TRACKS; VIRTUAL SPACE

To improve the methods available for predicting individual sensitivity to spatial cues, the impacts of hearing loss were further minimized by using headphone presentation, which permitted greater control over the audibility of the stimuli. A virtual spatial

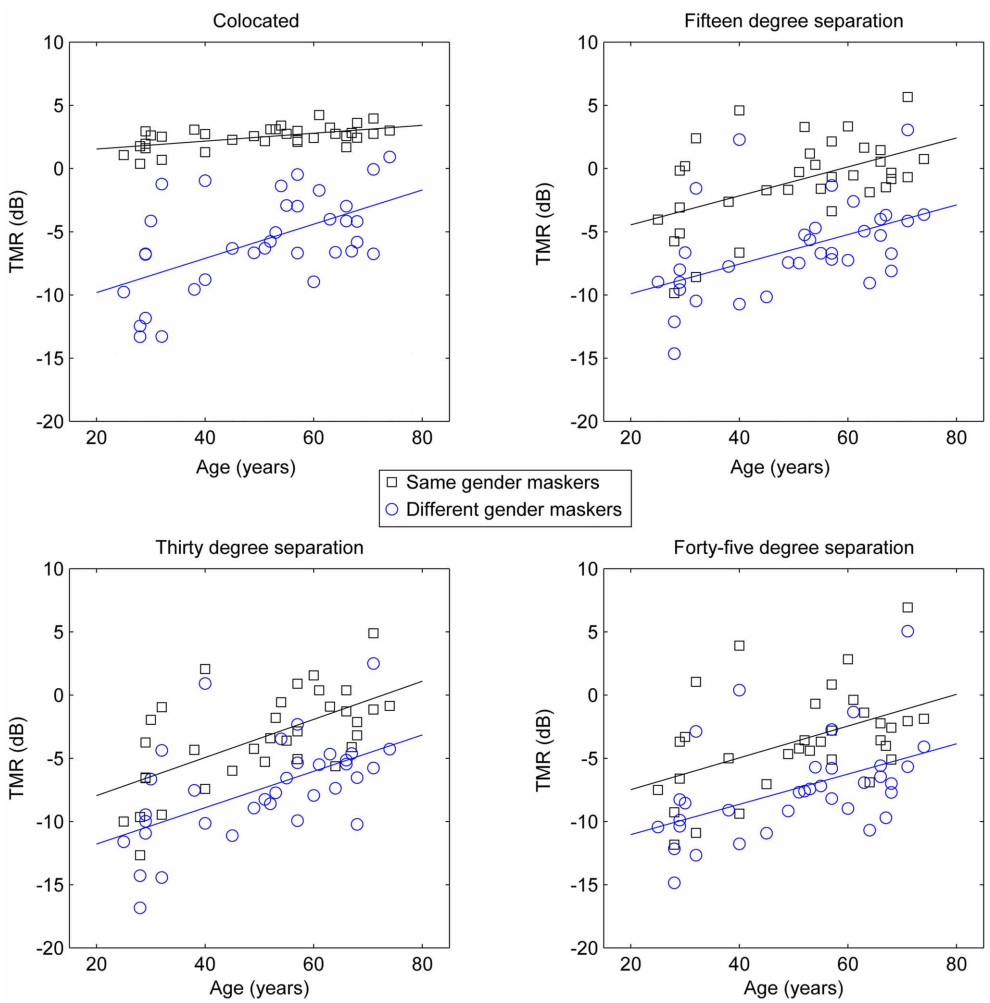


FIGURE 2 | Target-to-masker ratio (TMR) plotted as a function of age for the subjects in Experiment One. Each panel plots same-gender maskers (averaged across target gender) as squares and different-gender maskers as circles.

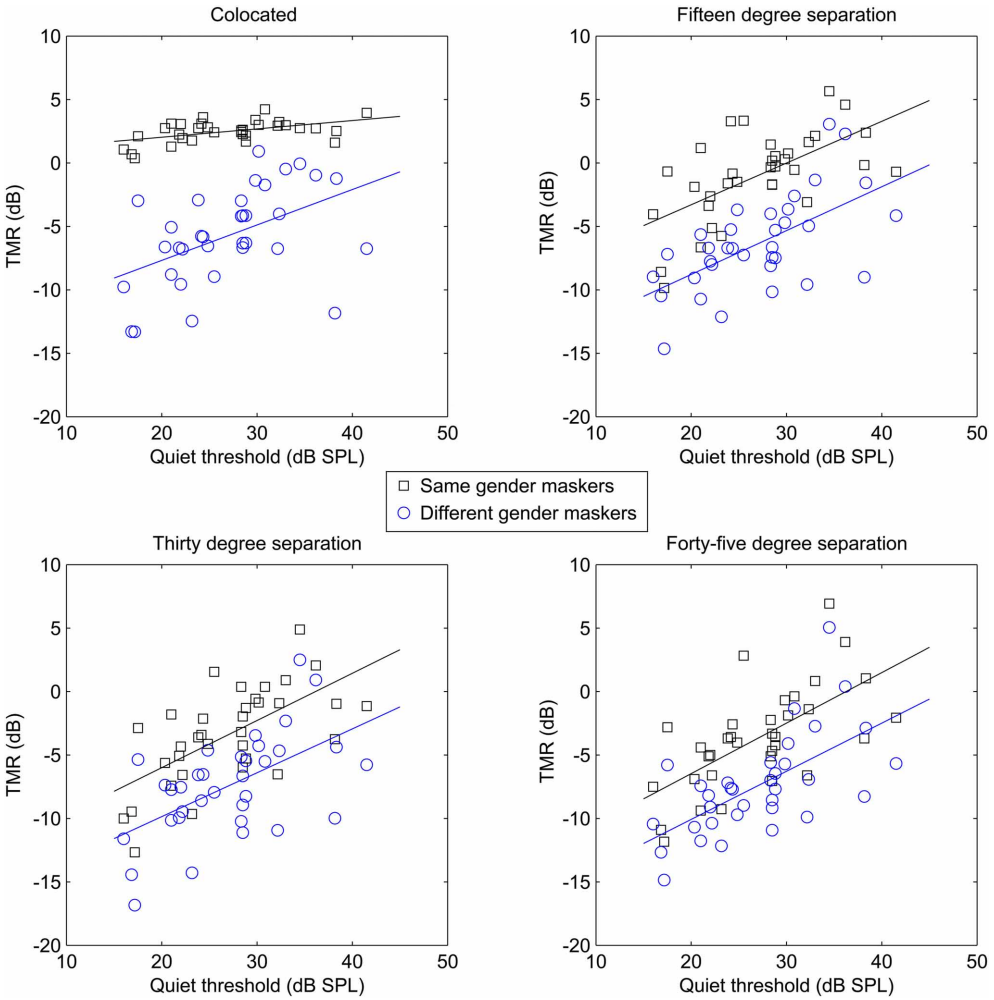


FIGURE 3 | Target-to-masker ratio (TMR) plotted as a function of quiet threshold for the subjects in Experiment One. Each panel plots same-gender maskers (averaged across target gender) as squares and different-gender maskers as circles.

array (VSA) was used to test the hypothesis that spatial release testing can produce the strong effects of aging seen in Experiment One without the use of an anechoic chamber. It was also hypothesized that the impacts of even mild hearing loss would be reduced by equating audibility of the stimuli at the two ears, and that even stronger effects of aging would be revealed.

MATERIALS AND METHODS

The methods were very closely matched to those of Experiment One, with a few notable exceptions. The first difference was the use of a VSA presented over insert earphones (Etymotic ER2). The VSA is a technology that depends on imposing the appropriate binaural time and level differences on a given source signal such that the resulting signals at the ears of the listener will resemble the signals that would occur if the sources had been presented from a loudspeaker at a given location (Xie, 2013). The approach used here was to use head-related impulse response (HRIR) measurements and MATLAB functions available from the Music and Audio Research Laboratory, at New York University (http://marl.smusic.nyu.edu/wordpress/projects/scanir/) and convolve each CRM wavefile to be presented with left ear and right ear HRIRs measured for a binaural manikin. The main differences between the VSA and loudspeaker presentation would have involved the spectral differences among locations and between listeners associated primarily with differences in vertical location. As none of the stimuli were simulated to be emanating from vertical locations off of the horizontal azimuth, it is unlikely that large differences in perceived location were present across listeners. Furthermore, the data from Experiment One indicate that once the spatial separation exceeds 15°, spatial separation only mildly affects performance. Consequently, even if one listener were to perceive a spatial separation of 40° and another listener 50° (which would be fairly unlikely), the impacts on performance might still be too small to detect with these methods.

The second important modification to the methods associated with Experiment One was the use of equal sensation level (SL) signals. This was achieved by first measuring the level at which each listener could just identify speech presented by the

audiologist over the audiometer, transforming that value from hearing level (HL) to dB SPL by adding 22 dB, and then adding 30 dB to that level to obtain the level of the target sentence, which was always fixed during the adaptive tracking procedure. As is shown in the results section below, this approximation for the HL to SPL transform resulted in a range of values in dB SPL that were essentially identical to the range of Quiet CRM values obtained in Experiment One. The two masker sentences were again presented at levels relative to the target, so they were appropriately scaled in SL as well. No listeners were tested for whom the 30 dB SL level would have resulted in maskers that exceeded 85 dB SPL.

In addition, minor changes were made to the protocol to reduce testing time and improve the accuracy of the estimates. Based on the results of the first experiment, performance was evaluated for the colocated (0°) and 45° spatial separation conditions, and the two conditions with female targets were not included. This eliminated roughly three-quarters of the conditions that would have to be tested, thus, allowing the length of each adaptive track to be extended by two reversals (to eight rather than six) and to repeat the entire set of tracks, resulting in eight rather than four adaptive tracks contributing to the estimate for each condition.

Responses were obtained using a monitor located on a table in front of the listener. The listener initiated each adaptive track and indicated the color and number keywords associated with the target callsign "Charlie" by using a mouse to depress a virtual button using the same interface as in Experiment One. Feedback was given after each presentation and the next trial was initiated roughly one second after the response was registered. All participants completed the testing in no more than 2 h, and only occasionally was the testing divided into two sessions. All procedures were approved by the Portland VA Medical Center Institutional Review Board and all participants gave written consent and were monetarily compensated for their time.

PARTICIPANTS

Fifty two listeners participated, varying in age from 19 to 76 years (mean of 45.3 years, standard deviation of 17.0 years). Nine of these individuals had participated in Experiment One and all had recently completed a similar listening experiment also using the CRM sentences with the same spatial separations as were employed in this experiment, thus, all of the listeners were very well trained on the task. All had hearing thresholds of 40 dB HL or better at octave frequencies of 4 kHz and below in both ears (mean values varied from 7 to 14 dB HL, with standard deviations of no more than 11 dB HL). At 8 kHz, greater losses were present (thresholds of up to 80 dB HL in a few listeners), however, the mean thresholds were all below 21 dB HL. All listeners had fairly symmetrical hearing at all audiometric frequencies. None had differences exceeding 10 dB HL at any one frequency or differences exceeding 5 dB HL at more than one frequency. Speech Reception Thresholds (SRTs) were tested by reducing the level of speech until listeners could only correctly report 50% of the words spoken. Thresholds obtained in this standard clinical test varied between 0 and 15 dB HL at the left ear (mean of 8.17 dB HL) and −5 and 20 dB HL (mean of 6.54 dB HL) at the right ear.

RESULTS

SRT values were set for each ear independently and ranged between 17 and 42 dB SPL, which is extremely similar to the range of SPL values found for the quiet speech in Experiment One (16–41.5 dB SPL). Target levels, which were set 30 dB above the SRT, varied from 47 dB SPL to 72 dB SPL, reflecting the 25 dB range of SRT values in the sample. In this case, SRTs at the left and right ears were significantly correlated with age ($R^2 = 0.30$, $p < 0.001$ for the left; $R^2 = 0.10$, $p < 0.002$ for the right) and were again significantly correlated with audiometric thresholds, as would be expected. All correlations were between $R^2 = 0.16$ and 0.81, with $p < 0.05$ for each. These values do not reflect corrections for multiple comparisons, but the majority of the correlations would still have been significant.

A within-subjects ANOVA (SPSS v.20) was used, with between-subject continuous variables entered as covariates. The within-subject factors were spatial separation (0 vs. 45°) and masker (same vs. different gender). Age and SRT (averaged across ears) were between-subject factors entered as covariates. Statistical results are shown in **Table 2**. Mean TMR thresholds for the colocated conditions were 2.0 dB (standard deviation of 1.6 dB) for the same-gender maskers and −6.9 dB (standard deviation of 3.4 dB) for the different-gender maskers. When the same-gender maskers were spatially separated, the mean TMR was −7.8 dB (standard deviation of 3.2 dB) and when the different-gender maskers were spatially separated the mean TMR was −10.8 dB (standard deviation of 3.5 dB). These values are substantially better than what was observed in Experiment One, by as much as 4.5 dB in the case of the same-gender maskers at 45° separation. Possible explanations for this difference are discussed below, but the most likely is the use of the equal SL target, which may have improved performance for many of the hearing-impaired participants.

The main effects of spatial separation and masker type were both statistically significant and the variance explained (measured by partial eta-squared) was greater than 75%. Main effects of age and SRT were also statistically significant, although age accounted

Table 2 | Within-subjects ANOVA for Experiment Two.

	$F_{(1, 49)}$	*p*-value	Partial eta squared
TESTS OF BETWEEN-SUBJECTS EFFECTS			
Age	16.848	<0.001	0.256
SRT	7.37	0.009	0.131
TESTS OF WITHIN-SUBJECTS EFFECTS			
Spatial separation	300.015	<0.001	0.86
Masker	172.016	<0.001	0.778
Spatial separation * Masker	80.01	<0.001	0.62
Masker * SRT	10.454	0.002	0.176
Spatial separation * SRT	7.937	0.007	0.139
Spatial separation * Masker * Age	4.414	0.041	0.083
Spatial separation * Age	4.031	0.05	0.076

Between-subjects factors entered as covariates. Statistical interaction effects that failed to reach significance are not shown.

for 26% of the variance while SRT accounted for only 13%, again likely due to the use of equal SL stimuli. **Figure 4** illustrates these main effects by plotting TMR for the four conditions as a function of age and SRT.

Three interactions among the within and between-subjects factors were able to account for greater than 10% of the variance: the spatial separation by masker interaction, the SRT by masker interaction, and the SRT by spatial separation interaction. **Figure 4** also illustrates these interactions.

DISCUSSION

The results of Experiment Two also support the hypothesis that age reduces spatial release from masking independently of hearing loss. Indeed, age accounted for nearly twice the variance than did SRT, suggesting that equating the audibility of the stimuli across listeners and across ears allowed the effects of age to be even more apparent. Furthermore, these results indicate that the use of HRIRs measured with a binaural manikin allows the creation of a virtual acoustic space with sufficient veridicality to capture the same amounts of spatial release and between-subject differences in performance as were found in the anechoic chamber presentation used in Experiment One. In fact, listener performance was better in the virtual acoustic space, both in terms of the average thresholds and in terms of the best performance achieved. This is likely due to two factors: the better overall hearing of the participants (PTAs for the frequencies 0.5, 1, and 2 kHz were 15.1 dB in Experiment One and 9.4 dB in Experiment Two) and the use of an equal SL target set at 30 dB above individual SRTs for each ear. However, it should be noted that there may have been a learning factor involved, as all of the participants in Experiment Two had already spent several hours performing a slightly different version of this task (not reported here) and thus, had substantially more practice than the listeners in Experiment One. Another possible explanation is that the HRIR convolution introduced additional cues that were of benefit to the listeners, although it is more often the case that actual loudspeaker presentation results in better performance than that found with generic HRIRs.

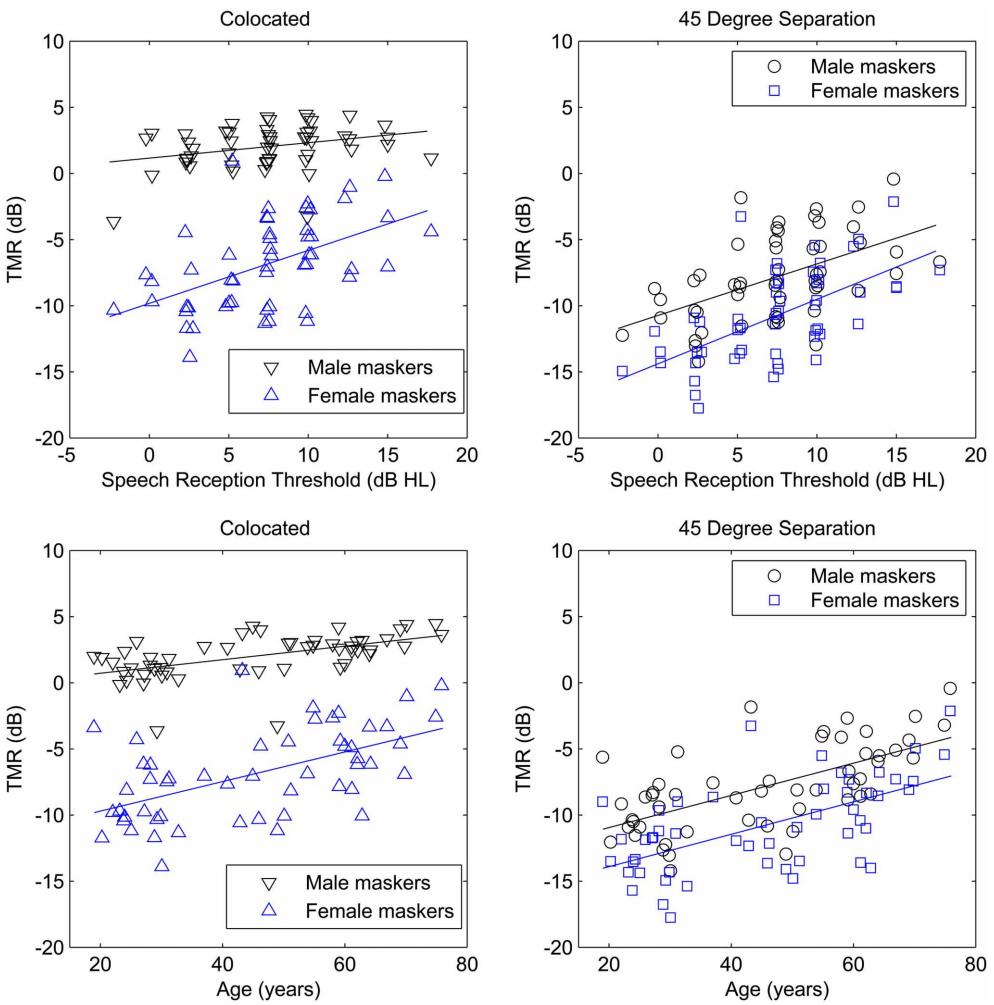


FIGURE 4 | TMR as a function of speech reception threshold (SRT; top panels) and age (bottom panels) for the participants in Experiment Two. Thresholds for male and female maskers are plotted separately for colocated targets (left panels) and spatially separated targets (right panels). SRT values are offset slightly to improve visualization.

Regardless of the reason for the improved performance associated with headphone presentation, it is clear from these data that older listeners perform less well, on average, in these tasks than do their younger counterparts, especially when the effects of hearing are sufficiently controlled through subject recruitment, sample size, and amplification of the stimuli. In addition, these data show that the VSA can support very good performance in a spatial release from masking task.

EXPERIMENT THREE: PROGRESSIVE TRACKS; VIRTUAL SPACE

The final experiment was designed to examine whether or not a rapid presentation of fixed TMRs could be used to reveal the same effects of age, independent of hearing loss, as were shown in the first two experiments with longer adaptive tracks. The hypothesis to be tested was that the effects of age and spatial separation were both so great that even a very rapid test with a minimal number of experimental trials could reveal the same pattern of performances across listeners as was found with more traditional methods. This would provide evidence both for the strength of the aging effects and for the utility of a rapid testing procedure that could be used either in combination with a larger battery of tests or could even be adapted for clinical use.

MATERIALS, METHODS, AND PARTICIPANTS

To reduce testing time, only the condition involving a male target with male maskers was examined. This was justified by the very similar results that were obtained for the two same-gender masking conditions, but removes the different-gender masking condition. This parameter (same- vs. different-gender) was not found to be as sensitive to the effects of aging and hearing loss as was the spatial separation parameter, and the interaction between gender of the masker and spatial separation was also not found to produce particularly large effects. Consequently, it was decided that comparing the colocated same-gender maskers with the spatially-separated same-gender maskers would provide the strongest test of the suitability of the progressive track for identifying differences across listeners.

The "progressive" tracking procedure involved the presentation of 20 trials, two at each of 10 TMR values, starting with 10 dB TMR and decreasing in steps of 2 dB until reaching a level of −8 dB TMR. This set of fixed TMRs would allow the stimuli to be designed in advance and presented via a fixed-track sound system such as a CD player or smart phone (provided the sound fidelity was appropriately verified). In this experiment, however, the tracks were still randomized for each participant and presented using the same computer-controlled audio system used in Experiment Two. Responses were recorded in the identical manner as well. From the perspective of the participant, the only difference was that the task started at a very easy TMR and became progressively more difficult, rather than moving back and forth between being relatively easier and more difficult across the length of the track. Using the VSA, each participant completed one progressive track with colocated maskers followed by one progressive track with maskers spatially separated from the target by plus and minus 45°.

Rather than measuring reversals, as in the adaptive procedure, performance was estimated using the much simpler metric of counting the total number of trials in which the color and number were both reported correctly. This was also used to provide a preliminary threshold TMR estimate by subtracting the number correct from 10 dB. At the extremes, this metric can be shown to be logically appropriate because if the number correct is zero, then the TMR is certainly no less than 10 dB, and if the number correct is 20, then the threshold is likely to be around −10 dB (given that the scale has an accuracy of roughly 2 dB). The appropriateness of this metric was evaluated by comparing the performance of each listener with the threshold estimate for that same listener in Experiment Two. To facilitate this comparison, the same 52 participants from Experiment Two all participated in Experiment Three.

RESULTS

A within -subjects ANOVA (SPSS v.20) was conducted on the estimated thresholds, with the between-subject continuous variables of age and SRT entered as covariates. The estimated threshold is based on the transformation described above, but as it is a linear subtraction, the same statistical results would be obtained from an analysis of the number of correct responses. Thresholds are reported primarily as they are more easily compared with the previous data. Statistical results are reported in **Table 3**.

As opposed to the 16 thresholds obtained in Experiment One and the four in Experiment Two, only 2 thresholds were obtained: the colocated condition (mean of 2.13 dB, sd of 2.2 dB) and the 45° separation (mean of −4.3 dB, sd of 3.8 dB). The colocated threshold is quite similar to the thresholds found in the first two experiments (2.6 and 2.0 dB), while the separated threshold is better than the −3.3 dB found in Experiment One and not as good as the −7.7 dB found in Experiment Two.

The main effects of spatial separation and age were both statistically significant, but the main effect of SRT was not. The interaction of age and spatial separation was significant, but the interaction of SRT and spatial separation was not. The relationships of age and SRT to thresholds in the two conditions are illustrated in the four panels of **Figure 5**.

DISCUSSION

Two aspects of the results of Experiment Three are of particular note. The first is that in this experiment the effect of age

Table 3 | Within-subjects ANOVA for Experiment Three.

	$F_{(1, 49)}$	*p*-value	Partial eta squared
TESTS OF BETWEEN-SUBJECTS EFFECTS			
Age	24.211	<0.001	0.331
SRT	0.022	0.882	0
TESTS OF WITHIN-SUBJECTS EFFECTS			
Spatial separation	55.91	<0.001	0.533
Spatial separation ∗ Age	7.274	0.01	0.129
Spatial separation ∗ SRT	0.33	0.568	0.007

Between-subjects factors entered as covariates.

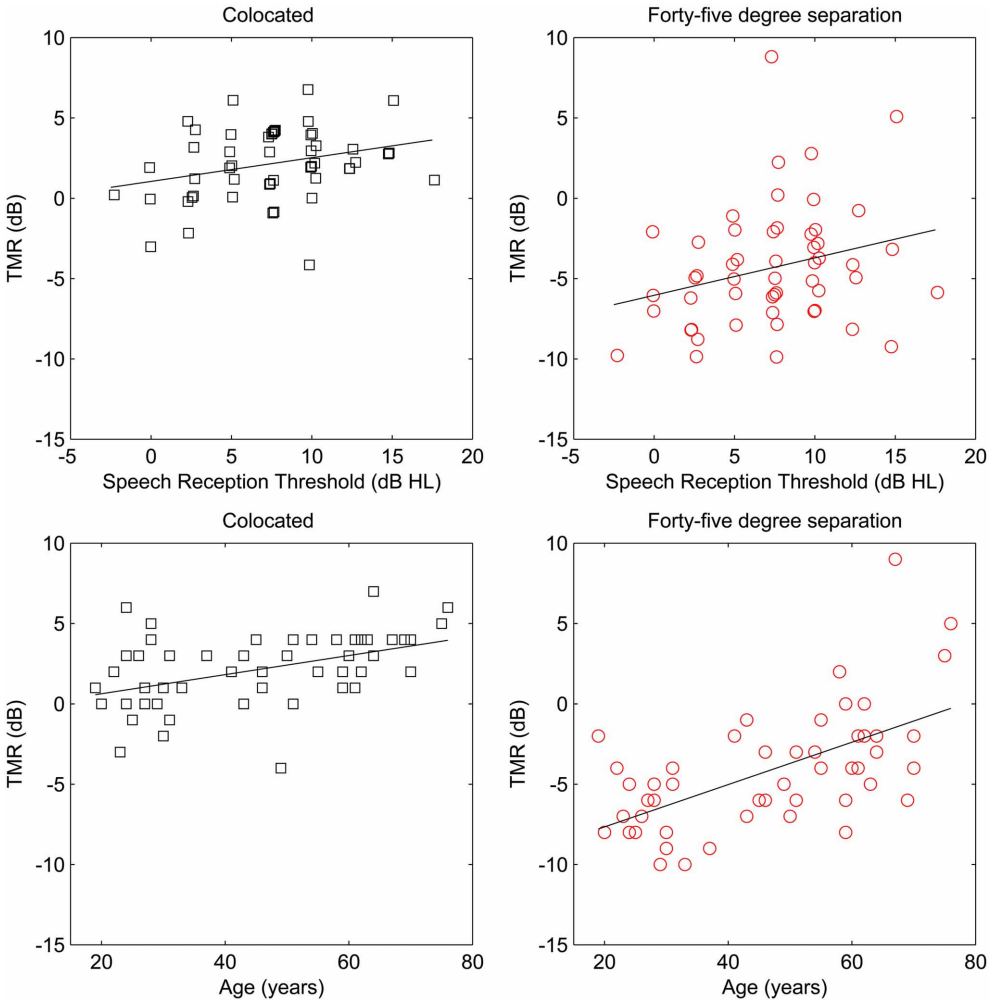


FIGURE 5 | Target-to-masker ratio (TMR) plotted as a function of speech reception threshold (SRT; top panels) and age (bottom panels) for the participants in Experiment Three. To further reduce testing time, only the same-gender maskers were tested. Although all targets and maskers were male, colocated target, and maskers (left panels) and spatially separated target and maskers (right panels) are plotted separately to illustrated the ability of the progressive tracking procedure to capture the same range of thresholds as were found in the first two experiments. SRT values are offset slightly to improve visualization.

completely eclipses the effect of hearing loss as a relevant factor. Not only does SRT fail to reach statistical significance, it literally explains 0% of the variance (see **Table 3**). The second result of note is that the range of thresholds revealed by the progressive tracker is similar to the ranges obtained in the first two experiments. This suggests that a rapid testing approach with fixed TMR values can reveal the same differences between subjects as can the adaptive approach.

GENERAL DISCUSSION

The results of these three experiments clearly demonstrate that the impacts of aging on spatial release can be substantial and can exist completely independently of effects of hearing loss, provided that the experiment is sufficiently powered due to a large sample size and a relatively narrow set of statistical tests. This provides one of the strongest pieces of behavioral evidence to date in support of the hypothesis, based on animal work (e.g., Walton, 2010), that age-related changes in the temporal properties of brainstem nuclei (likely along with cortical changes) can result in difficulties in auditory scene analysis independent of those associated with hearing loss. Future work using a battery of peripheral, central auditory processing, and cognitive tests will be needed to more fully distinguish these effects across a range of tasks, but the data presented here pave the way for further work on the central auditory processes associated with aging *per se*.

In addition, it is clear that the CRM corpus can be used to rapidly and efficiently assess the amount of spatial release a listener can achieve. By transforming the signals with HRIRs that are freely available for download, a VSA was presented over headphones and the performance of an individual listener could quickly be estimated. The relationships between performance in the conditions from Experiment Three and the same conditions as tested in Experiment Two are shown in **Figure 6**.

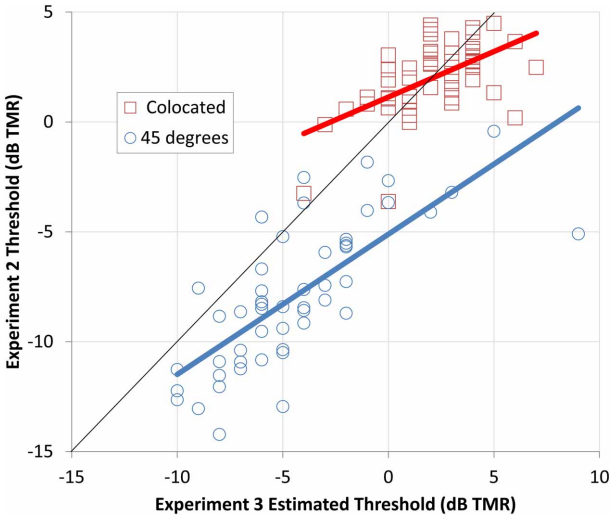


FIGURE 6 | The thresholds obtained in Experiment Two (y axis) for the male target and male maskers are plotted relative to the thresholds estimated from the progressive tracks used in Experiment Three (x axis). The squares indicate the thresholds for the colocated conditions and the circles indicate the spatially separated condition. Best-fitting linear predictors are shown as thick red (colocated) and blue (spatially separated) lines.

The relationship is strongest for the spatially separated conditions ($R^2 = 0.55$), but the colocated conditions are also strongly related ($R^2 = 0.32$). The better thresholds for the spatially separated condition in Experiment Two can be seen by the large number of points falling below the solid black line. These results indicate that the progressive tracking provides fairly accurate estimates of performance in the colocated condition, but underestimates the amount of spatial release a listener is likely to experience. Nonetheless, the progressive tracking appears to provide a rapid method of distinguishing those who are better at separating talkers in a complex auditory scene from those who are worse at this important communicative task.

CONCLUSIONS

These data show clearly that the impacts of age on spatial release from masking can be substantial and independent of the amount of hearing loss a listener suffers. This finding supports the hypothesis, drawn from animal models (e.g., Walton, 2010), that aging results in changes in the temporal processing occurring in brainstem nuclei essential for the encoding of acoustical cues associated with spatial locations of sound sources. To identify these changes in the cortical and subcortical structures of individual patients, and improve communicative rehabilitation strategies, it is important that clinical tests of spatial hearing be developed. Such tests should be designed so that very little time is taken away from the standard clinical test battery, and should use methods that are easily understood by patient and clinician alike. It is hoped that the approach developed in our laboratory and described here can eventually be incorporated into the clinic both as a diagnostic tool and as an outcome measure to assess rehabilitative approaches such as the benefit of a new hearing aid fitting or prescribed auditory training regimen.

ACKNOWLEDGMENTS

We are grateful to the many participants, both Veterans and non-Veterans, who volunteered their time to this study. Assistance in the areas of data collection, database management, and compliance assurance was rendered by Erin Conner, Patrick Tsukuda, and Sara Blankenship. The work was supported by the Department of Veterans Affairs Rehabilitation Research and Development Service (Career Development Award C4963W) and the National Institutes of Health's National Institute for Deafness and Communication Disorders (R01 DC011828). The work was supported with resources and the use of facilities at VA RR&D National Center for Rehabilitative Auditory Research, which is located at the Portland VA Medical Center. The contents of this article are the private views of the authors and should not be assumed to represent the views of the Department of Veterans Affairs or the United States Government.

REFERENCES

Arbogast, T. L., Mason, C. R., and Kidd, G. Jr. (2005). The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.* 117 (4 pt 1), 2169–2180. doi: 10.1121/1.1861598

Best, V., Gallun, F. J., Mason, C. M., Kidd, G., and Shinn-Cunningham, B. G. (2010). The impact of noise and hearing loss on the processing of simultaneous sentences. *Ear Hear* 31, 213–220. doi: 10.1097/AUD.0b013e3181c34ba6

Best, V., Ozmeral, E., Gallun, F. J., Sen, K., and Shinn-Cunningham, B. G. (2005). Spatial unmasking of birdsong in human listeners: energetic and informational factors. *J. Acoust. Soc. Am.* 118, 3766–3773. doi: 10.1121/1.2130949

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (2000). A speech corpus for multitalker communications research. *J. Acoust. Soc. Am.* 107, 1065–1066. doi: 10.1121/1.428288

Brungart, D. S. (2001). Evaluation for speech intelligibility with the coordinate response measure. *J. Acoust. Soc. Am.* 109, 2276–2279. doi: 10.1121/1.1357812

Brungart, D. S., and Simpson, B. D. (2007). Effect of target-masker similarity on across-ear interference in a dichotic cocktail-party listening task. *J. Acoust. Soc. Am.* 122, 1724. doi: 10.1121/1.2756797

Buus, S., Scharf, B., and Florentine, M. (1984). Lateralization and frequency selectivity in normal and impaired hearing. *J. Acoust. Soc. Am.* 76, 77–86. doi: 10.1121/1.391010

Carhart, R., Tillman, T. W., and Greetis, E. S. (1969). Perceptual masking in multiple sound backgrounds. *J. Acoust. Soc. Am.* 45, 694–703. doi: 10.1121/1.1911445

Caspary, D. M., Ling. L., Turner, J. G., and Hughes, L. F. (2008). Inhibitory neurotransmission, plasticity and aging in the mammalian central auditory system. *J. Exp. Biol.* 211 (pt 11), 1781–1791. doi: 10.1242/jeb.013581

Caspary, D. M., Milbrandt, J. C., and Helfert, R. H. (1995). Central auditory aging: GABA changes in the inferior colliculus. *Rev. Exp. Gerontol.* 30, 349–360. doi: 10.1016/0531-5565(94)00052-5

Dubno, J. R., Ahistrom, J. B., and Horwitz, A. R. (2002). Spectral contributions to the benefit from spatial separation of speech and noise. *J. Speech Lang. Hear. Res.* 45, 1297–1310. doi: 10.1044/1092-4388(2002/104)

Dubno, J. R., Ahlstrom, J. B., and Horwitz, A. R. (2008). Binaural advantage for younger and older adults with normal hearing. *J. Speech Lang. Hear. Res.* 51, 539–556. doi: 10.1044/1092-4388(2008/039)

Durlach, N. I., Thompson, C. L., and Colburn, H. S. (1981). Binaural interaction of impaired listeners. A review of past research. *Audiology* 20, 181–211. doi: 10.3109/00206098109072694

Gabriel, K. J., Koehnke, J., and Colburn, H. S. (1992). Frequency dependence of binaural performance in listeners with impaired binaural hearing. *J. Acoust. Soc. Am.* 91, 336–347. doi: 10.1121/1.402776

Gallun, F. J., and Diedesch, A. C. (2013). Exploring the factors predictive of informational masking in a speech recognition task. *Proc. Meet. Acoust.* 19, 060145. doi: 10.1121/1.4799107

Gallun, F. J., Mason, C. R., and Kidd, G. Jr. (2005) Binaural release from informational masking in a speech identification task. *J. Acoust. Soc. Am.* 118, 1614–1625. doi: 10.1121/1.1984876

Gelfand, S. A., Ross, L., and Miller, S. (1988). Sentence reception in noise from one versus two sources: effects of aging and hearing loss. *J. Acoust. Soc. Am.* 83, 248–256. doi: 10.1121/1.396426

Glyde, H., Cameron, S., Dillon, H., Hickson, L., and Seeto, M. (2013). The effects of hearing impairment and aging on spatial processing. *Ear Hear.* 34, 15–28. doi: 10.1097/AUD.0b013e3182617f94

Grose, J. H., and Mamo, S. K. (2010). Processing of temporal fine structure as a function of age. *Ear Hear.* 31, 755–760. doi: 10.1097/AUD.0b013e3181e627e7

Hawkins, D. B., and Wightman, F. L. (1980). Interaural time discrimination ability of listeners with sensorineural hearing loss. *Audiology* 19, 495–507. doi: 10.3109/00206098009070081

Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: effect of location and type of interferer. *J. Acoust. Soc. Am.* 115, 833–843. doi: 10.1121/1.1639908

Helfert, R. H., Sommer, T. J., Meeks, J., Hofstetter, P., and Hughes, L. F. (1999). Age-related synaptic changes in the central nucleus of the inferior colliculus of Fischer-344 rats. *J. Comp. Neurol.* 406, 285–298. doi: 10.1002/(SICI)1096-9861(19990412)406:3<285::AID-CNE1>3.3.CO;2-G

Hopkins, K., and Moore, B. C. (2010).The importance of temporal fine structure information in speech at different spectral regions for normal-hearing and hearing-impaired subjects. *J. Acoust. Soc. Am.* 127, 1595–1608. doi: 10.1121/1.3293003

Koehnke, J., Culotta, C. P., Hawley, M. L., and Colburn, H. S. (1995). Effects of reference interaural time and intensity differences on binaural performance in listeners with normal and impaired hearing. *Ear Hear.* 16, 331–353. doi: 10.1097/00003446-199508000-00001

Lacher-Fougere, S., and Demany, L. (2005). Consequences of cochlear damage for the detection of interaural phase differences. *J. Acoust. Soc. Am.* 118, 2519–2526. doi: 10.1121/1.2032747

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.* 49 (Suppl. 2), 467+. doi: 10.1121/1.1912375

Marrone, N., Mason, C. R., and Kidd, G. Jr. (2008). The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms. *J. Acoust. Soc. Am.* 124, 3064–3075. doi: 10.1121/1.2980441

Moore, B. C. J. (2007). *Cochlear Hearing Loss: Physiological, Psychological, and Technical Issues*. 2nd Edn. West Sessex, England: Wiley & Sons. doi: 10.1002/9780470987889

Neher, T., Laugesen, S., Jensen, N. S., and Kragelund, L. (2011). Can basic auditory and cognitive measures predict hearing-impaired listeners' localization and spatial speech recognition abilities? *J. Acoust. Soc. Am.* 130, 1542–1558. doi: 10.1121/1.3608122

Ross, B., Fujioka, T., Tremblay, K. L., and Picton, T. W. (2007). Aging in binaural hearing begins in mid-life: evidence from cortical auditory-evoked responses to changes in interaural phase. *J. Neurosci.* 27, 11172–11178. doi: 10.1523/JNEUROSCI.1813-07.2007

Ruggles, D., Bharadwaj, H., and Shinn-Cunningham, B. G. (2011). Normal hearing is not enough to guarantee robust encoding of suprathreshold features important to everyday communication. *Proc. Natl. Acad. Sci.* 108, 15516–15521. doi: 10.1073/pnas.1108912108

Scheidt, R. E., Kale, S., and Heinz, M. G. (2010). Noise-induced hearing loss alter the temporal dynamics of auditory-nerve responses. *Hear. Res.* 269, 23–33. doi: 10.1016/j.heares.2010.07.009

Smoski, W. J., and Trahiotis, C. (1986). Discrimination of interaural temporal disparities by normal-hearing listeners and listeners with high-frequency sensorineural hearing loss. *J. Acoust. Soc. Am.* 79, 1541-1547. doi: 10.1121/1.393680

Strelcyk, O., and Dau, T. (2009). Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing. *J. Acoust. Soc. Am.* 125, 3328–3345. doi: 10.1121/1.3097469

Walton, J. P. (2010). Timing is everything: temporal processing deficits in the aged auditory brainstem. *Hear. Res.* 264, 63–69. doi: 10.1016/j.heares.2010.03.002

Wang, H., Turner, J. G., Ling, L., Parrish, J. L., Hughes, L. F., and Caspary, D. M. (2009). Age-related changes in glycine receptor subunit composition and binding in dorsal cochlear nucleus. *Neuroscience* 160, 227–239. doi: 10.1016/j.neuroscience.2009.01.079

Xie, B. (2013). *Head-Related Transfer Function and Virtual Auditory Display, 2nd Edn*. Plantation, FL: J. Ross Publishing.

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

Received: 24 October 2013; accepted: 06 December 2013; published online: 23 December 2013.

Citation: Gallun FJ, Diedesch AC, Kampel SD and Jakien KM (2013) Independent impacts of age and hearing loss on spatial release in a complex auditory environment. Front. Neurosci. 7:252. doi: 10.3389/fnins.2013.00252

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

frontiers in
NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 28 February 2014
doi: 10.3389/fnins.2014.00025

Stable individual characteristics in the perception of multiple embedded patterns in multistable auditory stimuli

Susan Denham[1,2]*, Tamás M. Bõhm[3,4], Alexandra Bendixen[5], Orsolya Szalárdy[3,6], Zsuzsanna Kocsis[3,6], Robert Mill[1] and István Winkler[3,7]

[1] Cognition Institute, University of Plymouth, Plymouth,UK
[2] School of Psychology, University of Plymouth, Plymouth, UK
[3] Research Centre for Natural Sciences, Institute of Cognitive Neuroscience and Psychology, Hungarian Academy of Sciences, Budapest, Hungary
[4] Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary
[5] Auditory Psychophysiology Lab, Department of Psychology, Cluster of Excellence "Hearing4all," European Medical School, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany
[6] Department of Cognitive Science, Budapest University of Technology and Economics, Budapest, Hungary
[7] Institute of Psychology, University of Szeged, Szeged, Hungary

Edited by:
Susann Deike, Leibniz Institute for Neurobiology, Germany

Reviewed by:
Claude Alain, Rotman Research Institute, Canada
Peter Lakatos, Hungarian Academy of Sciences, Hungary

***Correspondence:**
Susan Denham, School of Psychology, University of Plymouth, Drake Circus, Plymouth, PL4 8aa, UK
e-mail: s.denham@plymouth.ac.uk

The ability of the auditory system to parse complex scenes into component objects in order to extract information from the environment is very robust, yet the processing principles underlying this ability are still not well understood. This study was designed to investigate the proposal that the auditory system constructs multiple interpretations of the acoustic scene in parallel, based on the finding that when listening to a long repetitive sequence listeners report switching between different perceptual organizations. Using the "ABA-" auditory streaming paradigm we trained listeners until they could reliably recognize all possible embedded patterns of length four which could in principle be extracted from the sequence, and in a series of test sessions investigated their spontaneous reports of those patterns. With the training allowing them to identify and mark a wider variety of possible patterns, participants spontaneously reported many more patterns than the ones traditionally assumed (Integrated vs. Segregated). Despite receiving consistent training and despite the apparent randomness of perceptual switching, we found individual switching patterns were idiosyncratic; i.e., the perceptual switching patterns of each participant were more similar to their own switching patterns in different sessions than to those of other participants. These individual differences were found to be preserved even between test sessions held a year after the initial experiment. Our results support the idea that the auditory system attempts to extract an exhaustive set of embedded patterns which can be used to generate expectations of future events and which by competing for dominance give rise to (changing) perceptual awareness, with the characteristics of pattern discovery and perceptual competition having a strong idiosyncratic component. Perceptual multistability thus provides a means for characterizing both general mechanisms and individual differences in human perception.

Keywords: auditory scene analysis, multistability, auditory streaming, perceptual switching, individual differences

INTRODUCTION

Most sound sources of interest in the world around us emit sequences of sound events, e.g., the notes in a birdsong or the words spoken in conversation. These sounds are seldom present in isolation and typical sound events consist of many time-varying components. The problem for auditory perception is to parse this complex scene, and to do so in a timely manner that allows us to interact appropriately with the sound emitting objects of interest. The problem of grouping, both the simultaneously present components that belong to the same sound event, and the sequential associations between events emitted by the same source, is known as auditory scene analysis (Bregman, 1990). Understanding this seemingly effortless process of perceptual organization is an essential step toward explaining what determines our conscious perceptions of the world. Most of the studies in this field to date have focused on trying to identify the general processing strategies used by the human brain for parsing the auditory scene. In doing so, inter-individual differences have typically been treated as a source of noise in the experimental data, as is the case in many cognitive studies (cf. Kanai and Rees, 2011). In the current study, we asked whether the patterns of responses obtained from individual listeners are stable characteristics of the person.

Sequential grouping in auditory scene analysis has typically been studied using the auditory streaming paradigm: ABA-ABA- where A and B denote different sounds and "-" stands for a silent interval with the same duration as the two sounds (van Noorden, 1975; Bregman, 1990). There has been a long-standing assumption that in listening to such a sound sequence, listeners make a perceptual decision between *integration* (the grouping of

all sounds into the same stream, perceived as a repeating ABA- pattern) and *segregation* (the parsing of the sequence into two separate streams, perceived as repeating A- and B--- patterns, with one in the foreground and the other in the background). This perceptual decision is influenced by the distinctiveness (or similarity) of the A and B tones (e.g., differences in frequency, location or timbre) and the rate at which the sounds are presented (for a review see Moore and Gockel, 2012). The trade-off between similarity and presentation rate has led to the suggestion that in audition the Gestalt principle of similarity (Köhler, 1947) is mediated by time; i.e., similarity and good continuation combine to determine the likelihood of sequential grouping (Jones, 1976; Winkler et al., 2012; Denham and Winkler, 2014).

In most studies it has been assumed that *integration* and *segregation* are the only perceptual organizations possible, and that they are mutually exclusive (van Noorden, 1975). In addition, because it was thought that integration was always perceived first and that the build-up of segregation was a rather slow process (on the order of several seconds) (Anstis and Saida, 1985), it was suggested that perceptual organization could be viewed as a process in which the auditory system accumulates evidence in favor of an appropriate perceptual decision (Bregman, 1990). However, a number of recent experiments have challenged these ideas. Firstly, it has been found that there are other possible perceptual organizations, and, when given the possibility to do so, listeners also report hearing repeating patterns, sometimes for rather long periods of time, which do not match either of the patterns described above (Bendixen et al., 2010; Denham et al., 2013). Secondly, rather than fixing on a single perceptual decision, given sufficient time, perception switches between alternative interpretations of the sequence (Denham and Winkler, 2006; Pressnitzer and Hupé, 2006) and does so for all combinations of frequency difference and presentation rate tested to date, even those that have been assumed to be strongly biased toward either *integration* or *segregation* (Denham et al., 2013). Thirdly, *segregation* is often reported first for some parameter combinations (Deike et al., 2012; Denham et al., 2013), and the gradual build-up of *segregation* has been shown to be to some extent an artifact of the analysis and visualization methods used (Deike et al., 2012).

Based on these new findings it has been proposed that perceptual organization is a process in which the auditory system continually attempts to discover patterns (or regularities) in the incoming sequence (Winkler et al., 2012; Denham and Winkler, 2014). Multiple such patterns may be detected embedded in a sequence and represented in parallel. Consistent with theories of binocular rivalry (for a review see Logothetis et al., 1996), the proposal is that a sequence of conscious perceptual states arises as a result of ongoing competition for perceptual dominance between concurrent rivaling percepts; evidence for which has been found in an auditory mismatch negativity experiment (Horváth et al., 2001). The ease with which each pattern is discovered (related to the notion of similarity described above) determines how likely it is that the pattern will be perceived, especially at the beginning of a sound sequence. The auditory system uses each detected pattern to generate expectations of future events, that, if violated, signal new (i.e., as yet unmodeled) information in the sequence. Patterns that predict the same events compete for dominance. Compatible patterns, i.e., those that do not attempt to predict the same events, form cooperative groups that give rise to the perceptual organizations reported by listeners. Perceptual switching between these cooperative groups, and the corresponding changes in perceptual awareness, are caused by the competition between incompatible patterns (Winkler et al., 2012). A computational model based on these principles successfully replicated many of the phenomena of perceptual bi-stability in the auditory streaming paradigm (Mill et al., 2013).

EXPERIMENT 1

In our previous experiments, having realized that participants sometimes experienced organizations other than *integration* and *segregation*, we used instructions that provided participants with a wider range of possible reports (e.g., see Bendixen et al., 2010; Denham et al., 2013). Participants were instructed to report *integration* if all tones in the sequence were perceived as belonging to a single repeating pattern (i.e., a single stream); *segregation,* if the sequence was perceived as consisting of two repeating patterns (or streams), one containing only the A tones and the other only the B tones; *both*, if the sequence was perceived as consisting of two streams, one containing both A and B tones, and the other only A or only B tones; and finally, *none*, if no repeating pattern was detected. Over the course of many experiments we found that (1) *both* is reported on average between 10 and 30% of the total duration, (2) *both* is almost never reported as the first percept, (3) the incidence of *both* percepts varies considerably between listeners, and (4) there is a tendency for *both* reports to be more common for parameter combinations supporting more evenly balanced proportions of *integration* and *segregation* (Bendixen et al., 2010, 2013; Denham et al., 2013; Szalárdy et al., 2013).

Although these experiments have supported the idea that perceptual organizations other than *integration* and *segregation* can be perceived when listening to the ABA- sequence, precisely what patterns listeners perceive when they report *both* has not been previously investigated. One possibility is that only *integration* and *segregation* are perceived but there is sometimes very rapid switching between them, and since there is some sluggishness in the system, listeners simply report *both*. In this case we would expect to find no explicit reports of distinctive patterns other than ABA-, A-, or B---. Another possibility, suggested by our modeling studies, is that the auditory system finds many patterns embedded even within this simple sequence, which results in the emergence of other compatible groups and thus other perceptual organizations; see **Figure 1**. In the experiments reported here, we investigated whether listeners spontaneously report perceiving these more uncommon patterns and, if so, to what extent their perception is influenced by stimulus parameters.

In order for participants to be able to quickly and reliably report what they perceived and to associate the possible patterns with the user interface controls, each participant attended a series of training sessions prior to the commencement of the main experiment. Previous work (Rogers and Bregman, 1993; Snyder et al., 2008, 2009b; Haywood and Roberts, 2011, 2013) has shown strong contextual effects of prior learning on auditory streaming. However, these studies of contextual influences on perception have typically involved within-session manipulations, and the

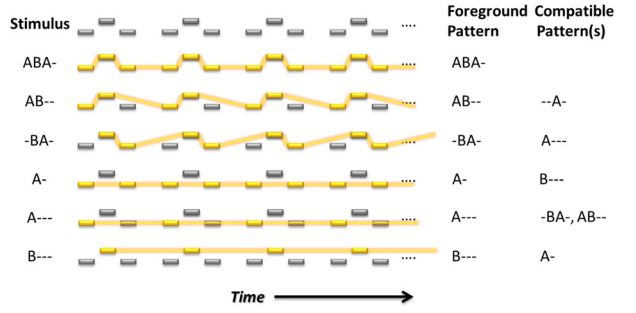


FIGURE 1 | Embedded repeating patterns up to length four that can, in theory, be perceived by listeners in the auditory streaming paradigm. Yellow rectangles indicate the foreground pattern. Feature differences, in this case frequency differences, are indicated by displacement in a vertical direction for the tones in each sequence. The foreground and compatible (background) patterns for each dominant (emphasized) pattern are indicated to the right in letter form.

perceptual effects have been probed using stimuli of rather short duration (generally <10 s). To our knowledge the experiments in this report are the first to include extensive multi-session training. In addition, we are concerned here not with differential effects of prior training on perception, but rather with the reliability of participants' ability to correctly categorize the specific patterns that may occupy their perceptual awareness.

Although the characteristics of perceptual switching are known to be very stochastic (Levelt, 1968), it is also known that some characteristics, e.g., typical switching rates, may be rather idiosyncratic (Aafjes et al., 1966; Kanai et al., 2010). We therefore investigated whether there was internal consistency in the data; i.e., whether individual listeners' perceptual switching behavior was similar across sessions. If so, this would provide some measure of confidence that listeners were reporting what they perceived, and were reliably engaged in the task. Addressing this question required us to conduct numerous (10 or more) sessions with each individual listener, and hence we involved only a small number of experimental participants ($N = 6$).

In summary, the two aims of the study were: (1) to determine whether listeners at least occasionally experience sequences created according to the auditory streaming paradigm in terms of percepts outside the traditional *integrated* and *segregated* sound organizations, and (2) to assess whether the perceptual reports of individual listeners show stable consistent characteristics (i.e., whether the patterns of perceptual reports are more similar within a single listener across sessions than between different listeners).

METHODS

Participants

Six healthy volunteers (mean age 22.3 years, range 19–25 years; all right-handed; 4 male, 2 female) took part in Experiment 1, which was conducted over numerous sessions over a period of approximately 1 month. All participants had reportedly normal hearing. None of the participants were taking any medication affecting the central nervous system. In compliance with the Declaration of Helsinki, participants gave written informed consent after the experimental procedures had been explained to them. Participants received modest financial compensation for their participation.

Stimuli

Sinusoidal tones of 75 ms (ms) duration (including 10 ms rise and fall times) and with an intensity of 40 dB sensation level (above hearing threshold, adjusted individually for each participant) were arranged according to the auditory streaming paradigm (a cyclically repeating "ABA-" pattern) in five stimulus conditions, with frequency difference (Δf) and stimulus onset asynchrony (SOA, onset to onset time interval) as follows: (1) $\Delta f = 3$ semitones (st), SOA = 100 ms; (2) $\Delta f = 16$ st, SOA = 100 ms; (3) $\Delta f = 7$ st, SOA = 150 ms; (4) $\Delta f = 3$ st, SOA = 200 ms; (5) $\Delta f = 16$ st, SOA = 200 ms. The frequency of the "A" tones was 400 Hz, and the frequency of the "B" tones was n semitones higher, depending on condition. At each test session, participants were presented with 10 min long ABA- tone sequences, one for each of the five conditions, delivered in a randomized order. An extra 30 s verification segment (see the *Test procedure* section) was appended to the end of each 10-min long stimulus block.

Apparatus and procedures

Participants were seated in an acoustically shielded chamber. Sounds were presented binaurally via headphones. Responses were given using a touch-screen monitor. As illustrated in **Figure 1**, restricting the patterns to length no longer than four, there are six possible patterns that can be extracted from the ABA- sequence; ABA-, AB--, -BA-, A-, A---, and B---. A specific area of the display was assigned to each of these response options (indicated by color and graphical icon). A further area (gray, "0") allowed listeners to indicate when they could not decide between the patterns (*confused*). Participants were required to use the index finger of their right hand to press the button corresponding to the pattern they experienced, and to keep the button depressed for as long as they continued hearing the pattern. The interface did not allow multiple responses to be reported simultaneously. A screenshot of the response screen with one "response button" pressed can be seen in **Figure 2**.

Training procedure

In order to make sure that listeners understood how each pattern sounds, and could reliably report their percepts, they attended a number of training sessions before the main experiment. In these sessions they learnt to use the response interface, and to report the dominant (foreground) pattern that they perceived. A detailed log of each participant's training history was kept. The experimenter screened performance to decide when to stop training. Only once listeners could reliably report the entire set of patterns were they ready to take part in the main experiment.

In each training session the experimenter adaptively adjusted the training procedure in accord with the participants' level of understanding and performance. Two pattern repetition speeds were pre-assigned by varying the silent period; slow ABA-- and normal ABA-, where "-" is a silent period corresponding to the SOA. Participants started with putatively easier tasks, and then proceeded to more difficult ones as their performance improved. The teaching procedure started with a demonstration in which

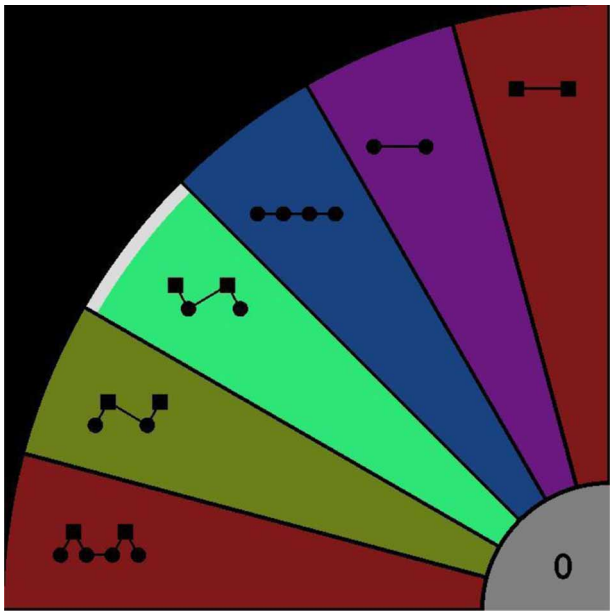


FIGURE 2 | Response interface. Participants were instructed to press the colored area corresponding to their dominant (foreground) percept, as indicated by the icons. Starting from the lowest area these patterns can be expressed using letters as ABA-, AB--, -BA-, A-, A---, B---, *confused* (gray area "0"). In this screen shot, the -BA- pattern is being reported as the currently perceived pattern (indicated by higher intensity and the white segment at the outer edge of the region).

the participants listened to each of the possible patterns where the tones not included in the pattern were left out and the corresponding button was depressed on the screen. Next, they were presented with blocks in which the order of the patterns were randomized and the duration of the patterns were also different. Participants had to press the button corresponding to their percept. These blocks were repeated for as long as the experimenter considered it necessary.

Next, the different patterns were introduced by means of emphasis, i.e., tones not part of the pattern were attenuated (−18 dB). Once again participants were asked to observe each of the possible pattern and button combinations, as explained above. Then they were presented with blocks in which the order of the patterns were randomized and the duration of the patterns were also different. Participants had to press the button corresponding to their percept. These blocks were again repeated for as long as the experimenter considered it necessary.

Over the course of three or four sessions these tasks were repeated until the participants reached the point where they were able to confidently identify the patterns. Finally, one or two blocks identical to the test procedure were administered to prepare them for the test sessions.

Test procedure

Once their training had been completed, participants attended seven test sessions spread over a period of approximately 1 month with at least 2 days between consecutive sessions.

At the start of each test session the experimenter made sure that participants remembered the patterns they were required to report using both auditory and visual illustrations. They were also reminded of their task by means of two to four training blocks in which the various patterns were emphasized, as described above.

Participants were instructed to listen to the tone sequences and to continuously indicate their percepts by depressing the region of the touch-screen corresponding to the pattern that was currently most prominent. Participants were encouraged to employ a neutral listening set, and to refrain from attempting to hear out one or another pattern. A break of at least 30 s separated successive stimulus blocks, with additional time given to participants as needed. Each test session lasted up to 1.5 h.

A criticism that has been leveled at the auditory streaming paradigm is that it relies on listeners being able to make accurate subjective reports of their perceptions. It is of course difficult to verify whether someone is actually reporting what they hear, rather than simply pressing buttons randomly or in order to satisfy the experimenter. We tried to address this issue in two ways. Firstly, an extra 30 s verification segment was appended to the end of each 10-min long stimulus block. In these verification segments, one of the six patterns (ABA-, AB--, -BA-, A-, A---, B---), randomly chosen, was emphasized by attenuating all non-pattern tones (−18 dB); i.e., the stimulation was identical to those used for training. If listeners are correctly reporting their percepts, then they should report the emphasized pattern during this period. Secondly, listeners performed the main experiment repeatedly and on different occasions (i.e., in seven separate sessions spread over a period of approximately 1 month). The rationale was that random button pressing should not lead to consistent response patterns across sessions. Altogether, including training sessions, four participants took part in 10 sessions, while two participants took part in 11.

Data recording and analysis

The state of the response buttons was continuously recorded at a nominal sampling rate of 250 Hz. However, due to the use of the graphical user interface for collecting data, it was not possible to guarantee a strictly regular sampling. For this reason the raw data was resampled at a regular 4 ms sampling period. Before analysing the button presses, because there was an explicit button for participants to use if they were confused or heard none of the patterns, we removed the phases where no button was pressed. In addition, all cases in which the duration between successive changes in response (termed a *perceptual phase*) was shorter than 300 ms were discarded because these were assumed not to result from intentional reports (Moreno-Bote et al., 2010). In Experiment 1 the data removed in these ways amounted to 3.6% of the total data duration.

To check that participants were performing the task correctly, the verification data was analyzed to determine the total proportion of time spent reporting the emphasized pattern. The latency for switching to the emphasized pattern from the start of the verification section was also extracted in order to take account of the time taken for the emphasized pattern to overcome the dominance of whatever pattern participants perceived at the time the verification section started. The combination of

these two durations gives a good account of the accuracy with which participants could identify and report each of the patterns.

To investigate the dynamics of the discovery of alternative patterns, the latency for the first report of each pattern was extracted for each participant for each session and condition. It is not necessarily the case that all patterns are experienced, but if they are this gives a measure of how quickly they are discovered by the auditory system. Perceptual phase durations in both auditory and visual multistability experiments are typically log normally distributed (Pressnitzer and Hupé, 2006). Therefore, mean phase durations are usually calculated by finding the mean value in the log domain and then converting back to the linear domain. However, here we found that in some cases the duration data (especially the latency to the first report of each pattern) was not log normally distributed, so for consistency, summary durations are given as the median of the corresponding data.

To characterize the perceptual switching patterns of participants in response to the tone sequences and to analyse the parameter dependence of perception, transition matrices were constructed using the method described by Denham et al. (2012). Each transition matrix is a 7×7 matrix with elements that represent the probability of switching from one percept to another percept (i.e., seven possibilities in all: six patterns and *confused*), with the percept of origin corresponding to the column and the destination percept to the row of the element. A *global transition matrix* that summarizes all of the switching patterns found in the experiment was constructed by counting the number of occurrences of each transition for all of the data recorded during the experiment. From this matrix, the overall proportions and phase durations of each percept were extracted. A set of five *condition transition matrices* were constructed by counting the number of occurrences of each possible transition, pooling all participants and all test sessions for each condition separately. To analyse individual differences between participants, *participant transition matrices* were constructed, one per participant per session. The participant transition matrices were calculated by counting the number of occurrences of each possible transition, pooling responses from all conditions for each participant within each test session. As explained in Denham et al. (2012), the global transition matrix provides neutral default values in the case of missing transitions for the more restricted data sets used to construct the condition and participant matrices.

Individual differences were investigated using the participant transition matrices. The Kullback-Leibler divergence (Kullback, 1959) between each matrix from an individual participant (intra-participant distances) and between each individual participant's matrices and all other participants' transition matrices (inter-participant distances) was used as a measure of similarity between perceptual switching patterns. This is a richer characterization of perceptual switching than the switching rate measure usually used in this regard.

The overall distributions of the proportions of each of the patterns were compared using a repeated-measures ANOVA. The effects of condition on the proportion of each pattern were analyzed using a repeated-measures ANOVA followed by *post-hoc* pairwise comparisons using a pairwise sign test to analyse the influence of stimulus parameters on the proportion of *segregated*, *integrated* and *both* phases. The distributions of overall phase durations were analyzed using a repeated-measures ANOVA and pairwise Wilcoxon rank sum tests. Wilcoxon rank sum tests were used to compare cumulative latency distributions. Finally, the distributions of intra- vs. inter-individual differences were compared separately for each participant using a Wilcoxon rank sum test. The significance of all statistical tests was assessed at the 95% confidence level ($\alpha = 0.05$). All analyses were carried out using Matlab and the Matlab Statistics Toolbox.

RESULTS

Unusable data

Three sessions were affected by problems with the user interface, as indicated by the occurrence of blocks in which no responses were recorded. These sessions (participant 1, session 8, participant 2, sessions 7 and 8) were excluded from the analysis.

Verification data

The mean proportion of the total duration of the verification responses (i.e., those made during the 30 s verification sequences appended to the end of the test block, in which particular patterns were emphasized) matching the emphasized patterns was 88.9%; 96.2%, if the latency to switch to the emphasized pattern is included. The latency to the first switch to the emphasized response accounted for a mean of 2.2 s. **Figure 3** shows this data according to pattern and according to participant. Overall, these "catch" sections showed that participants reliably categorized each emphasized pattern.

Proportion of each perceptual pattern

The first question we wished to address was whether participants would perceive all of the patterns they encountered during training, or whether only the conventional patterns of *integration* (ABA-) and *segregation* (A-, B---) would be reported. The perceptual reports of all participants, across all test sessions and all conditions were pooled to investigate the overall occurrence of each pattern. We found that all of the patterns in the training data were reported during the experimental sequences. The distribution of the proportions of each varied widely [comparing the pattern proportions using a One-Way repeated-measures ANOVA, $F_{(6,\ 35)} = 30.85$, $p < 0.0001$]; **Figure 4** (top). Pairwise comparisons between patterns show that all pattern proportions (except AB-- and A--- $p = 0.2$, AB-- and *confused* $p = 0.14$, -BA- and A- $p = 0.22$, -BA- and B--- $p = 0.46$, and A--- and *confused*, $p = 0.32$) are significantly different from each other, $p < 0.05$. **Figure 4** (bottom), shows the proportions resulting from pooling the response alternatives into the categories used in our previous studies, *integrated*, *segregated* and *both (none* was excluded because of the very low incidence of *confused*, and the removal of all instances of *no button press)*; these proportions are comparable with those found in previous experiments (e.g., Bendixen et al., 2010; Denham et al., 2013).

The parameter dependence of the proportion of time during which each pattern was perceived was analyzed using the condition transition matrices, both for the original responses and for the pooled responses (i.e., pooled for the response categories used

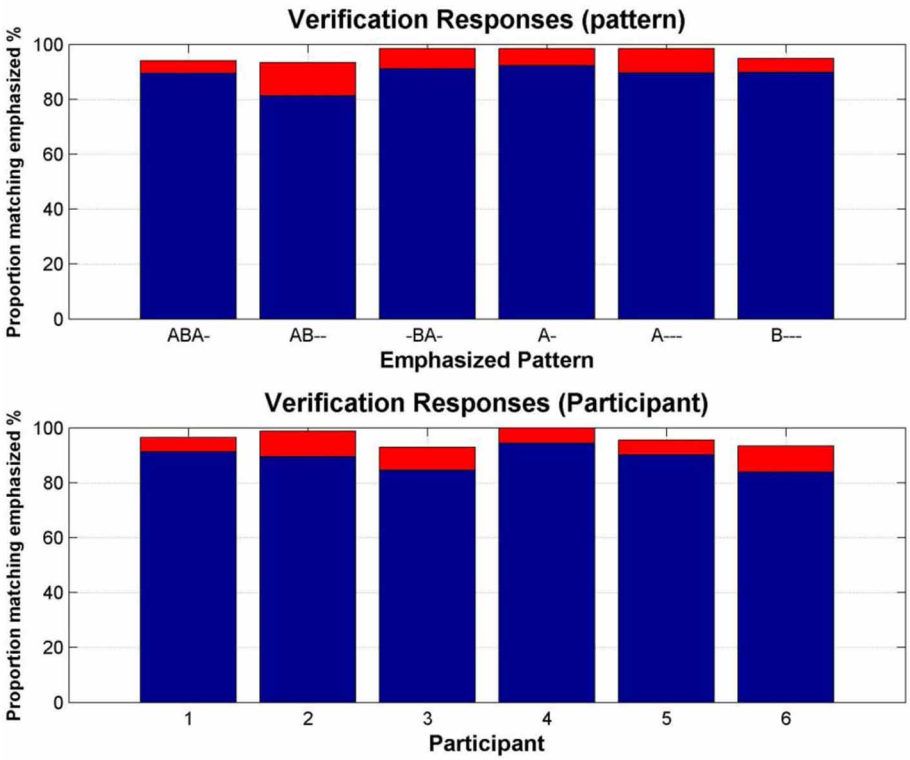


FIGURE 3 | Verification of pattern recognition. Top: Proportion of the verification sections in which each emphasized pattern was reported. **Bottom:** Proportion of the verification sections in which each participant reported the emphasized patterns. Blue bars show the mean proportion of matching responses, red bars show the additional proportion accounted for by the latency to the first switch to the matching response.

in previous studies). As expected from previous experiments (e.g., Denham et al., 2013), a fast rate of change of stimulus parameters ($\Delta f = 16$ st, SOA = 100 ms) increases the proportion of *segregation* [comparing the effect of condition using a One-Way repeated-measures ANOVA, $F_{(4,\ 25)} = 12.7$, $p < 0.0001$; pairwise sign test comparing the proportion of *segregation* in condition 2 vs. the proportion in every other condition, all $p < 0.05$], a slow rate of change of stimulus parameters ($\Delta f = 3$ st, SOA = 200 ms) increases the proportion of *integration* [comparing the effect of condition using a One-Way repeated-measures ANOVA, $F_{(4,\ 25)} = 13.96$, $p < 0.0001$; pairwise sign test comparing the proportion of *integration* in condition 4 vs. proportion in every other condition, all $p < 0.05$]. By considering the grouped percepts, we can also see that if the patterns hypothesized to correspond to the *both* responses in previous experiments are pooled, then we find that there tends to be a higher incidence of *both* in the regions of intermediate rate of feature change, here conditions 1, 3, and 5. This seems to confirm Denham et al's. (2013) observations; however, a One-Way repeated-measures ANOVA did not show a significant effect of condition, $F_{(4,\ 25)} = 0.73$, $p > 0.5$, therefore the planned pairwise comparisons were not performed.

Phase duration of each perceptual pattern

Another way to explore the data is to consider the statistics of perceptual phase durations. Phase durations can provide insights into the dynamics of perceptual switching not always apparent from the proportions; the same proportions can be achieved by many short phases or by fewer longer phases. In **Figure 6** the median phase durations for each pattern are shown for the entire data set; for clarity the *confused* phases are omitted from this analysis. This plot shows that although the proportion of the patterns (AB--, -BA-, and A---) is rather low, if participants do report them, then the phase durations during which they are experienced can be comparable with those of the *segregated* percepts [comparing the median phase durations of the six different patterns using a One-Way repeated-measures ANOVA, $F_{(5,\ 30)} = 2.32$, $p = 0.06$; pairwise Wilcoxon rank sum test comparing distributions shows the durations of AB-- and -BA- not to be not significantly different from any other pattern, $p > 0.05$ for all comparisons, and the durations of A--- not to be significantly different from any other pattern, $p > 0.05$ for all comparisons except ABA-, $p = 0.004$].

Latency

The time taken to discover each of the perceptual patterns varied as a function of stimulus parameters (condition) and between individual participants; see **Figure 7**. With the exception of the A--- pattern, all patterns are eventually reported for all conditions, and all participants (with the exception of participant 4: AB--) reported all of the patterns at some time during the experiment.

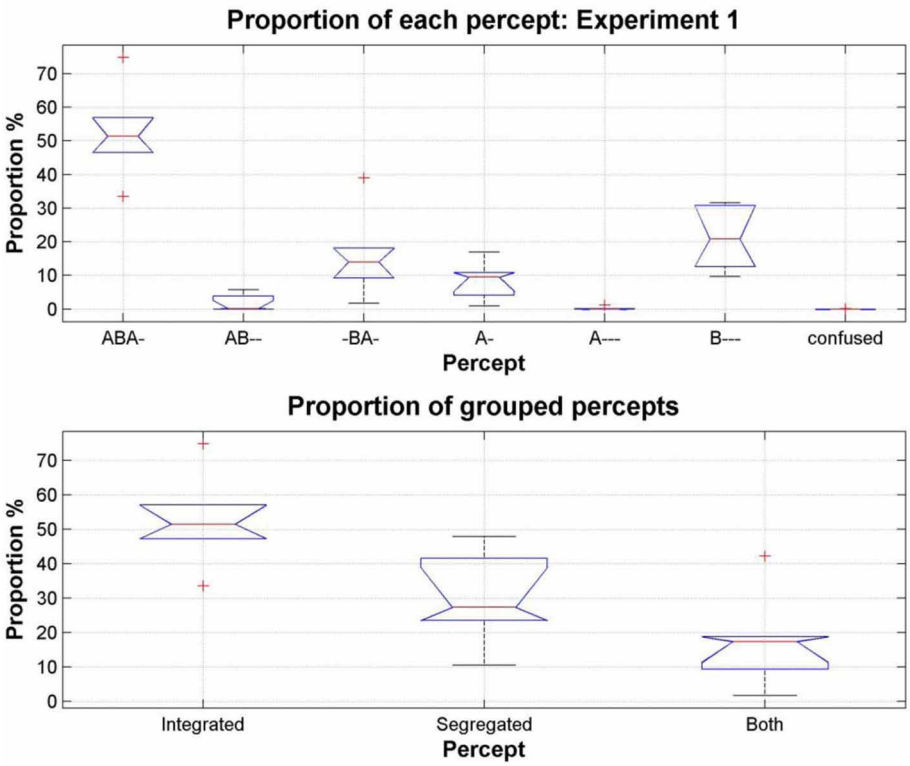


FIGURE 4 | Total proportion of each pattern, pooling the data from all participants, all sessions and all conditions in Experiment 1. On each box, the central red line is the median, the upper and lower edges of the boxes are the 25th and 75th percentiles, the whiskers extend to the most extreme data points considered not to be outliers; outliers are plotted individually as red crosses. **Top:** Proportion of each pattern reported by participants (median values: ABA- 51.39%, AB-- 0.13%, -BA- 14.02%, A- 9.49%, A--- 0.02%, B--- 20.78%, *confused* 0.0008%). **Bottom:** Patterns grouped according to the perceptual organizations that participants were allowed to report in our previous studies: *integrated* (ABA-), *segregated* (A-, B---), and *both* (AB--, -BA-, A---).

The latencies for reporting each pattern are not normally distributed; plotting the latency distribution in the log domain clearly shows its bimodal nature (**Figure 8**, top). This occurs because there is a tendency in auditory streaming experiments (as reported in vision too, e.g., Mamassian and Goutcher, 2005) for the first phase to be much longer than subsequent phases (Denham et al., 2013); the short latency peak in the distribution corresponds to the initial responses (typically ABA-, A-, or B---), with the later peak corresponding to the first reports of percepts in subsequent phases.

The empirical cumulative distributions of latencies, which show on the y axis the proportion of latencies less than any given latency (in seconds) on the x axis (**Figure 8**, bottom), demonstrate the strong tendency for the *integrated* ABA- pattern to be reported with a far shorter latency than the other patterns, followed by similar latency distributions for the B---, A-, and -BA- patterns, and finally, much later the AB-- and A--- patterns. For example, by reading off the intercepts of each cumulative plot with the dashed red line we can see that 80% of first reports of ABA- occur within 5.5 s, of B--- within 120 s, -BA 204 s, A- 222 s, AB—308 s, and A--- 440 s. This may explain why the other patterns have not been reported in past experiments, as most of them have used sequences of short duration (typically <20 s). Wilcoxon rank sum test comparing cumulative distributions shows ABA- to be significantly different from all other patterns, $p < 0.0001$, B---, A-, and -BA- to be significantly different from ABA-, A---, and AB--, $p < 0.0001$, but not from each other (B---/-BA- $p = 0.11$, A-/-BA- $p = 0.44$, except B---/A- $p = 0.013$), and A--- and AB—to be significantly different from all other patterns, $p < 0.0001$, but not from each other, $p = 0.098$.)

Consistency of individual perceptual switching behavior

The transition matrices constructed for each participant for each test session by pooling the data for all conditions were used to examine the consistency of individual behavior. **Figure 9** below shows a comparison between intra- and inter-individual differences in terms of explicit difference measures. This demonstrates that individual participant behavior tends to be idiosyncratic; i.e., the switching behavior of an individual is more similar to their own behavior in a different test session than it is to the perceptual switching behavior of other participants (Wilcoxon rank sum test: participant 1 $p = 0.0002$, participants 2–6 $p < 0.0001$). It should be noted that by comparing the transition matrices of individual participants we go beyond just comparing switching rates; in addition to switching rates, the transition matrices capture the likelihood of reporting and switching between different perceptual patterns.

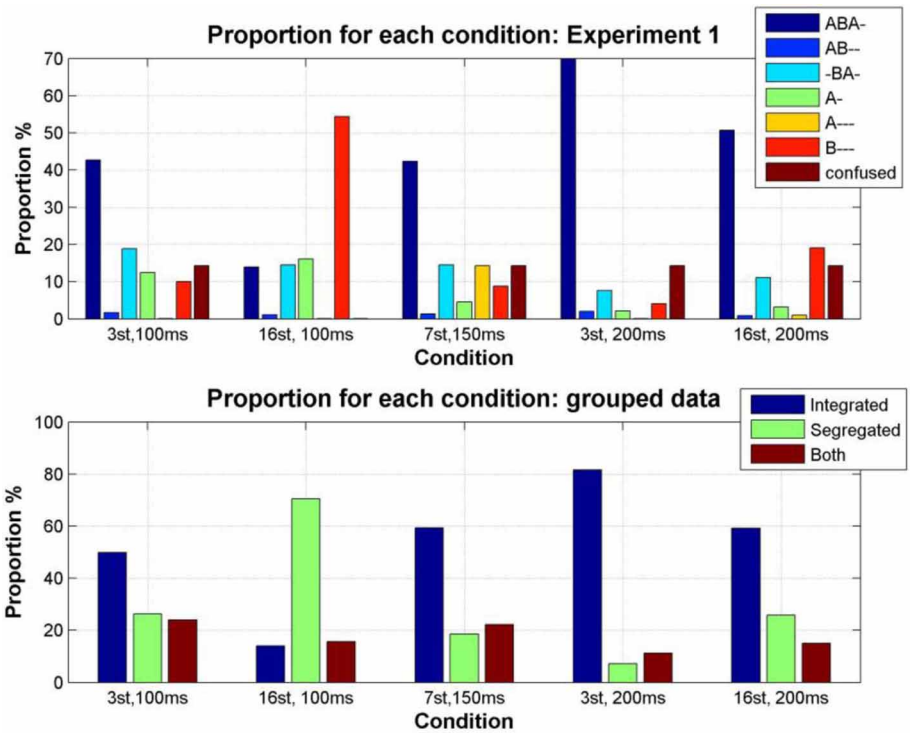


FIGURE 5 | Proportion of the total stimulus duration during which each pattern was perceived (color coded as indicated by the accompanying legend) for each condition, calculated from the condition transition matrices constructed by pooling the data from all participants for each condition (Denham et al., 2012). Top: Proportion of each pattern reported by participants. **Bottom:** Proportion of each pattern grouped according to the perceptual organizations that participants were allowed to report in our previous studies; *integrated* (ABA-), *segregated* (A-, B---), and *both* (AB--, -BA-, A---).

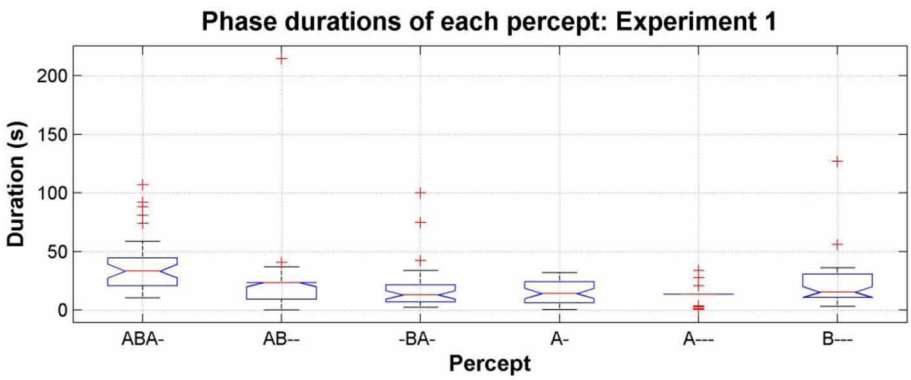


FIGURE 6 | Phase durations of each pattern, pooling the data from all participants, all sessions and all conditions in experiment 1; see Figure 4 for figure conventions.

INTERIM DISCUSSION

When listening to tone sequences of the form ABA-, participants report patterns other than the traditional integrated (ABA-) and segregated (A- or B---) ones if they are given the possibility to do so. Although the experiment is rather complex and participants require a number of training sessions before they are reliably able to perform the task, perceptual reports in the verification sections appended to the end of the stimuli provide confidence that participants were engaged in the task and accurately reported the patterns they perceived. Individual behavior was also consistent. We found that the perceptual switching behavior of each participant was very similar across sessions; median differences between an individual participant's transition matrices were small. In comparison, the differences between the transition matrices of different participants tended to be much larger; median differences were more than twice as large (except for participant 1, 1.8 times larger). This stable idiosyncratic behavior is interesting in its own right, as well as providing further confidence that

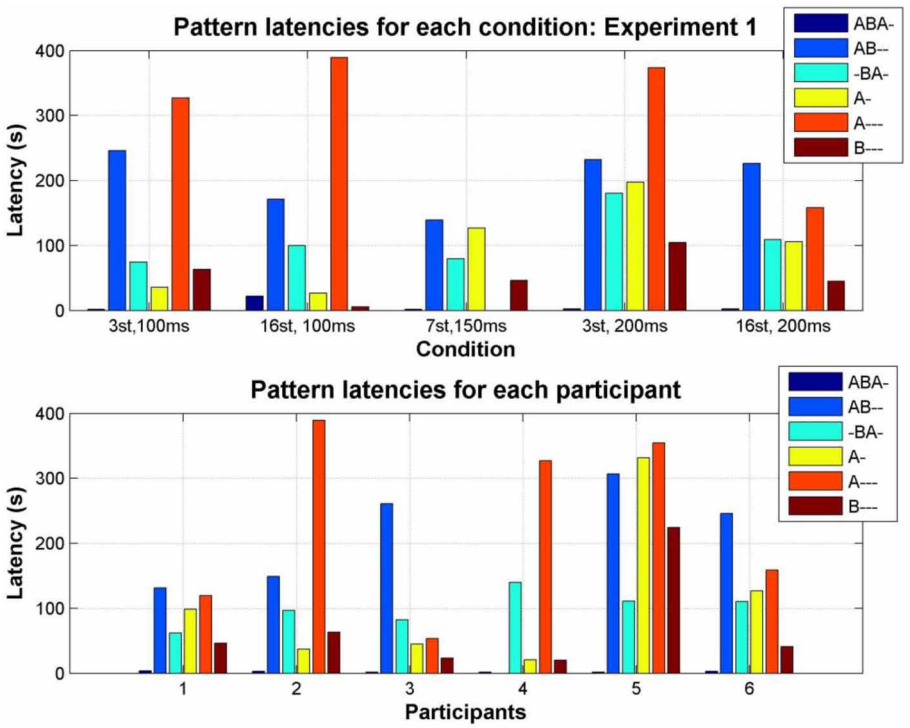


FIGURE 7 | Median latency of each pattern (color coded as indicated by the accompanying legend). Top: Median latency for the first report of each pattern in each condition, pooling all participants. **Bottom:** Median latency for the first report of each pattern by each participant, pooling all conditions.

participants were engaged in the task and were reliably reporting their perceptual experiences.

The choice of embedded patterns that we tested here was motivated in part by our modeling studies (Mill et al., 2013). The CHAINS model, which was used to successfully simulate the dynamics of the discovery and perceptual switching between the integrated and segregated organizations, can in principle also discover the embedded AB--, -BA-, and A--- patterns. Here we tested the perception of all repeating embedded patterns up to length 4. It is possible that listeners can spontaneously perceive even longer patterns if the sequence allows it, but we have not explored that possibility yet. Because we considered that participants may find the distinction too difficult to perceive, we did not ask participants to distinguish between the A--- and --A- patterns (i.e., complementary to -BA- and AB--, respectively), although the CHAINS model would predict that they are both perceived. The other limitation of the experimental design we used here is that we did not actually try to establish which perceptual organizations participants experienced; all we asked them to report was which pattern they perceived as dominant. On the basis of our theoretical proposals (Winkler et al., 2012) and the CHAINS model (Mill et al., 2013) we may infer the perceptual organizations, but these predictions remain to be properly tested in future experiments.

The distribution of the patterns other than *integration* and *segregation* (i.e., AB--, BA--, A---) in relation to the stimulus parameters, and the latency with which they are discovered, provide support for our hypothesis that the *both* response found in previous experiments (Bendixen et al., 2010; Denham et al., 2013) can be explained by the perception of embedded patterns not previously considered, rather than rapid switching between *integration* and *segregation*. The results are not directly comparable, as the participants and range of stimulus parameters tested were somewhat different. However, qualitatively the relationship between stimulus parameters and the probability of reporting *both* seems to be well explained by a combination of the AB--, -BA-, and A--- patterns. *Both* is most commonly reported for small frequency differences and fast presentation rates, and is less common when the stimulus parameters strongly promote *segregation* or *integration* (Denham et al., 2013). Furthermore, *both* was hardly ever reported as the first percept (Denham et al., 2013). Here the median latency for reporting the AB--, -BA-, and A--- patterns tended to be rather long, and in all conditions in this experiment the median latency for at least one of the set {ABA-, A-, B---} was always less than the minimum median latency for the set {AB--, -BA-, A---}. These factors lead us to believe that the *both* response in previous experiments corresponds to the dominant perception of one of the following patterns AB--, -BA-, or A---.

It could be argued that participants only perceived all the patterns reported here because they were trained to do so, and that without this training they would not have heard patterns such as AB--, -BA-, or A---. There are a number of factors that argue against this objection. Firstly, in a pilot experiment (reported in Denham et al., 2013), listeners were asked to verbally report all repeating patterns that they experienced during

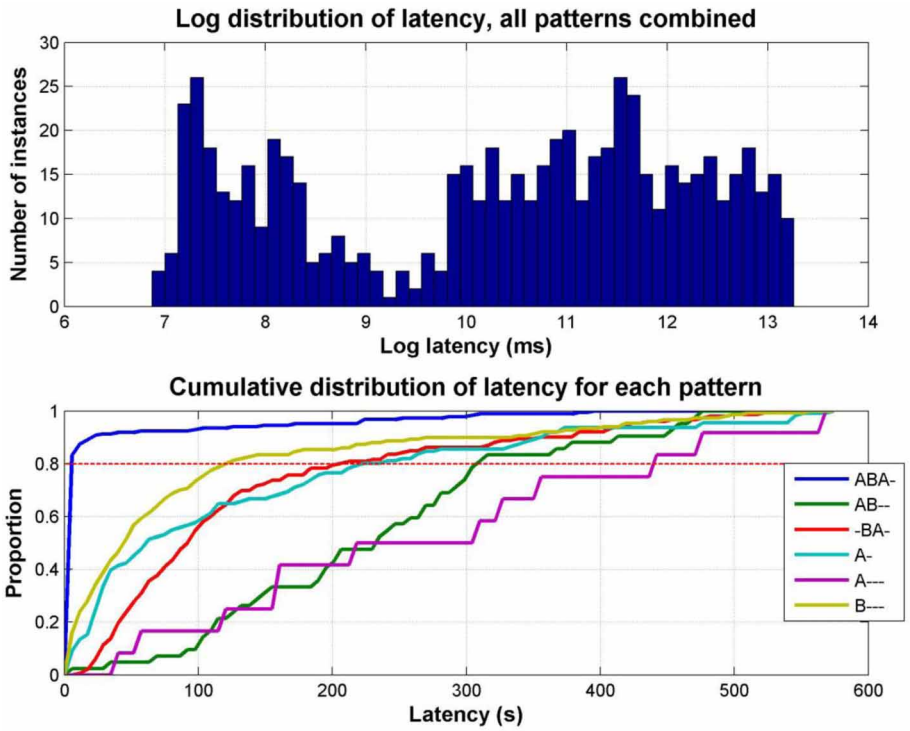


FIGURE 8 | Distribution of response latencies. Top: Distribution of latencies for all patterns combined, for all data in Experiment 1; for clarity latency is plotted as the log of the latency in milliseconds. **Bottom:** Cumulative distribution of latency for each pattern separately, combining data from all participants and sessions. Each line represents the proportion of latencies of the pattern indicated by the color code (see legend) less than the latency indicated along the x axis. For example, at a latency of 100 s, less than 20% of first reports of A--- have occurred, while more than 90% of first ABA- reports have occurred.

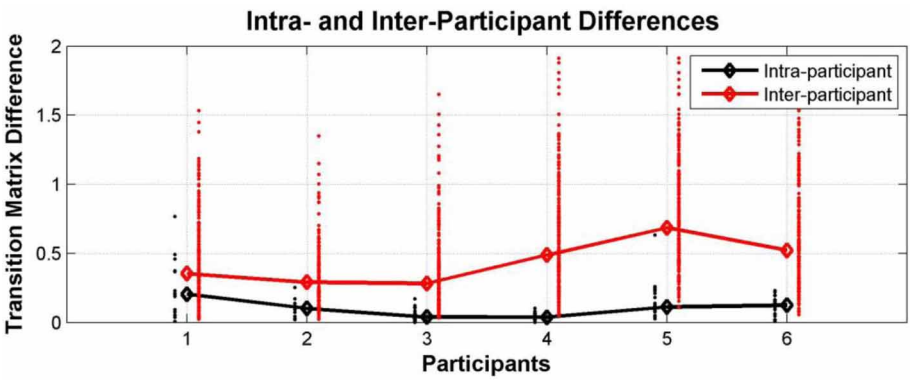


FIGURE 9 | Individual differences within Experiment 1. Comparison between intra- and inter-individual perceptual switching behavior, characterized by participant transition matrices extracted for each participant for each test session, pooling data from all conditions; differences between transition matrices from the same participant at different sessions (black dots; median—black diamonds, line), differences between the transition matrices of each participant and those of all other participants (red dots; median—red diamonds, line).

a 4-min long ABA- sequence. Most of the patterns trained in the current experiment were described spontaneously by listeners in this pilot study. Secondly, in our previous experiments *both* was reported without extensive pattern-specific prior training. Thirdly, the proportion and latency characteristics of *both* responses correspond well to those of the grouped {AB--, -BA-, A---} responses. On this basis, we argue here that *both* corresponds to one or other of the patterns AB--, -BA-, or A---. Finally, participants were instructed to adopt a neutral listening approach and not attempt to hear one or other pattern. Therefore, we would argue that the influence of training on the reports of these other patterns was largely limited to facilitating their categorization and reporting, rather than increasing their incidence.

Experiment 1 raised two important questions that we investigated in two follow-up experiments. Firstly, we decided to probe the stability of individual consistency in perceptual switching behavior and individual differences between participants over a longer time scale. If these are truly individual differences (i.e., differences of a physical nature), we would expect to find that the perceptual switching behavior of participants is similar even when tested in sessions separated by rather long periods of time, and that the individual differences found here also remain reliably detectable. Secondly, we were surprised by the relative prominence of the B--- pattern relative to the A- pattern. If anything the CHAINS model would predict that the A- pattern should be more prominent as it occurs more often (in CHAINS terms, makes more successful predictions per unit time) than the B--- one does. We hypothesized that the higher frequency of the B tones relative to the A tones in all conditions may have made the B tones perceptually more salient, and that this was the cause of the higher prominence of the B--- pattern. For these reasons we conducted two recall experiments (Experiments 2 and 3) approximately 1 year after Experiment 1.

EXPERIMENT 2

INTRODUCTION

Individuals are known to differ markedly in their perceptual behavior in visual multistability experiments; e.g., individual differences in perceptual switching rate in binocular rivalry have been known for many years (Aafjes et al., 1966). More recently, genetic markers (Miller et al., 2010) and differences in brain structure (e.g., Kanai et al., 2010; Genc et al., 2011), have been associated with differences in typical individual switching rates. Biases toward different perceptual decisions have also been reported and shown to relate, in the case of bistable motion, to differences in inter-hemispheric connectivity (Kanai and Rees, 2011). However, although there are some pointers toward such differences in audition (e.g., see Kondo and Kashino, 2009; Kashino and Kondo, 2012), a systematic investigation of the genetic or physiological basis for differences in auditory perceptual multistability has not yet been attempted. In Experiment 2, we sought to investigate whether the individual differences found in Experiment 1 were stable over a prolonged period. We hypothesized that if, as in vision, auditory individual perceptual differences are a result of stable physiological or even genetic differences, then we should find that the individual differences in perceptual switching behavior reported in Experiment 1 would be detectable 1 year later.

METHODS

Participants

Five (mean age 23.2 years, range 20–26 years; all right-handed; 3 male, 2 female) of the original six participants took part in Experiment 2, which was conducted over four sessions approximately 1 year after Experiment 1.

All equipment and procedures were as described for Experiment 1.

Training procedure

Participants attended one training session, which was similar in form to the training sessions for Experiment 1.

Testing procedure

Participants attended three test sessions over a duration of 2 weeks with at least 2 days between consecutive sessions during which the five 10-min experimental blocks were presented. All instructions and procedures were as for Experiment 1.

Analysis

To compare the effects of experiment (i.e., comparing Experiment 1 with the recall Experiment 2) we used a Two-Way repeated-measures ANOVA with two factors, experiment comparison (2 levels: same vs. other) × participant comparison (2 levels: self vs. other). For each of the 4 (2 × 2) comparisons we calculated the mean of all possible KL distances between the corresponding transition matrices. The ANOVA was computed on the resulting 20 values, i.e., 5 participants × 2 experiment levels × 2 participant levels.

RESULTS

From the entire data set for Experiment 2 the removal of phases where no button was pressed or duration was less than 300 ms resulted in the removal of 5.1% of the total data.

Participant transition matrices were constructed for each participant for each test session. **Figure 10** compares intra- and inter-individual differences in perceptual switching behavior in Experiments 1 and 2. The ANOVA analysis showed that the effect of participant (self vs. other) was highly significant [$F_{(1,\ 16)} = 20.49$, $p = 0.0003$], while the effect of experiment (1 vs. 2) was not significant [$F_{(1,\ 16)} = 0.38$, $p = 0.55$], and there was no interaction between these factors [$F_{(1,\ 16)} = 0.06$, $p = 0.81$].

INTERIM DISCUSSION

Individual differences in perceptual switching behavior remained consistent and detectable a year after the first experiment. All of the participants behaved in a way that was more similar to their behavior at all other sessions, including those separated by a year from each other, than to any of the other participants. This internal consistency over such a long period suggests that switching patterns in the current multistable auditory paradigm reflect some stable perceptual or higher-level traits, possibly stemming from physiological and maybe even genetic differences between listeners.

EXPERIMENT 3

INTRODUCTION

In Experiment 3 we investigated whether the frequency relation of the two tones (A lower than B, or *vice versa*) could influence the extent to which the A- and B--- patterns were reported as the foreground pattern by listeners. In Experiment 1 the frequency of the A tones was 400 Hz, and that of the B tones 476, 599, or 1008 Hz. Equal loudness curves of normal listeners (e.g., see Moore, 2003), suggest that the B tones would be perceptually louder than the A tones, and thus more salient. Therefore, in this experiment we decided to test whether switching the frequencies of the A and B tones would result in a greater tendency for listeners to report the A- than the B--- pattern.

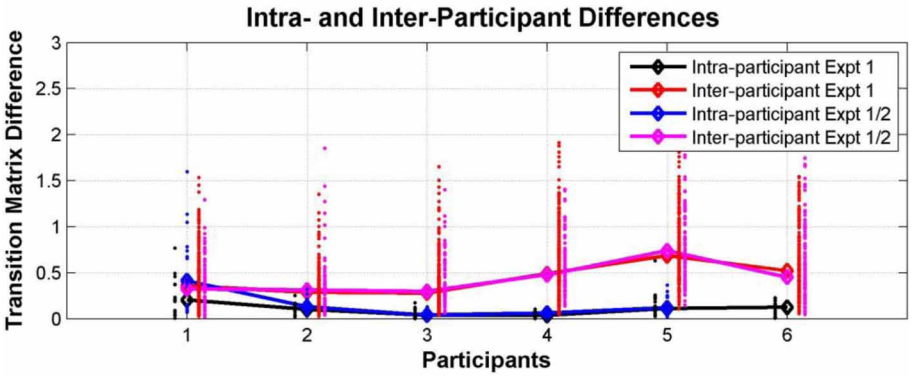


FIGURE 10 | Individual differences within and between Experiments 1 and 2. Comparison between intra- and inter-individual perceptual switching behavior, characterized by participant transition matrices extracted for each participant for each test session, pooling data from all conditions. Differences between transition matrices from the same participant in different sessions in Experiment 1 (black dots; median—black diamonds, line), differences between transition matrices of each participant in Experiment 1 vs. sessions from the same participant in Experiment 2 (blue dots; median—blue diamonds, line), differences between the transition matrices of each participant and those of all other participants in Experiment 1 (red dots; median—red diamonds, line), and differences between the transition matrices of each participant from Experiment 1 and those of all other participants in Experiment 2 (magenta dots; median—magenta diamonds, line).

METHODS

Testing

The same five participants as in Experiment 2 took part in three "reverse frequency" test sessions in Experiment 3. Stimuli were arranged in five conditions, with frequency difference (Δf) and stimulus onset asynchrony (SOA) as follows: (1) $\Delta f =$ 3 st, SOA = 100 ms; (2) $\Delta f = 16$ st, SOA = 100 ms; (3) $\Delta f =$ 7 st, SOA = 150 ms; (4) $\Delta f = 3$ st, SOA = 200 ms; (5) $\Delta f =$ 16 st, SOA = 200 ms. The "A" and "B" tones were delivered with a common duration of 75 ms (including 10 ms rise and fall times). The frequency of the "B" tones was 400 Hz, and the frequency of the "A" tones was n semitones higher, depending on condition. Participants were presented with all five conditions in a randomized order at each test session.

Analysis

To analyse the effect of changing the stimulus parameters (exchanging the A and B tone frequencies) a Two-Way repeated-measures ANOVA with factors experiment (1/3) and condition (1–5) was conducted. The dependent measure was the difference between the proportion of B--- and A- reported by each participant for each condition in Experiments 1 and 3. *Post-hoc* Wilcoxon rank sum tests were used to compare the effect of experiment in each condition separately. To analyse whether intra-individual similarities and inter-individual differences were preserved in Experiment 3, we performed the same Wilcoxon rank sum tests used for Experiment 1 to compare the distributions of intra- vs. inter-individual differences for each participant separately.

RESULTS

From the entire data set for Experiment 3 the removal of phases where no button was pressed or duration was less than 300 ms resulted in the removal of 5.3% of the total data.

Condition transition matrices were constructed as described for Experiment 1, and used to plot the mean proportion of each pattern as a function of condition. As predicted, exchanging the frequencies of the A and B tones resulted in the A- pattern becoming more prominent than the B--- pattern; see **Figure 11** in comparison with **Figure 5** (a repeated-measures Two-Way ANOVA with factors experiment (1/3) and condition (1–5) showed a significant effect of experiment on the difference between the proportion of B--- and A-, $F_{(4,\ 40)} = 16.75$, $p = 0.0002$. The interaction between experiment and condition was also significant, $F_{(4,\ 40)} = 3.85$, $p = 0.0097$. This was caused by the significant effect of experiment in condition 2 (Wilcoxon rank sum test $p = 0.03$), and no significant effect of experiment in other conditions (Wilcoxon rank sum test $p > 0.15$ for conditions 1, 3, 4, and 5).

We also examined intra- and inter- individual differences in Experiment 3 by constructing participant transition matrices as described in Experiment 1. As shown in **Figure 12**, these characteristics are still present when the frequencies are reversed; i.e., once more we find for all participants that, the switching behavior of each individual is more similar to their own behavior in a different test session than it is to the perceptual switching behavior of other participants (Wilcoxon rank sum test: $p = 0.02$, 0.008, 0.0003, 0.001, 0.007 for participants 1–5, respectively).

INTERIM DISCUSSION

Exchanging the A and B tone frequencies led to increased prominence of the A- pattern and decreased prominence of the B--- pattern. This supports the idea that perceptual saliency of the A and B tones (determined, amongst other factors, by their relative perceived loudness) can cause one or other of these patterns to become more dominant, i.e., occupy the perceptual foreground. This provides further support for a competition account of auditory streaming (Winkler et al., 2012), and suggests that models of

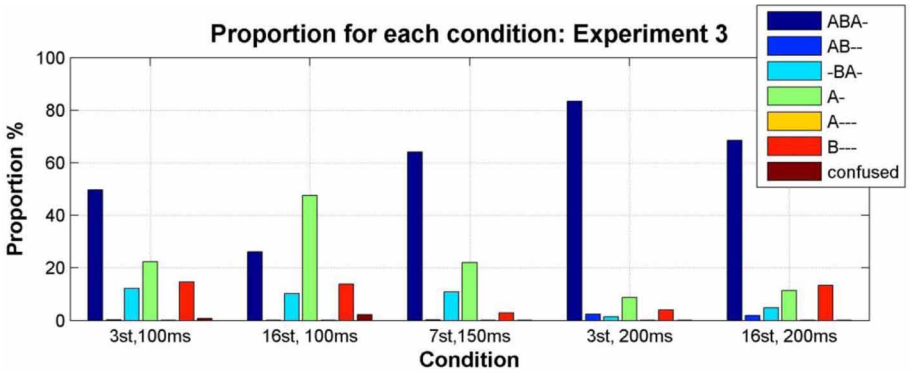


FIGURE 11 | Proportion of the total stimulus duration during which each pattern was perceived (color coded as indicated by the accompanying legend) for each condition, calculated from the condition transition matrices constructed by pooling the data from all participants for each condition (Böhm et al., 2013).

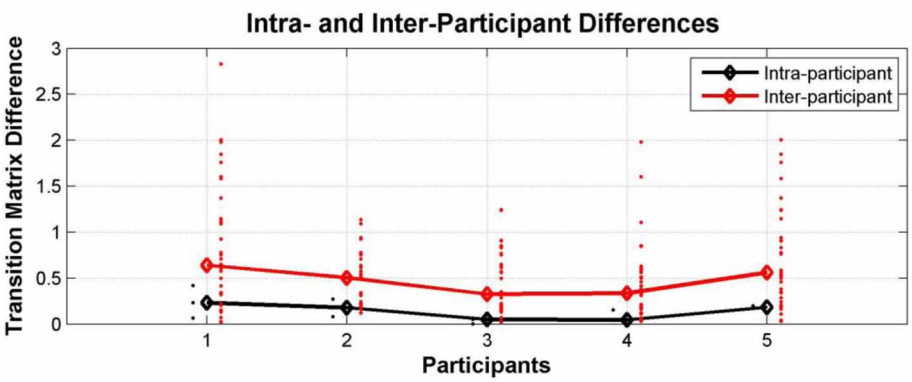


FIGURE 12 | Comparison between intra- and inter-individual perceptual switching behavior in Experiment 3; differences between transition matrices from the same participant at different sessions (black dots; median—black diamonds, line), and differences between the transition matrices of each participant and those of all other participants (red dots; median—red diamonds, line).

auditory scene analysis should also take account of the influence of event saliency on perceptual organization.

GENERAL DISCUSSION

The experiments presented here provide support for the hypothesis that auditory perception involves the extraction of patterns (regularities) from incoming sequences of sound events and that multiple patterns can be detected and held in parallel. Even when listening to simple repeating ABA- sequences, participants reported up to six different foreground patterns. This suggests that given sufficient time, pattern discovery appears to be exhaustive, i.e., all possible patterns within a certain length are perceived. We did not attempt to investigate whether there is a limit on the length of the patterns that are discovered and used by the auditory system to parse the auditory scene; this remains to be explored, probably with more complex sequences. However, our inclusion of patterns up to length four is supported by related ERP studies (Boh et al., 2011).

Once patterns have been discovered they come and go from conscious perception for as long as the stimulus sequence continues. This is consistent with the proposal that the contents of perceptual awareness are the result of an on-going competitive process between the set of patterns that have been discovered (Winkler et al., 2012). The distribution of these patterns in relation to the stimulus parameters, and the latency with which they were reported in Experiment 1 leads us to conclude that the *both* reports in previous experiments are consistent with the perception of foreground patterns AB--, -BA-, or A---. Based on our modeling studies we infer that the background pattern perceived in each of these cases was --A-, A---, and -BA-, respectively, although we did not attempt to investigate the perception of background patterns here.

Although bi-/multi-stable perceptual switching is highly stochastic, the switching patterns of individuals could be distinguished from each other. This is the first time to our knowledge that intra-individual similarities and inter-individual differences have been documented for *patterns* of perceptual switching in auditory streaming. However, individual differences in the number of perceptual switches have been previously reported (Kondo et al., 2012), and in vision individual differences in binocular rivalry have been known for some time (Aafjes et al., 1966). The method we use to distinguish individuals is different from the

switching rate measure that has previously been used. Here, we characterized the difference between two individuals in terms of a single distance measure between their transition matrices. However, much remains to be investigated regarding the details of these differences, which are likely to stem from some combination of switching rate, perceptual biases in the proportion of the various patterns perceived, and perhaps even higher order relationships such as idiosyncratic perceptual transitions.

The finding that individuals behave in a measurably consistent manner even between sessions separated by a year leads us to suggest that relatively stable bases such as anatomical or genetic differences, similar to those found for other multi-stable phenomena (Kanai et al., 2010; Genc et al., 2011), may be responsible. However, the neural correlates of perceptual dominance and perceptual switching in auditory perception are not yet well understood, although there is some reason to suppose that there may be some aspects that are shared with vision (e.g., Cusack, 2005). We suggest that the paradigm and analysis methods we present here may prove useful in the future for investigations of the neural basis for auditory perceptual organization.

Another question of interest for future investigations is whether, and if so what, other individual perceptual or cognitive characteristics are related to these individual patterns of switching behavior. If our assumption regarding the mechanisms underlying auditory perception are correct (i.e., discovery and on-going competition between alternative interpretations of the input), then these individual differences may underlie variation in other characteristics, such as perceptual abilities, cognitive style, personality or creativity.

The results of these experiments have implications for models of auditory scene analysis; in particular, any comprehensive model should account for the multitude of patterns that participants report, the parameter dependence of the pattern distributions, and the latencies with which they are discovered. No model has been developed yet which can account for all of these aspects. The popular temporal coherence model in its current formulation makes a fixed perceptual decision (e.g., Shamma et al., 2011). While this could undoubtedly be modulated by the introduction of noise and adaptation, it is more difficult to see how the other embedded patterns reported here, e.g., -BA- could be discovered, as the temporal coherence measure would either group or not group A's with B's. The CHAINS model (Mill et al., 2013) in its current formulation cannot discover the AB--, -AB-, or A--- patterns reported here either, but this is easily fixed with a simple modification to the pattern discovery function. However, there is a more fundamental problem; since the links in CHAINS form probabilistically between events, patterns involving two events, e.g., AB--, will be easier to discover than patterns involving three events, ABA-. Therefore, CHAINS would predict that AB-- is reported with a shorter latency than ABA-, which is not the case.

CONCLUSION

Although on the face of it the results we present here appear to challenge our everyday experience of perception as stable and veridical, we suggest that it is precisely by having the ability to construct multiple interpretations of a scene that perception is able to achieve robust performance. The first reported pattern corresponds to the most likely interpretation; here, typically that there is one object in the world producing sounds of different pitch but similar timbre, although if the frequency difference becomes too great then the possibility of two sound sources becomes more likely. It is important to note that this initial perceptual decision is not simply a function of the physical stimulus; relevant contextual information (Snyder et al., 2008, 2009b) and prior learning (van Zuijen et al., 2005; Snyder et al., 2009a) also exert an influence. It also makes sense that perception never fixes on a single solution. If either of these were the case then our perceptions would be entirely determined by external factors, leaving no room for autonomous behavior. Since perception is essentially about trying to extract information from the world around us, a better strategy than simply choosing the "best" interpretation is to explore other interpretations if time allows in case they offer insights not available in the most likely scenario. By setting up a competition between alternatives and ensuring that no solution can dominate forever, the perceptual system essentially performs a probabilistic likelihood sampling of the perceptual space (Moreno-Bote et al., 2011). This view of perception resonates with a number of earlier perceptual theories, including Helmholtz's view of the role of inferential processes in perception (Helmholtz, 1885), Gregory's notion of perception as hypotheses (Gregory, 1980), and more recent instantiations in work on predictive coding theory by Friston and colleagues amongst others (e.g., Friston and Kiebel, 2009). Competition between patterns, rather than individual elements, also fits well with Gestalt ideas of perception that emphasize the importance of the whole (i.e., patterns) relative to the parts (i.e., individual tones) (Köhler, 1947; Wagemans et al., 2012). In conclusion, studies of perceptual multistability can provide new and useful insights into general mechanisms as well as individual differences in human perception.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fnins.2014.00025/abstract

REFERENCES

Aafjes, M., Hueting, J. E., and Visser, P. (1966). Individual and interindividual differences in binocular retinal rivalry in man. *Psychophysiology* 3, 18–22. doi: 10.1111/j.1469-8986.1966.tb02674.x

Anstis, S., and Saida, S. (1985). Adaptation to auditory streaming of frequency-modulated tones. *J. Exp. Psychol. Hum. Percept. Perform.* 11, 257–271. doi: 10.1037/0096-1523.11.3.257

Bendixen, A., Bõhm, T. M., Szalárdy, O., Mill, R., Denham, S. L., and Winkler, I. (2013). Different roles of similarity and predictability in auditory stream segregation. *Learn. Percept.* 2, 37–54. doi: 10.1556/LP.5.2013.Suppl2.4

Bendixen, A., Denham, S. L., Gyimesi, K., and Winkler, I. (2010). Regular patterns stabilize auditory streams. *J. Acoust. Soc. Am.* 128, 3658–3666. doi: 10.1121/1.3500695

Boh, B., Herholz, S. C., Lappe, C., and Pantev, C. (2011). Processing of complex auditory patterns in musicians and nonmusicians. *PLoS ONE* 6:e21458. doi: 10.1371/journal.pone.0021458

Bõhm, T. M., Shestopalova, L., Bendixen, A., Andreou, A. G., Georgiou, J., Garreau, G., et al. (2013). Spatial location of sound sources biases auditory stream segregation but their motion does not. *J. Learn. Percept.* 5(Suppl 2), 55–72. doi: 10.1556/LP.5.2013.Suppl2.5

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Cusack, R. (2005). The intraparietal sulcus and perceptual organization. *J. Cogn. Neurosci.* 17, 641–651. doi: 10.1162/0898929053467541

Deike, S., Heil, P., Böckmann-Barthel, M., and Brechmann, A. (2012). The build-up of auditory stream segregation: a different perspective. *Front. Psychol.* 3:461. doi: 10.3389/fpsyg.2012.00461

Denham, S., Bendixen, A., Mill, R., Tóth, D., Wennekers, T., Coath, M., et al. (2012). Characterising switching behaviour in perceptual multi-stability. *J. Neurosci. Methods* 210, 79–92. doi: 10.1016/j.jneumeth.2012.04.004

Denham, S. L., Gyimesi, K., Stefanics, G., and Winkler, I. (2013). Multistability in auditory stream segregation: the role of stimulus features in perceptual organisation. *Learn. Percept.* 2, 73–100. doi: 10.1556/LP.5.2013.Suppl2.6

Denham, S. L., and Winkler, I. (2006). The role of predictive models in the formation of auditory streams. *J. Physiol. Paris* 100, 154–170. doi: 10.1016/j.jphysparis.2006.09.012

Denham, S. L., and Winkler, I. (2014). "Auditory perceptual organization," in *Oxford Handbook of Perceptual Organization*, ed J. Wagemans (Oxford: Oxford University Press). (in press).

Friston, K., and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364, 1211–1221. doi: 10.1098/rstb.2008.0300

Genc, E., Bergmann, J., Tong, F., Blake, R., Singer, W., and Kohler, A. (2011). Callosal connections of primary visual cortex predict the spatial spreading of binocular rivalry across the visual hemifields. *Front. Hum. Neurosci.* 5:161. doi: 10.3389/fnhum.2011.00161

Gregory, R. L. (1980). Perceptions as hypotheses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 290, 181–197. doi: 10.1098/rstb.1980.0090

Haywood, N. R., and Roberts, B. (2011). Effects of inducer continuity on auditory stream segregation: comparison of physical and perceived continuity in different contexts. *J. Acoust. Soc. Am.* 130, 2917–2927. doi: 10.1121/1.3643811

Haywood, N. R., and Roberts, B. (2013). Build-up of auditory stream segregation induced by tone sequences of constant or alternating frequency and the resetting effects of single deviants. *J. Exp. Psychol. Hum. Percept Perform.* 39, 1652–1666. doi: 10.1037/a0032562

Helmholtz, H. V. (1885). *On the Sensations of Tone as a Physiological Basis for the Theory Of Music.* London: Longmans, Green, and Co.

Horváth, J., Czigler, I., Sussman, E., and Winkler, I. (2001). Simultaneously active pre-attentive representations of local and global rules for sound sequences in the human brain. *Brain Res. Cogn. Brain Res.* 12, 131–144. doi: 10.1016/S0926-6410(01)00038-6

Jones, M. R. (1976). Time, our lost dimension: toward a new theory of perception, attention, and memory. *Psychol. Rev.* 83, 323–355. doi: 10.1037/0033-295X.83.5.323

Kanai, R., Bahrami, B., and Rees, G. (2010). Human parietal cortex structure predicts individual differences in perceptual rivalry. *Curr. Biol.* 20, 1626–1630. doi: 10.1016/j.cub.2010.07.027

Kanai, R., and Rees, G. (2011). The structural basis of inter-individual differences in human behaviour and cognition. *Nat. Rev. Neurosci.* 12, 231–242. doi: 10.1038/nrn3000

Kashino, M., and Kondo, H. M. (2012). Functional brain networks underlying perceptual switching: auditory streaming and verbal transformations. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 977–987. doi: 10.1098/rstb.2011.0370

Köhler, W. (1947). *Gestalt Psychology: An Introduction To New Concepts in Modern Psychology*. New York, NY: Liveright Publishing Corporation.

Kondo, H. Kitagawa, M., N., Kitamura, M. S., Koizumi, A., Nomura, M., and Kashino, M. (2012). Separability and commonality of auditory and visual bistable perception. *Cereb. Cortex* 22, 1915–1922. doi: 10.1093/cercor/bhr266

Kondo, H. M., and Kashino, M. (2009). Involvement of the thalamocortical loop in the spontaneous switching of percepts in auditory streaming. *J. Neurosci.* 29, 12695–12701. doi: 10.1523/JNEUROSCI.1549-09.2009

Kullback, S. (1959). *Information Theory and Statistics.*. New York, NY: John Wiley and Sons.

Levelt, W. J. M. (1968). *On Binocular Rivalry*. Paris: Mouton.

Logothetis, N. K., Leopold, D. A., and Sheinberg, D. L. (1996). What is rivalling during binocular rivalry? *Nature* 380, 621–624. doi: 10.1038/380621a0

Mamassian, P., and Goutcher, R. (2005). Temporal dynamics in bistable perception. *J. Vis.* 5, 361–375. doi: 10.1167/5.4.7

Mill, R. W., Bõhm, T. M., Bendixen, A., Winkler, I., and Denham, S. L. (2013). Modelling the emergence and dynamics of perceptual organisation in auditory streaming. *PLoS Comput. Biol.* 9:e1002925. doi: 10.1371/journal.pcbi.1002925

Miller, S. M., Hansell, N. K., Ngo, T. T.,. Liu, G. B., Pettigrew, J. D., Martin, N. G., et al. (2010). Genetic contribution to individual variation in binocular rivalry rate. *Proc. Natl. Acad. Sci. U.S.A.* 107, 2664–2668. doi: 10.1073/pnas.0912149107

Moore, B. C., and Gockel, H. E. (2012). Properties of auditory stream formation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 919–931. doi: 10.1098/rstb.2011.0355

Moore, B. C. J. (2003). *An Introduction to the Psychology of Hearing.* 5th edn. Amsterdam: Academic Press.

Moreno-Bote, R., Knill, D. C., and Pouget, A. (2011). Bayesian sampling in visual perception. *Proc. Natl. Acad. Sci. U.S.A.* 108, 12491–12496. doi: 10.1073/pnas.1101430108

Moreno-Bote, R., Shpiro, A., Rinzel, J., and Rubin, N. (2010). Alternation rate in perceptual bistability is maximal at and symmetric around equi-dominance. *J. Vis.* 10, 1. doi: 10.1167/10.11.1

Pressnitzer, D,. and Hupé, J. M. (2006). Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Curr. Biol.* 16, 1351–1357. doi: 10.1016/j.cub.2006.05.054

Rogers, W. L., and Bregman, A. S. (1993). An experimental evaluation of three theories of auditory stream segregation. *Percept Psychophys* 53, 179–189. doi: 10.3758/BF03211728

Shamma, S. A., Elhilali, M., and Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends Neurosci.* 34, 114–123. doi: 10.1016/j.tins.2010.11.002

Snyder, J. S., Carter, O. L., Hannon, E. E., and Alain, C. (2009a). Adaptation reveals multiple levels of representation in auditory stream segregation. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 1232–1244. doi: 10.1037/a0012741

Snyder, J. S., Carter, O. L., Lee, S. K., Hannon, E. E., and Alain, C. (2008). Effects of context on auditory stream segregation. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 1007–1016. doi: 10.1037/0096-1523.34.4.1007

Snyder, J. S., Holder, W. T., Weintraub, D. M., Carter, O. L., and Alain, C. (2009b). Effects of prior stimulus and prior perception on neural correlates of auditory stream segregation. *Psychophysiology* 46, 1208–1215. doi: 10.1111/j.1469-8986.2009.00870.x

Szalárdy, O., Bendixen, A., Tóth, D., Denham, S. L., and Winkler, I. (2013). Modulation-frequency acts as a primary cue for auditory stream segregation. *Learn. Percept.* 5(Suppl 2), 149–161. doi: 10.1556/LP.5.2013.Suppl2.9

van Noorden, L. P. A. S. (1975). *Temporal Coherence in the Perception of Tone Sequences*. Doctoral dissertation, Technical University Eindhoven.

van Zuijen, T. L., Sussman, E., Winkler, I., Näätänen, R., and Tervaniemi, M. (2005). Auditory organization of sound sequences by a temporal or numerical regularity—a mismatch negativity study comparing musicians and non-musicians. *Brain Res. Cogn. Brain Res.* 23, 270–276. doi: 10.1016/j.cogbrainres.2004.10.007

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M. et al. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychol. Bull.* 138, 1172–1217. doi: 10.1037/a0029333

Winkler, I., Denham, S., Mill, R., Bõhm, T. M., and Bendixen, A. (2012). Multistability in auditory stream segregation: a predictive coding view. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 1001–1012. doi: 10.1098/rstb.2011.0359

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

Received: 30 November 2013; paper pending published: 03 January 2014; accepted: 27 January 2014; published online: 28 February 2014.

Citation: Denham S, Bõhm TM, Bendixen A, Szalárdy O, Kocsis Z, Mill R and Winkler I (2014) Stable individual characteristics in the perception of multiple embedded patterns in multistable auditory stimuli. Front. Neurosci. ***8****:25. doi: 10.3389/fnins.2014.00025*

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

frontiers in
NEUROSCIENCE

REVIEW ARTICLE
published: 31 March 2014
doi: 10.3389/fnins.2014.00060

Predictability effects in auditory scene analysis: a review

Alexandra Bendixen*

Auditory Psychophysiology Lab, Department of Psychology, Cluster of Excellence "Hearing4all," European Medical School, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany

Edited by:
Susann Deike, Leibniz Institute for Neurobiology, Germany

Reviewed by:
Thomas F. Münte, University of Magdeburg, Germany
Michael Brosch, Leibniz Institute for Neurobiology, Germany

****Correspondence:***
Alexandra Bendixen, Auditory Psychophysiology Lab, Department of Psychology, Cluster of Excellence "Hearing4all," European Medical School, Carl von Ossietzky University of Oldenburg, D-26111 Oldenburg, Germany
e-mail: alexandra.bendixen@uni-oldenburg.de

Many sound sources emit signals in a predictable manner. The idea that predictability can be exploited to support the segregation of one source's signal emissions from the overlapping signals of other sources has been expressed for a long time. Yet experimental evidence for a strong role of predictability within auditory scene analysis (ASA) has been scarce. Recently, there has been an upsurge in experimental and theoretical work on this topic resulting from fundamental changes in our perspective on how the brain extracts predictability from series of sensory events. Based on effortless predictive processing in the auditory system, it becomes more plausible that predictability would be available as a cue for sound source decomposition. In the present contribution, empirical evidence for such a role of predictability in ASA will be reviewed. It will be shown that predictability affects ASA both when it is present in the sound source of interest (perceptual foreground) and when it is present in other sound sources that the listener wishes to ignore (perceptual background). First evidence pointing toward age-related impairments in the latter capacity will be addressed. Moreover, it will be illustrated how effects of predictability can be shown by means of objective listening tests as well as by subjective report procedures, with the latter approach typically exploiting the multi-stable nature of auditory perception. Critical aspects of study design will be delineated to ensure that predictability effects can be unambiguously interpreted. Possible mechanisms for a functional role of predictability within ASA will be discussed, and an analogy with the old-plus-new heuristic for grouping simultaneous acoustic signals will be suggested.

Keywords: predictive coding, sound processing, auditory stream segregation, integration, bistable perception, old-plus-new heuristic

INTRODUCTION: AUDITORY SCENE ANALYSIS PRINCIPLES

Our auditory system is often confronted with a mixture of sounds originating from several different sources. Relevant auditory information can only be retrieved if the system succeeds in decomposing this mixture into meaningful perceptual units termed *streams* (Bregman, 1990). Originally introduced as the *cocktail party problem* (Cherry, 1953), the sound stream decomposition problem was later coined *auditory scene analysis (ASA)* in an influential monograph by Bregman (1990). Theories and computational models of ASA increasingly incorporate structural and functional knowledge about the auditory system coming from other research areas (e.g., Haykin and Chen, 2005). In this spirit, the present contribution will link *predictive processing*, a central property of the auditory system extensively investigated in the neurosciences (e.g., Friston, 2005, 2010), with an important principle of ASA, the *old-plus-new heuristic*. On theoretical and empirical grounds, it will be argued that the formation and maintenance of stable sound source representations is facilitated by the capacity of the auditory system to extract predictability from the sound sources' signal emissions.

Disentangling a sound mixture requires inferring likely sources from physically overlapping signals. Two different types of signal decomposition are needed. First, for signals occurring at the same time, the listener must interpret whether the same or different sound source(s) emitted them. This process, called *concurrent* or *vertical grouping* (Bregman, 1990), rests on auditory cues such as location, harmonicity, and onset synchrony of the different components of a mixture (e.g., McDonald and Alain, 2005; Lipp et al., 2010; for review, see Micheyl and Oxenham, 2010b). As these cues are immediately informative of the possible relations between co-occurring auditory signals, they are called *simultaneous* or *instantaneous* cues.

Other cues carry no information in and of themselves, but are informative only in comparison with previous input. One such cue might be that several tones in succession all occupy the same frequency region. This second type of signal decomposition is termed *horizontal* or *sequential grouping* (Bregman, 1990). It is concerned with interpreting the relations between different auditory signals following each other in time. Again, the listener must interpret whether these signals were emitted by the same or different sound source(s). The focus of the present contribution is on this second situation, where the ambiguity lies in the spreading out of sound signals over time rather than in their physical overlap at one particular moment.

Indeed, many natural sound sources emit signals in a temporally discontinuous manner (e.g., a series of footsteps, or an utterance containing speech-inherent pauses). The auditory system of the listener is then confronted with discrete sound events that need to be bound together (*stream integration*). At the same time, binding events together that were actually emitted by two different sources needs to be avoided (*stream segregation*). The perceptual decision as to whether sounds in a mixture should

be grouped together or not has been suggested to be based on a number of heuristics, many of them originating from the Gestalt school of psychology (e.g., Wertheimer, 1923; Köhler, 1947). The feature *similarity* principle posits that sounds are more likely to have been emitted by the same source if they are similar in all of their acoustic features (for detailed discussion, see Moore and Gockel, 2002, 2012). Feature (dis)similarity is further evaluated against the temporal separation between consecutive sound events (Van Noorden, 1975) as sound sources rarely change the features of their sound emissions abruptly, but may do so over time. This relation can be expressed in quantitative terms by the temporal rate of feature change (Jones, 1976). Feature similarity can then be quantified as the inverse of this rate (Winkler et al., 2012; Mill et al., 2013). In doing so, the Gestalt rule of *similarity* merges into the principle of *continuity*, expressing the notion that sound sources show smooth variation over time rather than abrupt changes (Bregman, 1990).

Another important principle of ASA is called the *old-plus-new heuristic*. This heuristic was formulated by Bregman (1990, p. 222; going back to Helmholtz, 1859, p. 59f.) to express the idea that the auditory system, when confronted with a mixture of sounds, first looks for continuations of previous sounds and removes these from the mixture, and then analyzes the residue. According to Bregman's (1990) description, "mixture" in this case refers to the actual physical overlap of sounds; and "continuation" means uninterrupted continuation, i.e., one and the same sound that goes on while other sounds are added. With these specifications, it becomes evident that the old-plus-new heuristic is a cue for vertical (simultaneous) sound grouping. One may, however, easily have the idea that the same could be true for interrupted (i.e., discrete) forms of continuation, such as a sound source emitting the same sound event regularly every so many milliseconds (ms). An illustrative example is the repetitive acoustic signature of a train moving on the rails. In this situation, giving the continuation of old sound sources (the train sound) precedence over the identification of new sources (e.g., someone opening the cabin door) would again lead to a plausible decomposition of the auditory scene into known (old) and unknown (new and potentially relevant) information. The old-plus-new heuristic would then transfer from vertical to horizontal (sequential) sound grouping.

The Gestalt principles of *similarity* and *continuity* appear conceptually similar to the old-plus-new heuristic, in that they express the idea that a sound source continues with relatively unchanged attributes. However, in fact they do not constitute a sequential analogue to the old-plus-new heuristic. Instead, they are "unspecific" variants of the old-plus-new heuristic in that they do not distinguish between exact continuation and inexact but plausible continuation of the sound source's behavior (Bregman, 1990). In contrast, the old-plus-new heuristic appears to be based on the system assuming a very precise continuation for uninterrupted sounds (Darwin, 1995). Furthermore, the old-plus-new heuristic implies that sound continuations are subtracted from a mixture before running any other decomposition algorithms. In contrast, continuity does not take precedence over other grouping cues: The plausibility of continuation is checked only *after* the sequence has been partitioned (Rogers and Bregman, 1993). Hence according to Bregman's (1990) framework, vertical sound grouping appears to be equipped with a more powerful mechanism of dealing with signal continuation than horizontal sound grouping. This might be partly due to the fact that at the time, much less was known about the auditory system's capacity to detect continuity in a sequence of discrete sounds. Within the last two decades, this topic has experienced an upsurge in interest and research activity, leading to a much more comprehensive picture of how continuous (regular) properties can be extracted from discrete sounds.

PREDICTIVE AUDITORY PROCESSING

The concepts and empirical findings outlined in this section initially developed independently from ASA research. They were based on the experimental observation (Näätänen et al., 1978) and later theoretical conceptualization (Näätänen, 1992) of the auditory system's capacity to detect deviations from otherwise constant features in a sequence of discrete sounds. In a typical study, single tones of constant frequency (*standards*) would be repetitively presented. Occasionally, the frequency of one tone would be changed. These *deviant* tones would elicit a specific brain response originating in auditory cortical areas (Giard et al., 1990; Opitz et al., 2002): the mismatch negativity (MMN) component of the event-related potential (ERP). MMN was interpreted to indicate "mismatch" or *deviance* detection in otherwise regular sound sequences (Näätänen et al., 1978; see also Snyder and Hillyard, 1976).

Since the detection of deviant sounds requires the prior recognition of an invariant property in the standard sounds (e.g., their constant frequency), the argument went on to suggest that MMN elicitation can be taken as indirect evidence of *invariance* extraction (e.g., Picton et al., 2000). Following observations that not only constant feature values but also feature values changing in a regular manner are encoded as standards (e.g., alternation between feature values, "ABAB…"), the term "invariance" was replaced by "*regularity*" (Winkler, 2007). This leads to the notion of "regularity extraction" from sound sequences, which is equivalent to the detection of continuous properties in a source's discrete signal emissions referred to above. The types of regularities that can be extracted have been intensely studied, demonstrating the enormous potential of the auditory system to detect relations of various degrees of complexity between successive sounds in a sequence (cf. reviews by Näätänen et al., 2001, 2010).

Finally, the notion was put forward that regularity extraction is conceptually identical to the extraction of *predictability* from a sound sequence (Tiitinen et al., 1994; Winkler et al., 1996). It was further argued that the extracted information is used for *predicting* upcoming sounds (Baldeweg, 2006). This notion has gained momentum with the advent of predictive coding theory (Friston, 2005, 2010). Processing sensory input in a predictive manner has become an important element of general theories of perception (Gregory, 1980; Friston, 2005, 2010; Prinz, 2006; Schubotz, 2007). Abundant empirical evidence for predictive processing in the auditory system has been gathered, which is partly based on MMN and partly on more direct ERP indicators (for a recent review, see Bendixen et al., 2012a).

Whereas there is not yet consensus as to the precise underlying neuronal mechanisms and the terminology best used to

describe the relevant phenomena (e.g., Näätänen et al., 2005; May and Tiitinen, 2010), it is undisputed that the auditory system effortlessly acquires information about the regular structure of the surrounding sound sources. The term "effortlessly" is meant to imply that this information comes as an inherent property of auditory sensory information processing (as opposed to being made available only by actively searching for it). There is also general agreement that information about the regular characteristics of sound sources is available to the auditory system at early processing stages—within 150 ms after sound onset, when considering the latency of the MMN. More recent findings imply that the information is available even earlier, within 20–40 ms after sound onset (Grimm et al., 2011; cf. review by Escera et al., in press). In fact, the predictive notion implies that such information should be available even *before* the onset of the next signal emitted by a given sound source (Baldeweg, 2006; Bendixen et al., 2009, 2012a).

IMPLICATIONS OF PREDICTIVE PROCESSING FOR AUDITORY SCENE ANALYSIS

The early availability of information on the regular characteristics of sound sources implies that predictability could, in theory, be used as an early cue in ASA. If auditory input is indeed processed in a predictive manner, information about the predictable succession of sound events could act upon ASA processes before any other grouping cue would be able to exert its influence. This is because all other cues need at least some rudimentary analysis of stimulus input, whereas prediction-based grouping could start at or even before the expected time of stimulus arrival. Therefore, predictability could in theory be the earliest grouping cue in ASA. Predictability would then constitute a sequential analogue to the old-plus-new heuristic in simultaneous sound grouping.

Such considerations have led some researchers to propose theories linking predictive processing and sequential ASA (Denham and Winkler, 2006; Winkler, 2007; Winkler et al., 2009). It should be pointed out that in previous theoretical frameworks, the impact of predictive regularities on ASA was not negated altogether but was ascribed to high-level, "schema-based" ASA processes requiring the involvement of attention (Bregman, 1990, p. 411ff.). The idea of an attention-mediated role of predictability within ASA was strongly driven by the work of Mari Riess Jones and her colleagues on rhythmic attending (Jones, 1976; Jones et al., 1981, 1982; Jones and Boltz, 1989; Drake et al., 2000). It was, for instance, shown that an immediately apparent regularity (e.g., rhythmicity) in the emissions of a sound source helps listeners to focus their attention on that sound source. The impact of the regularity was, however, regarded as confined to a second stage of ASA, where top-down selections between the groupings offered by the first stage are made (Bregman, 1990; Bey and McAdams, 2002). In contrast, the possibility of predictability entering low-level, "primitive" ASA processes was distinctly ruled out (Bregman, 1990, p. 135f.; cf. p. 417ff. for a summary of the arguments). It seems justified to re-examine this conclusion based on the insights into auditory predictability extraction that have been gained within the last 20 years.

One might argue that the insights into predictive processing are not transferrable to ASA because almost all findings were obtained in artificially simplified experimental situations with only one active sound source. Indeed, the vast majority of MMN studies employed only one sequence of sounds in which a single regularity was embedded. Yet some studies have shown that the MMN-eliciting system can monitor regularities in at least three sound streams simultaneously (Nager et al., 2003; Winkler et al., 2003c; Sussman et al., 2005). Together with the complex forms of predictability that can be discovered without attentional involvement (Näätänen et al., 2001, 2010), it seems that supposedly "primitive" auditory analyses can process much more complex scenarios than previously assumed (Nelken, 2004).

Consequently, the hypothesis has been put forward that predictability contributes to the initial decomposition of the auditory input, and that the predictive models underlying MMN generation form the basis of the sequential grouping processes underlying ASA (Denham and Winkler, 2006; Winkler, 2007; Winkler et al., 2009). This idea nicely converges with the fact that predictive elements have successfully been implemented in computational modeling approaches of ASA (Ellis, 1999; Godsmark and Brown, 1999; Masuda-Katsuse and Kawahara, 1999; Grossberg et al., 2004; Coy and Barker, 2007). One must, however, concede that a functional role of predictability within ASA has not received much empirical support. Indeed, the results of early ASA studies (French-St. George and Bregman, 1989; Rogers and Bregman, 1993) suggest quite the opposite: that sound predictability does *not* affect auditory stream segregation or integration. These early studies will be examined in the following.

EARLY STUDIES ON PREDICTABILITY EFFECTS IN AUDITORY SCENE ANALYSIS

An influential study taken to demonstrate that ASA is unaffected by predictability was carried out by French-St. George and Bregman (1989). These authors presented a sequence of regularly alternating "A" and "B" sounds (i.e., "ABABABAB"), where "A" and "B" denote categories of sounds that could take one of four different frequency values each (A1–A4, B1–B4). The order of sounds was either arranged in a predictable manner (e.g., continuous repetition of the eight-tone cycle "A1B4A3B2A2B3A4B1") or chosen randomly anew for each eight-tone cycle (see **Figure 1** for illustration). Moreover, the delivery of the sounds was either isochronous (thus temporally predictable), or non-isochronous with unpredictable temporal intervals. The authors investigated whether these two manipulations of sound predictability would have an impact on participants' ability to perceive all the sounds together in one stream ("Integrated") as determined by subjective perceptual report. Results revealed that neither of the predictability manipulations had any effect on the perception of the sequence (French-St. George and Bregman, 1989). The authors concluded that predictability does not support the perceptual coherence of a putative sound stream.

These results were taken to imply that ASA is unaffected by sound predictability (French-St. George and Bregman, 1989). There is, however, a crucial confound in the experimental design as noted by the authors themselves (p. 386): Although the predictability manipulation was designed to pertain to the whole sequence of sounds ("ABABABAB"), this co-varied

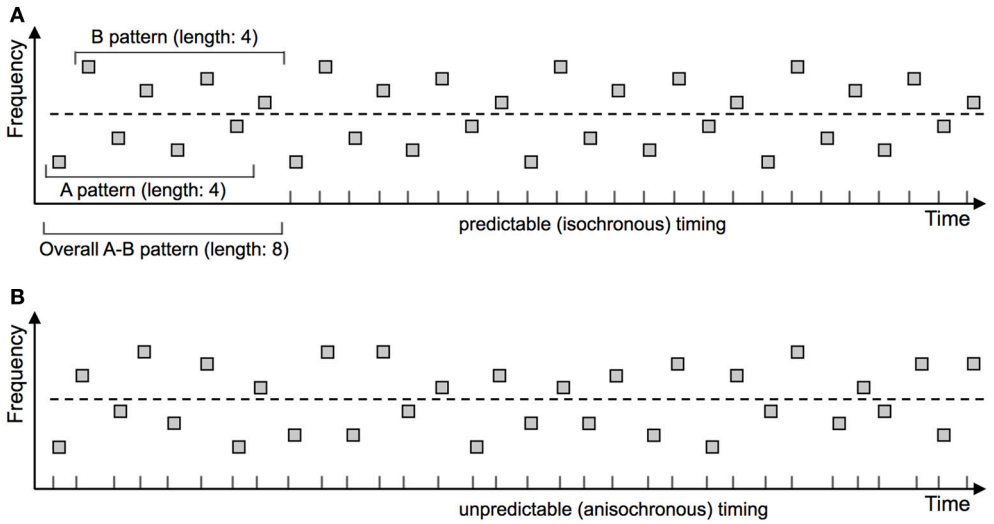


FIGURE 1 | Experimental paradigm confounding predictability of the "Integrated" and "Segregated" perceptual organizations. French-St. George and Bregman (1989) compared stimulus arrangements that were predictable or unpredictable with respect to stimulus timing and frequency. Sequences were predictable on both dimensions **(A)**, on neither dimension **(B)**, or on just one of the dimensions (not depicted). The cyclically repeating (thereby predictable) frequency patterns are marked in the upper panel. Stimulus timing is additionally marked by corresponding ticks on the X axis. The dashed line indicates the (nominal) separation between the "A" and "B" groups of tones. Participants were asked to try perceiving all tones as originating from one sound source and to press a button as long as they succeed in maintaining this percept. Predictability was assumed to increase the perceptual coherence of the tone set, thereby leading to a higher probability of perceiving the sequence as one stream ("Integrated"). Yet as illustrated in the upper panel, adding predictability to the sequence as a whole unavoidably renders the two separate streams more predictable, too. Thus their individual perceptual coherence might increase as well, leading to a higher probability of perceiving the sequence as "Segregated." These opposite effects might cancel each other out, leading to a null effect on average that would not be indicative of a general absence of predictability effects in ASA.

with the predictability of the two sub-streams ("A-A-A-A-" and "-B-B-B-B" separately). Therefore, if predictability acts as a cue favoring decompositions with predictable behavior of the resulting sound sources, in this specific design predictability would have favored not only the "Integrated" but also the "Segregated" perceptual organization. The null effect observed by the authors might have been due to these opposite effects canceling each other out. The authors suggest that experiments with independent manipulations of the predictability of the overall sequence and of the separate sub-streams are needed (French-St. George and Bregman, 1989).

A later study by Rogers and Bregman (1993) manipulated predictability in a different way, by presenting an induction sequence before the test sequence to be perceptually judged by the listener ("A_A_A_A_A_ABA_ABA_ABA," where "A_A_A" is the induction sequence, and "ABA_" is the test sequence). Predictability of the induction sequence was manipulated in terms of the occurrence times and duration values of the "A" stimuli. Consistent with the results of French-St. George and Bregman (1989), this manipulation had no effect on perceptual judgments of the test sequence (Rogers and Bregman, 1993). Yet again, a predictable induction sequence rendered not only the "A" tones of the test sequence, but also the whole "ABA_" pattern of the test sequence more predictable. Therefore, predictability may have supported both the "Integrated" and "Segregated" perceptual organizations, and the two effects may have canceled each other out. Consequently, empirical evidence in favor of or against sound predictability effects in ASA is inconclusive based on these early studies.

DE-CONFOUNDING PREDICTABILITY EFFECTS ON STREAM SEGREGATION AND INTEGRATION

Some further studies investigating the role of predictability within ASA emerged recently, motivated by the increased interest in auditory predictive processing (Bendixen et al., 2010, 2012b, 2013a, in revision; Devergie et al., 2010; Andreou et al., 2011; Rimmele et al., 2012; Rajendran et al., 2013). In these studies, care was taken not to introduce a confounding effect of influencing stream segregation and integration in parallel. To avoid this confound, one needs to acknowledge that it is difficult to manipulate a tone sequence such that only one of the perceptual organizations (but not the other) would change in terms of *formal* (mathematical) predictability. Yet it is much easier to achieve such a manipulation when considering *perceptual* rather than formal predictability. In some cases, formally predictable tone sequences are treated as if they were unpredictable by the auditory system because the predictable pattern contains too many elements or spans too much time, thereby exceeding memory limitations (e.g., Scherg et al., 1989; Sussman et al., 1999; Sussman and Gumenyuk, 2005; Boh et al., 2011). This knowledge can be exploited for introducing manipulations that render only the "Integrated" or only the "Segregated" perceptual organization predictable.

An example of such an independent manipulation is depicted in **Figure 2**. The predictability manipulation is based on changing the sequential linkage between successive tones within each set ("A-A-A" and "B-B-B") or across the sets ("A-B" and "B-A"), such that these sounds are predictive of each other in some conditions but not in other conditions. Sequential links within

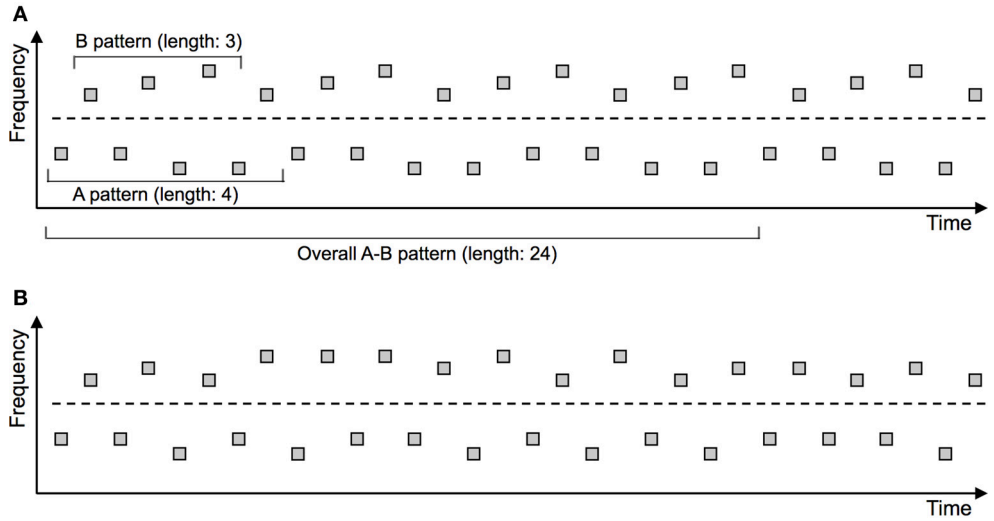


FIGURE 2 | Experimental paradigm disentangling predictability of the "Integrated" and "Segregated" perceptual organizations. The depicted stimulus arrangements are predictable **(A)** or unpredictable **(B)** with respect to stimulus frequency. The cyclically repeating (thereby predictable) frequency patterns are marked in the upper panel. The dashed line indicates the (nominal) separation between the "A" and "B" groups of tones. Critically, the number of elements included in the predictable patterns differs between the "A" and "B" group of tones. As a result, the length of the cyclically repeating overall pattern comprising "A" and "B" tones amounts to 24 elements, which is considerably too long to be picked up by the auditory system (e.g., Boh et al., 2011). Consequently, from a perceptual point of view the predictability manipulation is *directional*: It affects only predictability of the "Segregated" perceptual organization, whereas predictability of the "Integrated" organization remains unchanged. This directional manipulation allows for an unambiguous investigation of predictability effects in ASA (e.g., Bendixen et al., 2010, 2013a).

each set affect the predictability of the "Segregated" organization, while links across sets affect the predictability of the "Integrated" organization. Achieving one without the other leads to a *directional* predictability manipulation, avoiding the above-mentioned confound.

As depicted on **Figure 2**, one can include a regular pattern of a particular length (e.g., 4 elements) into the "A" sub-sequence, and a separate pattern of a different length (e.g., 3 elements) into the "B" sub-sequence (e.g., "B1B2B3B1B2B3…," with 1/2/3 representing slightly different frequency values within the "B" stream). Compared to a random arrangement of frequencies (e.g., "B1B1B3B1B2B3B2…"), this enhances the predictability of the "B" sub-sequence (and, by the same principle but independently, of the "A" sub-sequence). Hence the predictability of the "Segregated" perceptual organization increases. From a formal-mathematical point of view, predictability of the whole "AB" sequence (and hence of the "Integrated" organization) also increases, but the resulting pattern spans 24 elements, which is considerably too long to be picked up by the auditory system (e.g., Boh et al., 2011). Therefore, from a perceptual point of view, predictability of the "Integrated" perceptual organization remains unchanged. This type of manipulation thus selectively increases predictability of the "Segregated" perceptual organization, permitting a clear inference as to the role of predictability within ASA (Bendixen et al., 2010, 2013a).

Selectively manipulating predictability of the "Segregated" perceptual organization can also be achieved by changing one of the sub-sequences (only "A" or only "B") from random to predictable, while leaving the other sub-sequence random (Devergie et al., 2010; Andreou et al., 2011; Rimmele et al., 2012). In this case, predictability of the "Integrated" organization again remains unchanged, while the "Segregated" organization becomes partially predictable.

A closely related possibility is to start with two predictable sub-sequences and to then change one of the sub-sequences (only "A" or only "B") from predictable to random, while leaving the other sub-sequence predictable (Bendixen et al., 2012b; Rajendran et al., 2013). Provided that the regular patterns of the two predictable sub-sequences had been distinct from each other to preclude predictability of the whole sequence ("Integrated" organization), this would again lead to a selective manipulation of predictability of the "Segregated" organization (Bendixen et al., 2012b). If, on the other hand, the regular patterns of the two predictable sub-sequences had rendered the whole sequence predictable (e.g., because they were identical in length, forming an easily detectable overall pattern), changing one of the sub-sequences would decrease predictability of the "Integrated" and "Segregated" organizations in parallel, making the results of the predictability manipulation more difficult to interpret (Rajendran et al., 2013). This problem pertains to any design where the "predictable" condition is chosen to have no feature variation in the "A" and "B" sequences (as in the classical "ABA_" paradigm; Van Noorden, 1975). Constancy is of course the easiest form of predictability, but it is also the one that is most difficult for disentangling "Segregated" and "Integrated" predictability.

In summary, by exploiting knowledge on the pattern lengths that the auditory system recognizes as predictable, it is feasible to selectively manipulate predictability of the "Segregated" perceptual organization. The opposite case, manipulating predictability of the "Integrated" but not "Segregated" perceptual organization, is more difficult to achieve. Consider that the whole "ABA_" sequence, corresponding to the "Integrated" organization, is

designed to be predictable. In this case, the feature values of the first "A" tone predict those of the "B" tone, while the feature values of "B" in turn predict those of the second "A" tone. In order to avoid predictability of the "Segregated" organization, one would have to ensure that the feature values of the first "A" tone are not predictive of those of the second "A" tone. This is difficult, if not impossible to achieve when the predictability manipulation is based on only one feature. In a recent study, we proposed a compromise solution (Bendixen et al., in revision) where the "Integrated" organization is at least partially predictable, while the "Segregated" organization remains almost entirely unpredictable.

The various possibilities of manipulating predictability of the "Segregated" or "Integrated" organization (sometimes unwantedly affecting both in parallel) are summarized in **Figure 3**, alongside with studies that exemplify the resulting comparisons. Six different predictability conditions are distinguished on **Figure 3**. In the two conditions to the left, predictability of the "Integrated" organization is higher than predictability of the "Segregated" organization. The converse is true for the two conditions to the right. The two conditions in the middle represent cases where predictability of the "Integrated" and "Segregated" organizations do not differ (they are either both perfectly predictable or both unpredictable). As a consequence, empirical investigations comparing these two middle conditions must be regarded as uninformative with respect to the role of predictability as a cue in ASA. The fact that such studies (French-St. George and Bregman, 1989; Rogers and Bregman, 1993; Denham et al., 2010; Carl and Gutschalk, 2013) revealed no clear effects of the predictability manipulation can be attributed to the above-mentioned confound. It should be noted, though, that some of these studies employed the predictability manipulation with a different aim than the one in focus here, and that their results are nevertheless informative for models of ASA in a wider sense (see below).

Results of the studies changing predictability in a directional manner will be examined in the following. If the proposition that predictability contributes to the organization of the auditory scene is correct (Denham and Winkler, 2006; Winkler, 2007; Winkler et al., 2009, 2012), the empirical results should show an increase in the likelihood of stream segregation for those conditions where the predictability of the "Segregated" organization is selectively increased. In turn, the likelihood of stream integration should increase for those conditions where the predictability of the "Integrated" organization is selectively increased. Before reviewing the empirical data, the methods typically employed for studying ASA will be briefly summarized to point at some methodological advances that have contributed to the findings under review.

METHODOLOGICAL ASPECTS OF MEASURING AUDITORY SCENE ANALYSIS

Two main approaches are typically pursued to study ASA (Bregman, 1990; Micheyl and Oxenham, 2010a; Spielmann et al., 2013). Participants can be asked explicitly how they perceived a given auditory scene; in particular, whether they heard one or two sound sources at each particular moment (*subjective-report procedure*). This procedure has the advantage of providing a direct measurement of perceptual organization. However, it comes with the disadvantage of any introspective method: the limited possibility to validate whether participants are indicating their perception in a veridical manner. It is therefore beneficial to complement the subjective-report data with a second approach (*objective-listening tests*): Listening tasks can be set up that are easier to accomplish if participants are perceiving the auditory scene in one way (e.g., segregated into two sound sources) rather than another (e.g., integrated into one sound source). Poor task performance is then indicative of the listener not being able to segregate (or, in tasks applying the opposite logic, to integrate) the streams despite trying to do so volitionally. These tests have the advantage of permitting objective measurements, yet the disadvantage of being but indirect measures of the actual percept. It is, for instance, possible that participants succeed in stream segregation but still fail to perform the task on the foreground stream because they are too heavily distracted by the presence of the background stream. Perhaps more often neglected, the opposite is also possible: Participants might succeed in performing the task although they do not succeed in segregating the streams. This happens if participants find a way of solving the task without perceptually organizing the auditory scene in the way the experiment was set up to evoke. For instance, it has been reported that participants can adopt the strategy of listening out for every other tone in a mixture to circumvent the need of stream segregation (Dowling et al., 1987). In both cases, task performance would become an invalid indicator of perceptual organization. The two approaches offered by objective-listening tests and subjective-report procedures can therefore be seen as complementary in the type of evidence they provide as well as in the type of alternative explanations they can rule out.

Recent advances regarding the objective-listening tests have been brought about by improved analysis methods for assessing task performance under fast presentation conditions (Bendixen and Andersen, 2013). Moreover, it was demonstrated that objective-listening performance (and hence, success in stream segregation or integration) can be measured not only behaviorally but also in terms of ERPs, with close correspondence between these measures (Sussman et al., 2001; Winkler et al., 2003b). This reduces the reliance on giving participants a behavioral task, and thereby considerably widens the scope of studying ASA (e.g., toward ASA processes outside the focus of attention, cf. Sussman et al., 2007; or toward participant groups with difficulties in giving behavioral responses, cf. Winkler et al., 2003a). It also permits parallel acquisition of objective indicators of stream segregation and integration (Spielmann et al., in press) to be able to cross-validate these indirect measures of perception.

New methodological insights have also been gained throughout the past years in terms of the subjective-report procedure. It has been demonstrated that the perception of one vs. two streams is not fully determined by the temporal and feature characteristics of the stimulus sequence. Instead, stream perception underlies inter-individual and, more strikingly, intra-individual fluctuations. When presenting cyclically repeating "ABA_" sequences and asking participants to report their current percept in a

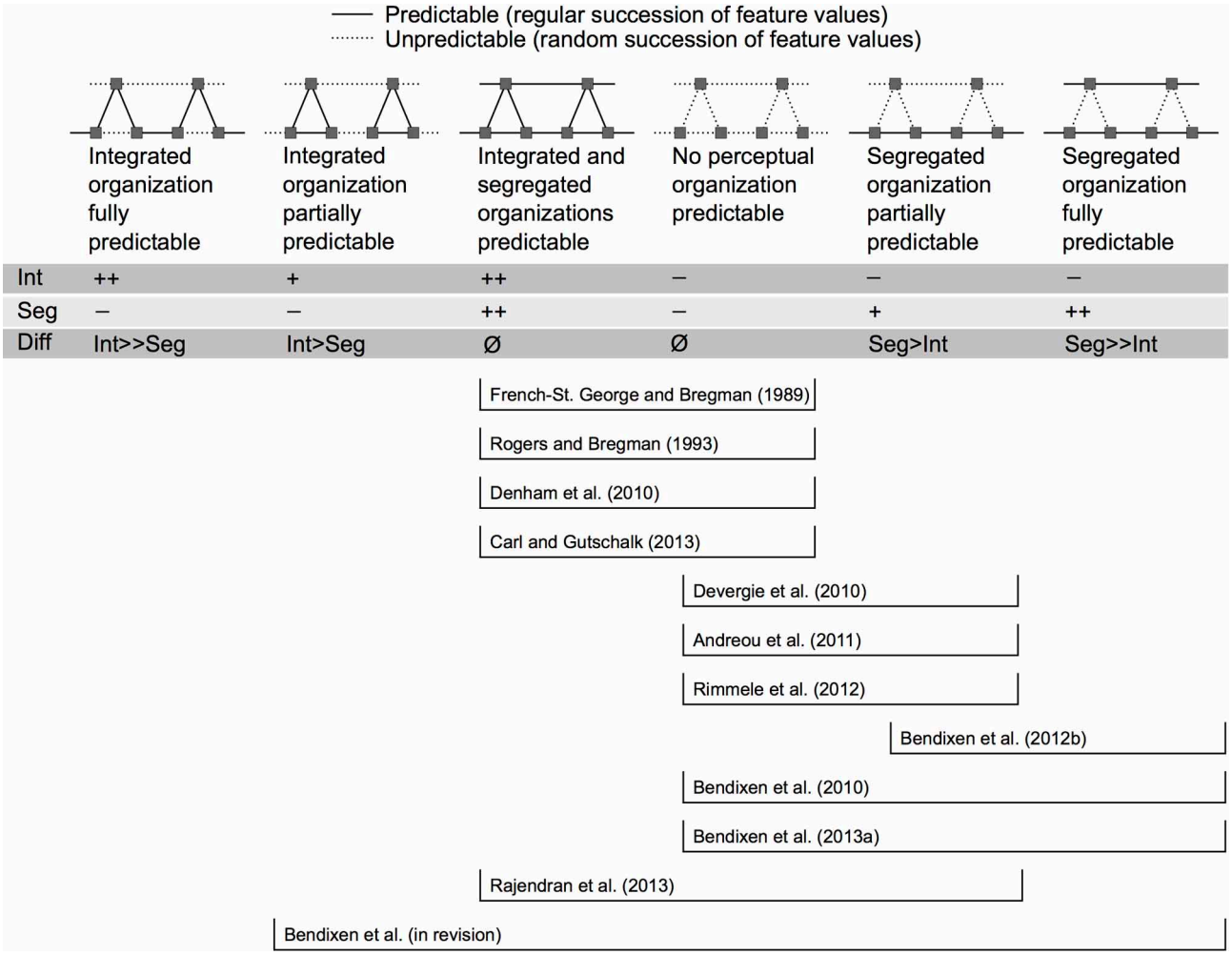


FIGURE 3 | Designs for studying predictability effects in auditory scene analysis. Six different levels of predictability are distinguished; and it is indicated for each of the previous studies which levels they have contrasted. Each level is schematically illustrated with a cutout of the corresponding stimulus sequence. Time is represented on the X axis and frequency on the Y axis of all panels. The schematic depiction uses the "ABA_" paradigm (Van Noorden, 1975), but studies have also used the "ABAB" paradigm or more irregular arrangements of "A" and "B" tones. Straight lines indicate the presence of predictive relations between successive tones (i.e., the feature values of one tone are predictive of the feature values of the next tone in one or both perceptual organizations). Dotted lines indicate random successions of feature values. The feature whose predictability was manipulated differs between studies (e.g., onset time, frequency, intensity, location). The effects on predictability of the "Integrated" (Int) and "Segregated" (Seg) organizations are marked with "++" (fully predictable), "+" (partially predictable), or "−" (unpredictable). The resulting predictability difference between the two organizations is marked in the "Diff" row. Following these differences, predictability conditions to the left should increase the likelihood of "Integrated" percepts, whereas predictability conditions to the right should increase the likelihood of "Segregated" percepts. Note that many studies have compared conditions in the middle of this scheme, and have revealed no clear effects of predictability on auditory perceptual organization. Studies employing directional manipulations have tended to investigate conditions where the "Segregated" organization was more predictable than the "Integrated" organization (depicted on the right of this scheme). No study has so far investigated a condition in which the "Integrated" but not the "Segregated" organization was fully predictable.

continuous manner, perception switches back and forth between the "Integrated" and "Segregated" alternatives (and possibly more, cf. Denham et al., 2014). Although initial attempts with this method (Anstis and Saida, 1985) have reported that perception settles on the "Segregated" interpretation rather than switching between the alternatives, more recent findings converge in showing that stochastic perceptual switching does occur with prolonged exposure times (e.g., Gutschalk et al., 2005; Winkler et al., 2005; Kondo and Kashino, 2009; Hill et al., 2011). Such changes in perception despite unchanged stimulus configuration constitute a case of bi- or multi-stability in audition that is analogous to bi-stable phenomena in vision (Pressnitzer and Hupé, 2006; Hupé and Pressnitzer, 2012; Kondo et al., 2012). Fluctuations between alternative percepts occur for a wide range of the acoustical parameters (Denham et al., 2013). They probably reflect the fact that the auditory system explores different alternatives of grouping the sounds (Denham and Winkler, 2006; Winkler et al., 2009, 2012; Denham et al., 2014).

The bi-stable nature of perception in the auditory streaming paradigm is becoming a popular tool in ASA research (cf. Pressnitzer et al., 2011). It can, for instance, be exploited for a characterization of the impact of different cues affecting ASA. By analyzing the time-course of perceptual switching, it is possible to separate cues that initiate perceptual switches from cues that stabilize a given percept but do not cause switching toward it (e.g., Bendixen et al., 2010, 2013a,b; Szalárdy et al., 2013, in press). The latter cues prolong the duration of experiencing a given percept, while the former cues additionally shorten the duration of experiencing the alternative percept(s). Both types of cue effects lead to an increase in the overall proportion of experiencing the associated percept (see **Figure 4** for illustration). The differentiation of the underlying mechanism speaks to the stage of ASA at which the cues exert their influence. Cues that can be shown to initiate switches must exert their influence at an early (partitioning) stage of ASA, at which the decomposition is initially decided upon. In contrast, for cues that merely stabilize a given percept but cannot initiate switches toward it, it can be inferred that they are not able to affect this early stage but are limited to providing feedback to a perceptual organization emerging spontaneously or by means of other cues.

RECENT STUDIES ON PREDICTABILITY EFFECTS IN AUDITORY SCENE ANALYSIS

With these new methodological prospects, the issue of predictability as a cue in ASA was revisited by some recent studies. In the majority of studies (Bendixen et al., 2010, 2013a; Devergie et al., 2010; Andreou et al., 2011; Rimmele et al., 2012), predictability of the "Segregated" organization was selectively manipulated. A selective increase in "Segregated" predictability was tested as an additional cue toward stream segregation, while a primary cue for stream segregation (in particular, spectral separation) was also present. The precise manipulations as well as the employed control conditions differed between studies, but when sorting them into the scheme developed in **Figure 3**, it becomes obvious that all of them contrasted one condition where the "Segregated" organization was more predictable than the "Integrated" organization with another condition where the two organizations were equally (un)predictable (note that this is also true for the study of Rajendran et al., 2013, though interpreted in a different framework by the authors). Notably, all the aforementioned studies found a higher proportion of stream segregation in those conditions where the predictability of the "Segregated" organization was selectively enhanced. These consistent results strongly support the view that predictability is used as a cue in ASA.

The result that the predictability of the "Segregated" organization enhances stream segregation was obtained both with subjective-report procedures (Bendixen et al., 2010, 2013a; Rajendran et al., 2013) and with objective-listening tasks (Devergie et al., 2010; Andreou et al., 2011; Rimmele et al., 2012). Hence predictability effects occur with neutral listening instructions (i.e., they bias perception of ambiguous auditory scenes toward configurations with higher predictability) as well as with active attempts to segregate the streams for solving a given task. Moreover, in the objective-listening tasks, predictability supported stream segregation both when it was embedded in the stream that the listener was instructed to focus upon (Rimmele et al., 2012; cf. earlier work by Jones et al., 1981) and when it was

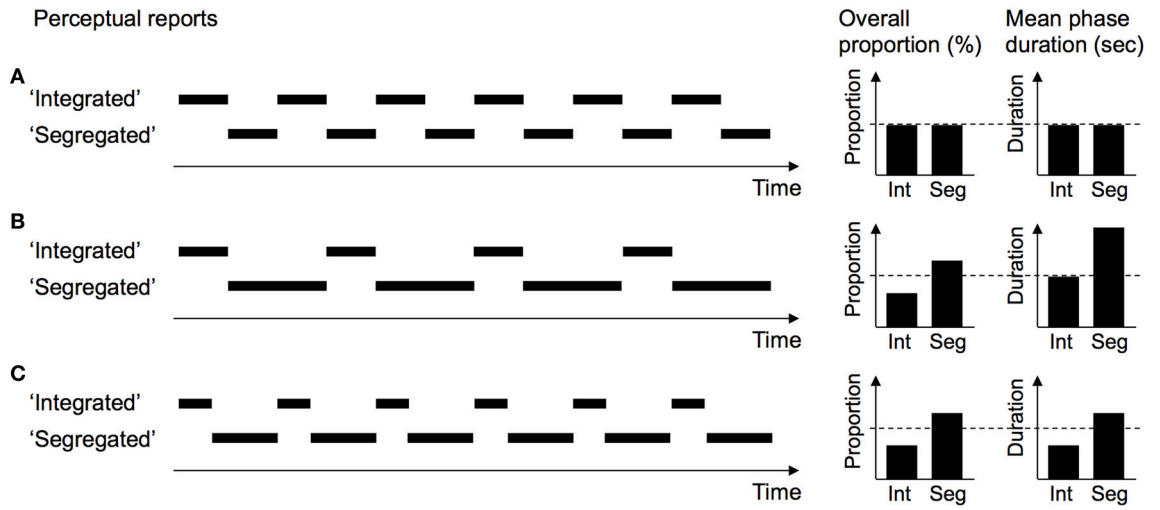


FIGURE 4 | Possible cue effects in the subjective-report procedure. Artificially simplified time-courses of perceptual switching were generated for illustration purposes. The upper row **(A)** reflects a balanced distribution of "Integrated" and "Segregated" percepts. The middle row **(B)** shows the impact of a percept-stabilizing cue in favor of stream segregation, which prolongs the duration (i.e., stability) of "Segregated" percepts but leaves the duration of "Integrated" percepts unaffected. The lower row **(C)** shows the impact of a percept-inducing cue in favor of segregation, which prolongs the duration of "Segregated" percepts and additionally shortens the duration of "Integrated" percepts by causing perceptual switches back to the "Segregated" percept. The dashed lines in each panel mark the proportion and duration values from the balanced condition for comparison. Note that in this example, percept-stabilizing and percept-inducing cues have identical effects on the proportions of the two percepts, hence analyzing only the average proportions cannot differentiate between these qualitatively different types of cues. The average phase durations are informative about the cause of the changes in proportion, and thus about the underlying mechanism of the cue.

embedded in another stream that was to be ignored by the listener (Devergie et al., 2010; Andreou et al., 2011; Rimmele et al., 2012). In other words, predictability of one sound source in a mixture helps both to voluntarily attend and to ignore this source. Predictability is thus a *symmetric* cue, which contrasts Bregman's (1990, p. 669) definition of higher-level cues in ASA.

A symmetry test at a different level concerns the question as to whether predictability can support not only stream segregation but also integration. One recent study (Bendixen et al., in revision) provided the so far missing evidence that selectively enhancing predictability of the "Integrated" organization leads to an increase in perceptual reports of stream integration—the counterpart of what was demonstrated for stream segregation many times before. The data of Rajendran et al. (2013) can also be interpreted in this way, although segregated and integrated predictability were not strictly disentangled in this study. Hence predictability can now be regarded as a *symmetric* cue on multiple levels, in that it flexibly supports either stream segregation or integration, and in that it is effective both when it is present in an attended foreground sound source and when it is present in an unattended background sound source.

The latter finding is qualified by the results of Rimmele et al. (2012) who demonstrated that the facilitation of ASA by predictability in a background sound source shows age-related decline. These authors applied an objective-listening task requiring stream segregation. In a between-subject design, older listeners (mean age of 67 years) were shown to benefit less from background predictability than younger listeners (mean age of 22 years). In contrast, the effect of foreground predictability was similar in both age groups. The authors explain their finding by the fact that older listeners are known to exhibit deficits in predictability extraction from sequences outside the focus of attention (cf. reviews by Pekkonen, 2000; Näätänen et al., 2012). It should, however, be conceded that age-related impairments in ASA were observed for only one out of two types of background predictability, with the reason for this difference between predictability conditions remaining unclear. Moreover, Rimmele et al. (2012) did not record measures of predictability extraction in their participants. Hence although the interpretation that impairments in predictability extraction translate into corresponding impairments in predictability-based stream segregation is suggestive, this link remains to be directly examined. If the assumed relation between predictability extraction capacities and the ability to exploit predictability for ASA could be demonstrated at a single-subject level, this would have strong implications for models of impaired ASA (e.g., Beutelmann et al., 2010).

Whether there is or is not a direct link with predictability extraction, one can conclude from the data of Rimmele et al. (2012) that there are age-related difficulties in predictability-related aspects of ASA when predictability needs to be processed outside the focus of attention. This contrasts previous observations of sequential ASA being preserved in older participants (Trainor and Trehub, 1989; Snyder and Alain, 2007a), which have led to the assumption that only concurrent ASA declines with age (Snyder and Alain, 2005; Alain and McDonald, 2007). The data of Rimmele et al. (2012; see also Hutka et al., 2012) might thus help in filling the gap between the troublesome hearing impairments experienced by older listeners (e.g., Schneider et al., 2002; Shinn-Cunningham and Best, 2008) and the surprising failure to replicate such difficulties in the laboratory. If future studies confirm the notion that some aspects of sequential ASA do show an age-related decline, this insight has the potential to inform hearing aid design or to develop training programs for counteracting such deficits.

UNDERSTANDING PREDICTABILITY EFFECTS: PERCEPTION OR ATTENTION?

Given that effects of predictability on ASA have now been consistently demonstrated across studies and laboratories, a next important step is to understand *how* predictability affects ASA. One imminent question pertains to whether all the above-reviewed studies indeed measured predictability effects on the perception of the tone sequences (i.e., on the processes of stream segregation and integration), or whether other aspects of auditory processing might have been influenced as well. For the objective-listening tasks, it is possible that predictability facilitated processes of attention rather than sound organization. It has already been argued by Bregman (1990) that predictable sequences are easier to hold in the focus of attention (cf. Jones and Boltz, 1989). Somewhat more cumbersome but still possible, one may also posit that a predictable sequence in the background is easier to shield from attentive processing in order to spare resources for foreground listening (Devergie et al., 2010). How the brain uses predictability for directing attention toward or away from incoming stimuli is an area of active exploration (for a recent review, see Henry and Herrmann, 2014). When considering objective-listening tests alone, one could thus argue that predictability benefits in task performance stem from processes that are unrelated to sound organization.

However, when considering the objective-listening together with the subjective-report data, this interpretation becomes less likely: The same predictability manipulation was shown to increase subjective perceptual reports of "Segregation" (without specific instruction in terms of attention) and to improve performance in a task designed to require a "Segregated" percept. This makes the interpretation that the predictability manipulation exerted its influence through acting upon auditory perceptual organization much more plausible. As pointed out above, the two approaches complement each other in the type of inference they allow.

Further evidence for the view that predictability effects on stream segregation comprise more than attentional allocation comes from a recent MMN study (Bendixen et al., 2012b). This study showed that a series of interleaved tone sequences could be disentangled solely on the basis of predictability when the sequences were presented outside the focus of attention. In contrast, listeners mostly failed to segregate the streams when attentively trying to do so. This failure during active listening makes it highly unlikely that attention could have contributed to the predictability effect during passive listening. Hence the "pre-attentive" (Sussman, 2007) auditory system appears to be equipped with a bottom-up mechanism disentangling a mixture of two sound streams solely based on the predictability of these streams.

Taken together, these results suggest that predictability is effortlessly taken into account for auditory perceptual organization and does not require attentive processing of the predictable sound source, nor of the predictability itself. This conclusion is consistent with the view of predictive processing as the "default mode" of the brain (Friston, 2005, 2010), and it exceeds Bregman's (1990) framework in which the role of predictability was confined to an attention-mediated stage of ASA.

The consistent results from objective and subjective listening procedures nevertheless do not imply that predictability effects measured in objective-listening tests are *entirely* attributable to processes of sound organization rather than attention. It remains possible that both effects contribute (cf. Lange, 2013), and this possibility should be carefully considered for each objective-listening study. This seems particularly important for the potentially translational perspective of understanding predictability-related deficits in older listeners (Rimmele et al., 2012). Difficulties of older listeners in the allocation of auditory attention are well described (e.g., Mager et al., 2005; Horváth et al., 2009; Lawo and Koch, in press), and hence they must be examined as a viable alternative explanation.

UNDERSTANDING PREDICTABILITY EFFECTS: EARLY OR LATE?

When asking *how* predictability affects sound organization processes in ASA, another imminent question pertains to whether it acts as an early grouping cue (affecting the initial decomposition of the auditory input) or as a late grouping cue (confirming or disconfirming the decomposition provided by the early cues). As outlined above, the studies employing subjective-report procedures can shed some further light on this. Analyzing the switching characteristics between the perceptual reports of "Integration" and "Segregation" offers a straightforward way of determining whether a given cue induces or merely stabilizes the associated perceptual organization.

Initial results with this method (Bendixen et al., 2010) showed that a directional predictability cue stabilized the "Segregated" perceptual organization (i.e., it prolonged the duration of "Segregated" percepts), but it did not cause switching toward a "Segregated" perceptual organization (i.e., it did not shorten the duration of "Integrated" reports due to perception switching back to "Segregation"). Within a two-stage framework of auditory scene analysis (Bregman, 1990), these results suggest that predictability does not act upon the first stage during which the auditory input is decomposed, but only upon the second stage during which the sound groups are evaluated. The effect of predictability could be conceptualized as giving support to the configurations provided by the first stage depending on how successfully they predict newly incoming sounds. The decomposition itself would be dominated by primary acoustic features (such as spectral separation) exerting their influence in the first stage.

This interpretation received further support from a study with an orthogonal manipulation of predictability and spectral separation (Bendixen et al., 2013a). This study showed qualitatively different effects of manipulating these two types of cues, in that predictability again had stabilizing effects on "Segregated" perceptual reports, whereas spectral separation not only stabilized but also induced stream segregation. Additional support for a dissociation between the cues was provided by the absence of statistical interactions between the effects of spectral separation and predictability. Furthermore, the two types of cues responded differently to changes introduced during the sequence: Changing spectral separation was associated with a contrastive carry-over effect (similar to Snyder et al., 2009a,b), while changing predictability led to no carry-over effect. These findings were taken to suggest that predictability exerts its influence upon ASA only after primary acoustic grouping cues have been considered (Bendixen et al., 2010, 2013a,b). This implies that the impact of predictability is not as large as theoretically possible when deriving hypotheses from a strict predictive-processing point of view (Friston, 2005, 2010). In particular, there is no compelling support for the idea that predictability could be the earliest grouping cue (based on its availability with stimulus onset) or that it acts in parallel with the acoustic grouping cues.

This initially clear-cut inference must be viewed with more caution now that more evidence regarding predictability effects in ASA has been gathered. First, there are now demonstrations of interactions between spectral separation and predictability (Andreou et al., 2011). Second, there is ERP (though no behavioral) evidence that stream-specific predictable patterns can induce stream segregation when spectral separation and other primary feature differences between the streams are removed (Bendixen et al., 2012b). Third, switching characteristics for predictability manipulations in a more recent study (Bendixen et al., in revision) showed not only inducing but also stabilizing properties (see also Szalárdy et al., in press, for a predictability-like manipulation based on musical structure). Hence, although the question of *how* predictability acts upon ASA received a seemingly simple answer initially, it appears more appropriate to view this as an issue for further investigation.

A possible explanation for the discrepant results comes from another recent line of studies on the so-called *hierarchical novelty detection* system (cf. reviews by Grimm and Escera, 2012; Escera et al., in press). These authors show that violations of simple-repetition types of predictability are detected by the auditory system considerably earlier (in the middle-latency range) than violations of complex-pattern types of predictability. Complex predictability (such as a regular alternation between two feature values: "121212…") does not affect processing until the MMN latency range (cf. Cornella et al., 2012; Althen et al., 2013). This implies that predictability information in terms of stimulus repetition is available to the auditory system as early as 20 ms post-stimulus while information in terms of more complex patterns is available only later in auditory processing, at around 150 ms post-stimulus. In other words, although the information about the next (predictable) sound is in theory available before stimulus onset once the predictability is extracted, the auditory system does not actually "pass down" this information from cortical structures to earlier processing stages. If this is the case, then it is not surprising to see that such complex forms of predictability cannot act upon ASA at early time-points. Together with the fact that stream formation based on strong acoustic cues (e.g., large spectral separation) occurs within less than 100 ms (Müller et al., 2005) and probably starts in peripheral structures of the auditory

pathway (Pressnitzer et al., 2008), complex forms of predictability would be bound to exert their influence only after the acoustic cues have been taken into account.

In that respect, it is worthy to note that those studies showing that predictability acts upon ASA *after* the primary features (Bendixen et al., 2010, 2013a) employed complex-pattern types of predictability. In contrast, those studies whose results suggest that predictability acts at least in parallel with the primary features (Andreou et al., 2011; Bendixen et al., 2012b; in revision) have employed simple forms of predictability based on repetition or gradual progression of feature values. It is thus possible that simple types of predictability show early effects in ASA whereas complex forms of predictability are indeed limited to acting at a later stage. The interpretation that some forms of predictability affect auditory processing earlier than other forms is consistent with the view that stream formation can be triggered on various levels of the auditory processing hierarchy depending on the types of available cues (Cusack et al., 2004; Griffiths and Warren, 2004; Snyder and Alain, 2007b; Pressnitzer et al., 2008). Yet a direct comparison of the effects of predictability on different levels of complexity remains to be performed in future studies.

PUTTING THE ROLE OF PREDICTABILITY INTO PERSPECTIVE

Although the precise mechanisms are not yet specified, the studies reviewed above clearly demonstrate that predictability affects ASA. These findings lend support to theories linking predictive processing and ASA (Denham and Winkler, 2006; Winkler, 2007; Winkler et al., 2009), and they have encouraged the formulation of further conceptual and computational models of ASA based on predictive-processing principles (Winkler and Czigler, 2012; Winkler et al., 2012; Mill et al., 2013; Schröger et al., in press). Such efforts should not be mistaken to suggest that strict predictability is necessary for stream formation to occur. It has been shown several times that with unpredictable tone arrangements, stream formation simply operates on the basis of other cues (French-St. George and Bregman, 1989; Müller et al., 2005; Denham et al., 2010; Carl and Gutschalk, 2013). In other words, auditory perception readily tolerates random variation in the signal emissions of putative sound sources.

Sound organizations with predictable behavior of the resulting sound sources are sizably but not exclusively preferred over random arrangements. This is evidenced by the sometimes modest effect sizes of predictability manipulations (e.g., ∼10% change in perceptual reports in Bendixen et al., 2010, 2013a). In order to put such numbers into perspective, it is important to acknowledge that the labels "predictable" and "unpredictable" exaggerate the differences between conditions in many studies. This is because tones in the "unpredictable" arrangement are usually still predictable in many properties. Typically, only one or two tone parameters are manipulated, and the lack of predictability is caused by random choice between a handful of feature values. Hence it may be more appropriate to see the above-reviewed studies as comparing predictability in a strict sense with predictability in a wider sense. With this in mind, it becomes more remarkable that such robust effects of predictability were achieved.

The predictability effects are consistent with the results of MMN studies showing that the auditory system prefers deterministic over stochastic sequential structures (Winkler et al., 1990; Schröger and Roeber, 2011; Daikhin and Ahissar, 2012). Knowing that one out of two possible feature values will come next in a sequence seems to be less beneficial than knowing exactly which one. In other words, ASA differentiates between the exact continuation and the inexact but plausible continuation of a stream. This provides another level of distinction between the role of predictability and earlier effects described for the Gestalt principles of *similarity* and *continuity* (Bregman, 1990). It also provides another case where predictability effects in ASA follow the same principles as auditory predictive processing *per se*, illustrating the prospects of combining these research lines (Denham and Winkler, 2006; Winkler et al., 2009, 2012; Schröger et al., in press).

PREDICTABILITY AS A SEQUENTIAL ANALOGUE TO THE SIMULTANEOUS OLD-PLUS-NEW HEURISTIC?

Starting with the notion of a sequential analogue to the simultaneous *old-plus-new heuristic*, this review explored the idea that the auditory system's capacity to detect predictability in a sequence of discrete sounds could be used to support sequential ASA. Such supportive effects are clearly apparent in the reviewed studies. Sequential ASA shows a bias for perceptual interpretations comprising "old" (previously identified, well-describable and thus predictable) behavior of sound sources over interpretations including "new" signal emissions (i.e., unpredictable, surprising signals that are not yet characterized or not possible to characterize within the limitations of auditory predictability extraction). A further effect not explored in the present review but corroborating the gist of "old-plus-new" in sequential ASA is the auditory system's tendency to maintain a current perceptual interpretation despite parameter changes (Sussman and Steinschneider, 2006; Rahne and Sussman, 2009; Snyder et al., 2009b).

Comparing the properties of predictability-related sequential ASA with the old-plus-new heuristic in simultaneous ASA, similarities and differences become apparent. Unlike assumed by Bregman (1990), sequential grouping distinguishes plausible continuation of a source's signal emissions (as expressed in the *similarity* and *continuity* Gestalt principles) from precise (i.e., predictable) continuation. This is equivalent to the expectation of precise signal continuation observed for simultaneous grouping in the old-plus-new heuristic (Darwin, 1995). However, the preference for precise continuation in simultaneous grouping is so strong that the expected continuation of a sound source is metaphorically subtracted from a mixture before applying any other decomposition principles (Bregman, 1990). In other words, the old-plus-new heuristic is the first grouping principle in simultaneous ASA, while the same does not seem to hold for sequential predictability. Despite this difference, it is probably helpful for theorizing to retain the notion of an analogy between predictability-related sequential ASA and the previously expressed old-plus-new heuristic in simultaneous grouping.

Interestingly in this respect, McDermott et al. (2011) proposed an extension of the old-plus-new heuristic in simultaneous ASA in a similar vein as suggested for the sequential analogue in the present review. Specifically, these authors propose that simultaneous grouping is facilitated by the recurrence of sound sources even if every single occurrence is accompanied by

concurrently overlapping signals (so-called *embedded repetition*). This is supported by empirical evidence that repetition-driven extraction of sound sources is indeed possible (McDermott et al., 2011). The data of McDermott et al. (2011) relieve the constraint that the simultaneous old-plus-new heuristic only applies to uninterrupted continuation (Bregman, 1990). Some earlier findings can also be interpreted in this vein (Darwin et al., 1995; Hukin and Darwin, 1995; Shinn-Cunningham et al., 2007; Lee and Shinn-Cunningham, 2008): These studies showed that a few presentations of one element of a complex sound (i.e., non-embedded repetition) prior to the presentation of the whole sound can lead to segregation of the element from the complex (i.e., a case where sequential grouping overcomes the temporal coherence principle; Shamma et al., 2011). This effect has been called *sequential capture*, and has been discussed in the framework of (competitive) interactions between simultaneous and sequential grouping cues. Interpreting the prior findings as demonstrations of an extended old-plus-new effect on simultaneous grouping offers an alternative perspective that can also integrate the present findings: Interrupted but repeated occurrence of a sound source's signal emissions might be equally beneficial for sequential and for simultaneous ASA.

Although this might provide a compelling perspective of a ubiquitous old-plus-new heuristic, two important differences between the respective sequential and simultaneous phenomena should not be neglected. First, old-plus-new effects on simultaneous ASA benefit from random changes in the concurrently present sources (McDermott et al., 2011) because in this case, the repetitive elements of the "old" source can be extracted more easily. In contrast, the corresponding effect in sequential ASA benefits from both sources being fully predictable (Bendixen et al., 2012b), lending itself more readily to a predictive-coding-based explanation.

Second, old-plus-new effects on simultaneous ASA have mainly been demonstrated for simple repetition (see also Best et al., 2008; Dyson and Alain, 2008), and have been shown to break down rapidly with more complex forms of predictability (Jones and Litovsky, 2008; Best et al., 2010). In contrast, sequential old-plus-new effects occur with predictable patterns of sound events exceeding mere repetition, as reviewed above. For postulating a common underlying principle, old-plus-new effects on the segregation of simultaneous sound sources would need to be demonstrated with regularities on different levels of complexity. Altogether, a further exploration into analogies of the old-plus-new heuristic in simultaneous and sequential sound grouping is expected to yield further insights into both types of grouping.

OUTSTANDING ISSUES

Now that the general case for predictability effects in ASA appears to be made, it is time for a more detailed investigation into the involved mechanisms to gain a better understanding of their properties and limitations. Hence this review will end with a set of open questions and issues to be explored in future studies.

ROLE OF THE TYPE OF PREDICTABILITY

As outlined above, future studies should explicitly contrast the effects of simple and complex regularities and investigate whether this affects the time-range at which predictability exerts its influence. This should be done while bearing in mind that for simple-repetition regularities, it is all the more difficult to separately manipulate predictability of the "Segregated" and "Integrated" organizations. Moreover, simple-repetition regularities typically confound predictability with the amount of feature variation, since the unpredictable condition used for comparison is unavoidably more variable in the feature values (i.e., the unpredictable tones are more dissimilar to each other than the predictable ones). Notwithstanding these interpretational difficulties, studying predictability effects on different levels of complexity is needed to settle the above-mentioned controversy regarding the timescale of the predictability effects on sequential ASA.

ROLE OF THE FEATURES CARRYING THE REGULARITIES

Predictability was manipulated by means of different stimulus features in the reviewed studies, including frequency, intensity, location, timing, and some combinations of these. Future studies should attempt to distinguish between the effects of these different features, as they are regarded as qualitatively distinct in theoretical frameworks of auditory processing. Näätänen and Picton (1987) coined the terms *temporal uncertainty* and *event uncertainty* to emphasize the qualitative difference between temporal regularities (reflecting the *when* aspect) and feature regularities (reflecting the *what* aspect; see also recent work by Sperduti et al., 2011; Arnal and Giraud, 2012; Hughes et al., 2013; Schwartze et al., 2013). Location, reflecting the *where* aspect, may constitute yet another qualitatively different feature (Rauschecker and Tian, 2000; Schwartz and Shinn-Cunningham, 2010). Thus, although the different features were initially treated as somewhat interchangeable to demonstrate the general principle of predictability effects in ASA, their specific effects should be disentangled in future studies (for an analogy in simultaneous ASA, see Kitterick et al., 2010).

MECHANISMS UNDERLYING PREDICTABILITY-RELATED ASA PROCESSES

The mechanisms supporting predictability effects in ASA are far from being understood. Further studies will be required to clarity the controversies alluded to above (perceptual vs. attentional mechanisms as well as early vs. late effects on perceptual organization). It would be desirable for these studies to receive further inspiration from the field of predictive processing, including the ever-growing understanding of how predictability is instantiated in neuronal terms (e.g., Besle et al., 2011; Arnal and Giraud, 2012; Wacongne et al., 2012), in order to generate more precise hypotheses as to how predictability might act within ASA.

RELATION OF PREDICTABILITY WITH OTHER GROUPING CUES

Obviously, predictability effects should not be treated in isolation but integrated into a comprehensive model of ASA. Integrating predictability into a general ASA framework will also require thorough investigation of how predictability effects relate to those of other grouping cues, both primary acoustic ones (e.g., spectral separation) and higher-level cues (e.g., familiarity with a sound source, cf. Johnsrude et al., 2013).

AGE EFFECTS

In view of the practical implications, it seems of utmost importance to understand the origin of the age-related decline in predictability-based ASA. A highly relevant question is how much this effect actually contributes to age-related deficits in ASA in challenging real-life listening situations. This immediately relates to the next and final outstanding issue.

TRANSFER TO NATURAL LISTENING CONDITIONS

It would be highly desirable to show a predictability benefit for ASA with more natural stimulus material; in other words, to demonstrate the ecological validity of the findings reviewed here. To what extent do listeners exploit natural predictability for forming a stable auditory stream of their conversation partner in a noisy environment? Natural auditory scenes obviously contain less strict forms of predictability, and the issue as to how much variability the brain tolerates while still treating a sound source's signal emissions as predictable is largely unresolved (but see Winkler et al., 1990; Daikhin and Ahissar, 2012). This issue must be addressed, however, if the present results shall be transferred to more applied domains. A perceptually inspired approach toward predictability in ASA might be highly informative for technical approaches of noise cancellation by predictive principles (e.g., Guldenschuh and Höldrich, 2013).

CONCLUSIONS

In summary, there is a growing body of evidence from studies with complementary methods demonstrating that the impact of sound predictability on ASA is considerably more extensive than previously assumed. First, perceptual grouping favors precise continuation of a sound source's signal emissions over inexact but plausible (i.e., within the range of the sound source's feature characteristics) continuation; predictability thereby differs from the Gestalt principles of *similarity* and *continuity*. Second, effects of predictability on ASA need not be mediated by attention, but are readily explained in a bottom-up framework of automatically extracting predictability and applying it for further processing. Third, effects of predictability on ASA are symmetric: Maintaining segregation of a foreground and background stream is facilitated by predictability in either stream. Fourth, predictability acts as a symmetric cue also with respect to supporting either stream segregation or integration. These effects can be observed only if *directional* manipulations are employed: Predictability of the "Segregated" and "Integrated" perceptual organizations must be disentangled in the experimental design to allow for unambiguous interpretations. The precise mechanisms for predictability effects in ASA as well as for possible age-related impairments of these effects remain to be determined. Deriving a joint theoretical framework for sequential predictability and the simultaneous old-plus-new heuristic that receives further inspiration from the field of auditory predictive processing is considered a promising avenue for future research.

ACKNOWLEDGMENTS

This article is partly based on a thesis submitted to the University of Leipzig, Germany, for fulfilling the requirements of habilitation [Bendixen (unpublished habilitation thesis). A new look at the old-plus-new heuristic: Sound predictability facilitates auditory scene analysis. Unpublished habilitation thesis, University of Leipzig, Germany.] The conversion into a review article was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG Cluster of Excellence 1077 "Hearing4all," and SFB/TRR 31 "The active auditory system").

REFERENCES

Alain, C., and McDonald, K. L. (2007). Age-related differences in neuromagnetic brain activity underlying concurrent sound perception. *J. Neurosci.* 27, 1308–1314. doi: 10.1523/JNEUROSCI.5433-06.2007

Althen, H., Grimm, S., and Escera, C. (2013). Simple and complex acoustic regularities are encoded at different levels of the auditory hierarchy. *Eur. J. Neurosci.* 38, 3448–3455. doi: 10.1111/ejn.12346

Andreou, L.-V., Kashino, M., and Chait, M. (2011). The role of temporal regularity in auditory segregation. *Hear. Res.* 280, 228–235. doi: 10.1016/j.heares.2011.06.001

Anstis, S., and Saida, S. (1985). Adaptation to auditory streaming of frequency-modulated tones. *J. Exp. Psychol. Hum. Percept. Perform.* 11, 257–271. doi: 10.1037/0096-1523.11.3.257

Arnal, L. H., and Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends Cogn. Sci.* 16, 390–398. doi: 10.1016/j.tics.2012.05.003

Baldeweg, T. (2006). Repetition effects to sounds: evidence for predictive coding in the auditory system. *Trends Cogn. Sci.* 10, 93–94. doi: 10.1016/j.tics.2006.01.010

Bendixen, A., and Andersen, S. K. (2013). Measuring target detection performance in paradigms with high event rates. *Clin. Neurophysiol.* 124, 928–940. doi: 10.1016/j.clinph.2012.11.012

Bendixen, A., Bőhm, T. M., Szalárdy, O., Mill, R., Denham, S. L., and Winkler, I. (2013a). Different roles of similarity and predictability in auditory stream segregation. *Learn. Percept.* 5, 37–54. doi: 10.1556/LP.5.2013.Suppl2.4

Bendixen, A., Denham, S. L., Gyimesi, K., and Winkler, I. (2010). Regular patterns stabilize auditory streams. *J. Acoust. Soc. Am.* 128, 3658–3666. doi: 10.1121/1.3500695

Bendixen, A., Denham, S. L., and Winkler, I. (2013b). "Sound predictability as a higher-order cue in auditory scene analysis," in *AIA-DAGA 2013*, *International Conference on Acoustics* (Berlin: DEGA), 705–708.

Bendixen, A., SanMiguel, I., and Schröger, E. (2012a). Early electrophysiological indicators for predictive processing in audition: a review. *Int. J. Psychophysiol.* 83, 120–131. doi: 10.1016/j.ijpsycho.2011.08.003

Bendixen, A., Schröger, E., Ritter, W., and Winkler, I. (2012b). Regularity extraction from non-adjacent sounds. *Front. Psychol.* 3:143. doi: 10.3389/fpsyg.2012.00143

Bendixen, A., Schröger, E., and Winkler, I. (2009). I heard that coming: event-related potential evidence for stimulus-driven prediction in the auditory system. *J. Neurosci.* 29, 8447–8451. doi: 10.1523/Jneurosci.1493-09.2009

Besle, J., Schevon, C. A., Mehta, A. D., Lakatos, P., Goodman, R. R., McKhann, G. M., et al. (2011). Tuning of the human neocortex to the temporal dynamics of attended events. *J. Neurosci.* 31, 3176–3185. doi: 10.1523/JNEUROSCI.4518-10.2011

Best, V., Ozmeral, E. J., Kopčo, N., and Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proc. Natl. Acad. Sci. U.S.A.* 105, 13174–13178. doi: 10.1073/pnas.0803718105

Best, V., Shinn-Cunningham, B. G., Ozmeral, E. J., and Kopčo, N. (2010). Exploring the benefit of auditory spatial continuity. *J. Acoust. Soc. Am.* 127, EL258–EL264. doi: 10.1121/1.3431093

Beutelmann, R., Brand, T., and Kollmeier, B. (2010). Revision, extension, and evaluation of a binaural speech intelligibility model. *J. Acoust. Soc. Am.* 127, 2479–2497. doi: 10.1121/1.3295575

Bey, C., and McAdams, S. (2002). Schema-based processing in auditory scene analysis. *Percept. Psychophys.* 64, 844–854. doi: 10.3758/BF03194750

Boh, B., Herholz, S. C., Lappe, C., and Pantev, C. (2011). Processing of complex auditory patterns in musicians and nonmusicians. *PLoS ONE* 6:e21458. doi: 10.1371/journal.pone.0021458

Bregman, A. S. (1990). *Auditory Scene Analysis. The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Carl, D., and Gutschalk, A. (2013). Role of pattern, regularity, and silent intervals in auditory stream segregation based on inter-aural time differences. *Exp. Brain Res.* 224, 557–570. doi: 10.1007/s00221-012-3333-z

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* 25, 975–979. doi: 10.1121/1.1907229

Cornella, M., Leung, S., Grimm, S., and Escera, C. (2012). Detection of simple and pattern regularity violations occurs at different levels of the auditory hierarchy. *PLoS ONE* 7:e43604. doi: 10.1371/journal.pone.0043604

Coy, A., and Barker, J. (2007). An automatic speech recognition system based on the scene analysis account of auditory perception. *Speech Commun.* 49, 384–401. doi: 10.1016/j.specom.2006.11.002

Cusack, R., Deeks, J., Aikman, G., and Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 643–656. doi: 10.1037/0096-1523.30.4.643

Daikhin, L., and Ahissar, M. (2012). Responses to deviants are modulated by subthreshold variability of the standard. *Psychophysiology* 49, 31–42. doi: 10.1111/j.1469-8986.2011.01274.x

Darwin, C. J. (1995). "Perceiving vowels in the presence of another sound: a quantitative test of the 'old-plus-new' heuristic," in *Levels in Speech Communication: Relations and Interactions: A Tribute to Max Wajskop*, eds C. Sorin, J. Mariani, H. Méloni and J. Schoentgen (Amsterdam: Elsevier), 1–12.

Darwin, C. J., Hukin, R. W., and Al-Khatib, B. Y. (1995). Grouping in pitch perception: Evidence for sequential constraints. *J. Acoust. Soc. Am.* 98, 880–885. doi: 10.1121/1.413513

Denham, S. L., Bőhm, T. M., Bendixen, A., Szalárdy, O., Kocsis, Z., Mill, R., et al. (2014). Stable individual characteristics in the perception of multiple embedded patterns in multistable auditory stimuli. *Front. Neurosci.* 8:25. doi: 10.3389/fnins.2014.00025

Denham, S. L., Gyimesi, K., Stefanics, G., and Winkler, I. (2010). "Stability of perceptual organisation in auditory streaming," in *The Neurophysiological Bases of Auditory Perception*, eds E. A. Lopez-Poveda, A. R. Palmer, and R. Meddis (New York, NY: Springer), 477–488. doi: 10.1007/978-1-4419-5686-6_44

Denham, S. L., Gyimesi, K., Stefanics, G., and Winkler, I. (2013). Perceptual bistability in auditory streaming: how much do stimulus features matter? *Learn. Percept.* 5, 73–100. doi: 10.1556/LP.5.2013.Suppl2.6

Denham, S. L., and Winkler, I. (2006). The role of predictive models in the formation of auditory streams. *J. Physiol. Paris* 100, 154–170. doi: 10.1016/j.jphysparis.2006.09.012

Devergie, A., Grimault, N., Tillmann, B., and Berthommier, F. (2010). Effect of rhythmic attention on the segregation of interleaved melodies. *J. Acoust. Soc. Am.* 128, EL1–EL7. doi: 10.1121/1.3436498

Dowling, W. J., Lung, K. M.-T., and Herrbold, S. (1987). Aiming attention in pitch and time in the perception of interleaved melodies. *Percept. Psychophys.* 41, 642–656. doi: 10.3758/BF03210496

Drake, C., Jones, M. R., and Baruch, C. (2000). The development of rhythmic attending in auditory sequences: attunement, referent period, focal attending. *Cognition* 77, 251–288. doi: 10.1016/S0010-0277(00)00106-2

Dyson, B. J., and Alain, C. (2008). It all sounds the same to me: Sequential ERP and behavioral effects during pitch and harmonicity judgments. *Cogn. Affect. Behav. Neurosci.* 8, 329–343. doi: 10.3758/CABN.8.3.329

Ellis, D. P. W. (1999). Using knowledge to organize sound: the prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures. *Speech Commun.* 27, 281–298. doi: 10.1016/S0167-6393(98)00083-1

Escera, C., Leung, S., and Grimm, S. (in press). Deviance detection based on regularity encoding along the auditory hierarchy: Electrophysiological evidence in humans. *Brain Topogr*. doi: 10.1007/s10548-013-0328-4

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

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

French-St. George, M., and Bregman, A. S. (1989). Role of predictability of sequence in auditory stream segregation. *Percept. Psychophys.* 46, 384–386. doi: 10.3758/BF03204992

Giard, M.-H., Perrin, F., Pernier, J., and Bouchet, P. (1990). Brain generators implicated in the processing of auditory stimulus deviance: a topographic event-related potential study. *Psychophysiology* 27, 627–640. doi: 10.1111/j.1469-8986.1990.tb03184.x

Godsmark, D., and Brown, G. J. (1999). A blackboard architecture for computational auditory scene analysis. *Speech Commun.* 27, 351–366. doi: 10.1016/S0167-6393(98)00082-X

Gregory, R. L. (1980). Perceptions as hypotheses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 290, 181–197. doi: 10.1098/rstb.1980.0090

Griffiths, T. D., and Warren, J. D. (2004). What is an auditory object? *Nat. Rev. Neurosci.* 5, 887–892. doi: 10.1038/nrn1538

Grimm, S., and Escera, C. (2012). Auditory deviance detection revisited: evidence for a hierarchical novelty system. *Int. J. Psychophysiol.* 85, 88–92. doi: 10.1016/j.ijpsycho.2011.05.012

Grimm, S., Escera, C., Slabu, L., and Costa-Faidella, J. (2011). Electrophysiological evidence for the hierarchical organization of auditory change detection in the human brain. *Psychophysiology* 48, 377–384. doi: 10.1111/j.1469-8986.2010.01073.x

Grossberg, S., Govindarajan, K. K., Wyse, L. L., and Cohen, M. A. (2004). ARTSTREAM: a neural network model of auditory scene analysis and source segregation. *Neural Netw.* 17, 511–536. doi: 10.1016/j.neunet.2003.10.002

Guldenschuh, M., and Höldrich, R. (2013). Prediction filter design for active noise cancellation headphones. *IET Signal Process.* 7, 497–504. doi: 10.1049/iet-spr.2012.0161

Gutschalk, A., Micheyl, C., Melcher, J. R., Rupp, A., Scherg, M., and Oxenham, A. J. (2005). Neuromagnetic correlates of streaming in human auditory cortex. *J. Neurosci.* 25, 5382–5388. doi: 10.1523/JNEUROSCI.0347-05.2005

Haykin, S., and Chen, Z. (2005). The cocktail party problem. *Neural Comput.* 17, 1875–1902. doi: 10.1162/0899766054322964

Helmholtz, H. (1859). *On the Sensations of Tone as a Physiological Basis for the Theory of Music. 2nd English Edn.* (Trans. A. J. Ellis, 1885) Whitefish, MT: Reprinted by Kessinger Publishing, 2005.

Henry, M. J., and Herrmann, B. (2014). Low-frequency neural oscillations support dynamic attending in temporal context. *Timing Time Percept.* 2, 62–86. doi: 10.1163/22134468-00002011

Hill, K. T., Bishop, C. W., Yadav, D., and Miller, L. M. (2011). Pattern of BOLD signal in auditory cortex relates acoustic response to perceptual streaming. *BMC Neurosci.* 12:85. doi: 10.1186/1471-2202-12-85

Horváth, J., Czigler, I., Birkás, E., Winkler, I., and Gervai, J. (2009). Age-related differences in distraction and reorientation in an auditory task. *Neurobiol. Aging* 30, 1157–1172. doi: 10.1016/j.neurobiolaging.2007.10.003

Hughes, G., Desantis, A., and Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: the role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychol. Bull.* 139, 133–151. doi: 10.1037/a0028566

Hukin, R. W., and Darwin, C. J. (1995). Effects of contralateral presentation and of interaural time differences in segregating a harmonic from a vowel. *J. Acoust. Soc. Am.* 98, 1380–1387. doi: 10.1121/1.414348

Hupé, J.-M., and Pressnitzer, D. (2012). The initial phase of auditory and visual scene analysis. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 942–953. doi: 10.1098/rstb.2011.0368

Hutka, S. A., Alain, C., Binns, M. A., and Bidelman, G. M. (2012). Age-related differences in the sequential organization of speech sounds. *J. Acoust. Soc. Am.* 133, 4177–4187. doi: 10.1121/1.4802745

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

Jones, G. L., and Litovsky, R. Y. (2008). Role of masker predictability in the cocktail party problem. *J. Acoust. Soc. Am.* 124, 3818–3830. doi: 10.1121/1.2996336

Jones, M. R. (1976). Time, our lost dimension: toward a new theory of perception, attention, and memory. *Psychol. Rev.* 83, 323–355. doi: 10.1037/0033-295X.83.5.323

Jones, M. R., and Boltz, M. (1989). Dynamic attending and responses to time. *Psychol. Rev.* 96, 459–491. doi: 10.1037/0033-295X.96.3.459

Jones, M. R., Boltz, M., and Kidd, G. (1982). Controlled attending as a function of melodic and temporal context. *Percept. Psychophys.* 32, 211–218. doi: 10.3758/BF03206225

Jones, M. R., Kidd, G., and Wetzel, R. (1981). Evidence for rhythmic attention. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 1059–1073. doi: 10.1037/0096-1523.7.5.1059

Kitterick, P. T., Bailey, P. J., and Summerfield, A. Q. (2010). Benefits of knowing who, where, and when in multi-talker listening. *J. Acoust. Soc. Am.* 127, 2498–2508. doi: 10.1121/1.3327507

Köhler, W. (1947). *Gestalt Psychology: An Introduction to New Concepts in Modern Psychology*. New York, NY: Liveright Publishing Corporation.

Kondo, H. M., and Kashino, M. (2009). Involvement of the thalamocortical loop in the spontaneous switching of percepts in auditory streaming. *J. Neurosci.* 29, 12695–12701. doi: 10.1523/JNEUROSCI.1549-09.2009

Kondo, H. M., Kitagawa, N., Kitamura, M. S., Koizumi, A., Nomura, M., and Kashino, M. (2012). Separability and commonality of auditory and visual bistable perception. *Cereb. Cortex* 22, 1912–1922. doi: 10.1093/cercor/bhr266

Lange, K. (2013). The ups and downs of temporal orienting: a review of auditory temporal orienting studies and a model associating the heterogeneous findings on the auditory N1 with opposite effects of attention and prediction. *Front. Hum. Neurosci.* 7:263. doi: 10.3389/fnhum.2013.00263

Lawo, V., and Koch, I. (in press). Examining age-related differences in auditory attention control using a task-switching procedure. *J. Gerontol. B Psychol. Sci. Soc. Sci*. doi: 10.1093/geronb/gbs107

Lee, A. K. C., and Shinn-Cunningham, B. G. (2008). Effects of frequency disparities on trading of an ambiguous tone between two competing auditory objects. *J. Acoust. Soc. Am.* 123, 4340–4351. doi: 10.1121/1.2908282

Lipp, R., Kitterick, P., Summerfield, Q., Bailey, P. J., and Paul-Jordanov, I. (2010). Concurrent sound segregation based on inharmonicity and onset asynchrony. *Neuropsychologia* 48, 1417–1425. doi: 10.1016/j.neuropsychologia.2010.01.009

Mager, R., Falkenstein, M., Störmer, R., Brand, S., Müller-Spahn, F., and Bullinger, A. H. (2005). Auditory distraction in young and middle-aged adults: a behavioural and event-related potential study. *J. Neural Transm.* 112, 1165–1176. doi: 10.1007/s00702-004-0258-0

Masuda-Katsuse, I., and Kawahara, H. (1999). Dynamic sound stream formation based on continuity of spectral change. *Speech Commun.* 27, 235–259. doi: 10.1016/S0167-6393(98)00084-3

May, P. J. C., and Tiitinen, H. (2010). Mismatch negativity (MMN), the deviance-elicited auditory deflection, explained. *Psychophysiology* 47, 66–122. doi: 10.1111/j.1469-8986.2009.00856.x

McDermott, J. H., Wrobleski, D., and Oxenham, A. J. (2011). Recovering sound sources from embedded repetition. *Proc. Natl. Acad. Sci. U.S.A.* 108, 1188–1193. doi: 10.1073/pnas.1004765108

McDonald, K. L., and Alain, C. (2005). Contribution of harmonicity and location to auditory object formation in free field: evidence from event-related brain potentials. *J. Acoust. Soc. Am.* 118, 1593–1604. doi: 10.1121/1.2000747

Micheyl, C., and Oxenham, A. J. (2010a). Objective and subjective psychophysical measures of auditory stream integration and segregation. *J. Assoc. Res. Otolaryngol.* 11, 709–724. doi: 10.1007/s10162-010-0227-2

Micheyl, C., and Oxenham, A. J. (2010b). Pitch, harmonicity and concurrent sound segregation: psychoacoustical and neurophysiological findings. *Hear. Res.* 266, 36–51. doi: 10.1016/j.heares.2009.09.012

Mill, R. W., Bőhm, T. M., Bendixen, A., Winkler, I., and Denham, S. L. (2013). Modelling the emergence and dynamics of perceptual organisation in auditory streaming. *PLoS Comput. Biol.* 9:e1002925. doi: 10.1371/journal.pcbi.1002925

Moore, B. C. J., and Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acust. United Acust.* 88, 320–333.

Moore, B. C. J., and Gockel, H. E. (2012). Properties of auditory stream formation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 919–931. doi: 10.1098/rstb.2011.0355

Müller, D., Widmann, A., and Schröger, E. (2005). Auditory streaming affects the processing of successive deviant and standard sounds. *Psychophysiology* 42, 668–676. doi: 10.1111/j.1469-8986.2005.00355.x

Näätänen, R. (1992). *Attention and Brain Function*. Hillsdale, NJ: Erlbaum.

Näätänen, R., Astikainen, P., Ruusuvirta, T., and Huotilainen, M. (2010). Automatic auditory intelligence: an expression of the sensory-cognitive core of cognitive processes. *Brain Res. Rev.* 64, 123–136. doi: 10.1016/j.brainresrev.2010.03.001

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

Näätänen, R., Jacobsen, T., and Winkler, I. (2005). Memory-based or afferent processes in mismatch negativity (MMN): a review of the evidence. *Psychophysiology* 42, 25–32. doi: 10.1111/j.1469-8986.2005.00256.x

Näätänen, R., Kujala, T., Escera, C., Baldeweg, T., Kreegipuu, K., Carlson, S., et al. (2012). The mismatch negativity (MMN) - a unique window to disturbed central auditory processing in ageing and different clinical conditions. *Clin. Neurophysiol.* 123, 424–458. doi: 10.1016/j.clinph.2011.09.020

Näätänen, R., and Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology* 24, 375–425. doi: 10.1111/j.1469-8986.1987.tb00311.x

Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., and Winkler, I. (2001). 'Primitive intelligence' in the auditory cortex. *Trends Neurosci.* 24, 283–288. doi: 10.1016/S0166-2236(00)01790-2

Nager, W., Teder-Sälejärvi, W., Kunze, S., and Münte, T. F. (2003). Preattentive evaluation of multiple perceptual streams in human audition. *Neuroreport* 14, 871–874. doi: 10.1097/00001756-200305060-00019

Nelken, I. (2004). Processing of complex stimuli and natural scenes in the auditory cortex. *Curr. Opin. Neurobiol.* 14, 474–480. doi: 10.1016/j.conb.2004.06.005

Opitz, B., Rinne, T., Mecklinger, A., von Cramon, D. Y., and Schröger, E. (2002). Differential contribution of frontal and temporal cortices to auditory change detection: fMRI and ERP results. *Neuroimage* 15, 167–174. doi: 10.1006/nimg.2001.0970

Pekkonen, E. (2000). Mismatch negativity in aging and in Alzheimer's and Parkinson's diseases. *Audiol. Neuro-Otol.* 5, 216–224. doi: 10.1159/000013883

Picton, T. W., Alain, C., Otten, L., Ritter, W., and Achim, A. (2000). Mismatch negativity: different water in the same river. *Audiol. Neuro-Otol.* 5, 111–139. doi: 10.1159/000013875

Pressnitzer, D., and Hupé, J.-M. (2006). Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Curr. Biol.* 16, 1351–1357. doi: 10.1016/j.cub.2006.05.054

Pressnitzer, D., Sayles, M., Micheyl, C., and Winter, I. M. (2008). Perceptual organization of sound begins in the auditory periphery. *Curr. Biol.* 18, 1124–1128. doi: 10.1016/j.cub.2008.06.053

Pressnitzer, D., Suied, C., and Shamma, S. A. (2011). Auditory scene analysis: the sweet music of ambiguity. *Front. Hum. Neurosci.* 5:158. doi: 10.3389/fnhum.2011.00158

Prinz, W. (2006). What re-enactment earns us. *Cortex* 42, 515–517. doi: 10.1016/S0010-9452(08)70389-7

Rahne, T., and Sussman, E. (2009). Neural representations of auditory input accommodate to the context in a dynamically changing acoustic environment. *Eur. J. Neurosci.* 29, 205–211. doi: 10.1111/j.1460-9568.2008.06561.x

Rajendran, V. G., Harper, N. S., Willmore, B. D., Hartmann, W. M., and Schnupp, J. W. H. (2013). Temporal predictability as a grouping cue in the perception of auditory streams. *J. Acoust. Soc. Am.* 134, EL98–E104. doi: 10.1121/1.4811161

Rauschecker, J. P., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11800–11806. doi: 10.1073/pnas.97.22.11800

Rimmele, J. M., Schröger, E., and Bendixen, A. (2012). Age-related changes in the use of regular patterns for auditory scene analysis. *Hear. Res.* 289, 98–107. doi: 10.1016/j.heares.2012.04.006

Rogers, W. L., and Bregman, A. S. (1993). An experimental evaluation of three theories of auditory stream segregation. *Percept. Psychophys.* 53, 179–189. doi: 10.3758/BF03211728

Scherg, M., Vajsar, J., and Picton, T. W. (1989). A source analysis of the late human auditory evoked potentials. *J. Cogn. Neurosci.* 1, 336–355. doi: 10.1162/jocn.1989.1.4.336

Schneider, B. A., Daneman, M., and Pichora-Fuller, M. K. (2002). Listening in aging adults: from discourse comprehension to psychoacoustics. *Can. J. Exp. Psychol.* 56, 139–152. doi: 10.1037/h0087392

Schröger, E., Bendixen, A., Denham, S. L., Mill, R., Bőhm, T. M., and Winkler, I. (in press). Predictive regularity representations in violation detection and auditory stream segregation: from conceptual to computational models. *Brain Topogr*. doi: 10.1007/s10548-013-0334-6

Schröger, E., and Roeber, U. (2011). A difference in the brain's ability to automatically encode serially deterministic and stochastic regularities. *Front. Hum. Neurosci. Conf. Abstr. XI Int. Conf. Cogn. Neurosci.* doi: 10.3389/conf.fnhum.2011.207.00031

Schubotz, R. I. (2007). Prediction of external events with our motor system: towards a new framework. *Trends Cogn. Sci.* 11, 211–218. doi: 10.1016/j.tics.2007.02.006

Schwartz, A. H., and Shinn-Cunningham, B. G. (2010). Dissociation of perceptual judgments of "what" and "where" in an ambiguous auditory scene. *J. Acoust. Soc. Am.* 128, 3041–3051. doi: 10.1121/1.3495942

Schwartze, M., Farrugia, N., and Kotz, S. A. (2013). Dissociation of formal and temporal predictability in early auditory evoked potentials. *Neuropsychologia* 51, 320–325. doi: 10.1016/j.neuropsychologia.2012.09.037

Shamma, S. A., Elhilali, M., and Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends Neurosci.* 34, 114–123. doi: 10.1016/j.tins.2010.11.002

Shinn-Cunningham, B. G., and Best, V. (2008). Selective attention in normal and impaired hearing. *Trends Amplific.* 12, 283–299. doi: 10.1177/1084713808325306

Shinn-Cunningham, B. G., Lee, A. K. C., and Oxenham, A. J. (2007). A sound element gets lost in perceptual competition. *Proc. Natl. Acad. Sci. U.S.A.* 104, 12223–12227. doi: 10.1073/pnas.0704641104

Snyder, E., and Hillyard, S. A. (1976). Long-latency evoked potentials to irrelevant, deviant stimuli. *Behav. Biol.* 16, 319–331. doi: 10.1016/S0091-6773(76)91447-4

Snyder, J. S., and Alain, C. (2005). Age-related changes in neural activity associated with concurrent vowel segregation. *Cogn. Brain Res.* 24, 492–499. doi: 10.1016/j.cogbrainres.2005.03.002

Snyder, J. S., and Alain, C. (2007a). Sequential auditory scene analysis is preserved in normal aging adults. *Cereb. Cortex* 17, 501–512. doi: 10.1093/cercor/bhj175

Snyder, J. S., and Alain, C. (2007b). Toward a neurophysiological theory of auditory stream segregation. *Psychol. Bull.* 133, 780–799. doi: 10.1037/0033-2909.133.5.780

Snyder, J. S., Carter, O. L., Hannon, E. E., and Alain, C. (2009a). Adaptation reveals multiple levels of representation in auditory stream segregation. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 1232–1244. doi: 10.1037/a0012741

Snyder, J. S., Holder, W. T., Weintraub, D. M., Carter, O. L., and Alain, C. (2009b). Effects of prior stimulus and prior perception on neural correlates of auditory stream segregation. *Psychophysiology* 46, 1208–1215. doi: 10.1111/j.1469-8986.2009.00870.x

Sperduti, M., Tallon-Baudry, C., Hugueville, L., and Pouthas, V. (2011). Time is more than a sensory feature: attenting to duration triggers specific anticipatory activity. *Cogn. Neurosci.* 2, 11–18. doi: 10.1080/17588928.2010.513433

Spielmann, M. I., Schröger, E., Kotz, S. A., and Bendixen, A. (in press). Attention effects on auditory scene analysis: insights from event-related brain potentials. *Psychol. Res*. doi: 10.1007/s00426-014-0547-7

Spielmann, M. I., Schröger, E., Kotz, S. A., Pechmann, T., and Bendixen, A. (2013). Using a staircase procedure for the objective measurement of auditory stream integration and segregation thresholds. *Front. Psychol.* 4:534. doi: 10.3389/fpsyg.2013.00534

Sussman, E., Ritter, W., and Vaughan, H. G. Jr. (1999). An investigation of the auditory streaming effect using event-related brain potentials. *Psychophysiology* 36, 22–34. doi: 10.1017/S0048577299971056

Sussman, E., Čeponienė, R., Shestakova, A., Näätänen, R., and Winkler, I. (2001). Auditory stream segregation processes operate similarly in school-aged children and adults. *Hear. Res.* 153, 108–114. doi: 10.1016/S0378-5955(00)00261-6

Sussman, E. S. (2007). A new view on the MMN and attention debate: the role of context in processing auditory events. *J. Psychophysiol.* 21, 164–175. doi: 10.1027/0269-8803.21.34.164

Sussman, E. S., Bregman, A. S., Wang, W. J., and Khan, F. J. (2005). Attentional modulation of electrophysiological activity in auditory cortex for unattended sounds within multistream auditory environments. *Cogn. Affect. Behav. Neurosci.* 5, 93–110. doi: 10.3758/CABN.5.1.93

Sussman, E. S., and Gumenyuk, V. (2005). Organization of sequential sounds in auditory memory. *Neuroreport* 16, 1519–1523. doi: 10.1097/01.wnr.0000177002.35193.4c

Sussman, E. S., Horváth, J., Winkler, I., and Orr, M. (2007). The role of attention in the formation of auditory streams. *Percept. Psychophys.* 69, 136–152. doi: 10.3758/BF03194460

Sussman, E., and Steinschneider, M. (2006). Neurophysiological evidence for context-dependent encoding of sensory input in human auditory cortex. *Brain Res.* 1075, 165–174. doi: 10.1016/j.brainres.2005.12.074

Szalárdy, O., Bendixen, A., Bőhm, T. M., Davies, L. A., Denham, S. L., and Winkler, I. (in press). The effects of rhythm and melody on auditory stream segregation. *J. Acoust. Soc. Am.* doi: 10.1121/1.4865196

Szalárdy, O., Bendixen, A., Tóth, D., Denham, S. L., and Winkler, I. (2013). Modulation frequency acts as a primary cue for auditory stream segregation. *Learn. Percept.* 5, 149–161. doi: 10.1556/LP.5.2013.Suppl2.9

Tiitinen, H., May, P., Reinikainen, K., and Näätänen, R. (1994). Attentive novelty detection in humans is governed by pre-attentive sensory memory. *Nature* 372, 90–92. doi: 10.1038/372090a0

Trainor, L. J., and Trehub, S. E. (1989). Aging and auditory temporal sequencing: ordering the elements of repeating tone patterns. *Percept. Psychophys.* 45, 417–426. doi: 10.3758/BF03210715

Van Noorden, L. P. A. S. (1975). *Temporal Coherence in the Perception of Tone Sequences.* Doctoral dissertation, Technical University Eindhoven, Eindhoven.

Wacongne, C., Changeux, J. P., and Dehaene, S. (2012). A neuronal model of predictive coding accounting for the mismatch negativity. *J. Neurosci.* 32, 3665–3678. doi: 10.1523/JNEUROSCI.5003-11.2012

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt II [Laws of organization in perceptual forms II]. *Psychol. Forsch.* 4, 301–350. doi: 10.1007/BF00410640

Winkler, I. (2007). Interpreting the mismatch negativity. *J. Psychophysiol.* 21, 147–163. doi: 10.1027/0269-8803.21.34.147

Winkler, I., and Czigler, I. (2012). Evidence from auditory and visual event-related potential (ERP) studies of deviance detection (MMN and vMMN) linking predictive coding theories and perceptual object representations. *Int. J. Psychophysiol.* 83, 132–143. doi: 10.1016/j.ijpsycho.2011.10.001

Winkler, I., Denham, S. L., Mill, R., Bőhm, T. M., and Bendixen, A. (2012). Multistability in auditory stream segregation: a predictive coding view. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 1001–1012. doi: 10.1098/rstb.2011.0359

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

Winkler, I., Karmos, G., and Näätänen, R. (1996). Adaptive modeling of the unattended acoustic environment reflected in the mismatch negativity event-related potential. *Brain Res.* 742, 239–252. doi: 10.1016/S0006-8993(96)01008-6

Winkler, I., Kushnerenko, E., Horváth, J., Čeponienė, R., Fellman, V., Huotilainen, M., et al. (2003a). Newborn infants can organize the auditory world. *Proc. Natl. Acad. Sci. U.S.A.* 100, 11812–11815. doi: 10.1073/pnas.2031891100

Winkler, I., Paavilainen, P., Alho, K., Reinikainen, K., Sams, M., and Näätänen, R. (1990). The effect of small variation of the frequent auditory stimulus on the event-related brain potential to the infrequent stimulus. *Psychophysiology* 27, 228–235. doi: 10.1111/j.1469-8986.1990.tb00374.x

Winkler, I., Sussman, E., Tervaniemi, M., Horváth, J., Ritter, W., and Näätänen, R. (2003b). Preattentive auditory context effects. *Cogn. Affect. Behav. Neurosci.* 3, 57–77. doi: 10.3758/CABN.3.1.57

Winkler, I., Takegata, R., and Sussman, E. (2005). Event-related brain potentials reveal multiple stages in the perceptual organization of sound. *Cogn. Brain Res.* 25, 291–299. doi: 10.1016/j.cogbrainres.2005.06.005

Winkler, I., Teder-Sälejärvi, W. A., Horváth, J., Näätänen, R., and Sussman, E. (2003c). Human auditory cortex tracks task-irrelevant sound sources. *Neuroreport* 14, 2053–2056. doi: 10.1097/00001756-200311140-00009

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

Received: 15 February 2014; accepted: 14 March 2014; published online: 31 March 2014.

Citation: Bendixen A (2014) Predictability effects in auditory scene analysis: a review. Front. Neurosci. ***8****:60. doi: 10.3389/fnins.2014.00060*

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

frontiers in NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 07 April 2014
doi: 10.3389/fnins.2014.00064

Do audio-visual motion cues promote segregation of auditory streams?

Lidia Shestopalova[1]*, Tamás M. Bőhm[2,3], Alexandra Bendixen[4], Andreas G. Andreou[5,6], Julius Georgiou[6], Guillaume Garreau[6], Botond Hajdu[2], Susan L. Denham[7] and István Winkler[2,8]

[1] Pavlov Institute of Physiology, Russian Academy of Sciences, St.-Petersburg, Russia
[2] Research Centre for Natural Sciences, Institute of Cognitive Neuroscience and Psychology, Hungarian Academy of Sciences, Budapest, Hungary
[3] Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary
[4] Auditory Psychophysiology Lab, Department of Psychology, Cluster of Excellence "Hearing4all," European Medical School, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany
[5] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA
[6] Department of Electrical and Computer Engineering, University of Cyprus, Nicosia, Cyprus
[7] School of Psychology, Cognition Institute, University of Plymouth, Plymouth, UK
[8] Department of Cognitive and Neuropsychology, Institute of Psychology, University of Szeged, Szeged, Hungary

Edited by:
Susann Deike, Leibniz Institute for Neurobiology, Germany

Reviewed by:
Adrian K. C. Lee, University of Washington, USA
Torsten Rahne, University Hospital Halle (Saale), Germany

***Correspondence:**
Lidia Shestopalova, Pavlov Institute of Physiology, Russian Academy of Sciences, Makarova emb.,6, St.-Petersburg 199034, Russia
e-mail: shestolido@mail.ru

An audio-visual experiment using moving sound sources was designed to investigate whether the analysis of auditory scenes is modulated by synchronous presentation of visual information. Listeners were presented with an alternating sequence of two pure tones delivered by two separate sound sources. In different conditions, the two sound sources were either stationary or moving on random trajectories around the listener. Both the sounds and the movement trajectories were derived from recordings in which two humans were moving with loudspeakers attached to their heads. Visualized movement trajectories modeled by a computer animation were presented together with the sounds. In the main experiment, behavioral reports on sound organization were collected from young healthy volunteers. The proportion and stability of the different sound organizations were compared between the conditions in which the visualized trajectories matched the movement of the sound sources and when the two were independent of each other. The results corroborate earlier findings that separation of sound sources in space promotes segregation. However, no additional effect of auditory movement *per se* on the perceptual organization of sounds was obtained. Surprisingly, the presentation of movement-congruent visual cues did not strengthen the effects of spatial separation on segregating auditory streams. Our findings are consistent with the view that bistability in the auditory modality can occur independently from other modalities.

Keywords: auditory stream segregation, bistable perception, auditory motion, audio-visual integration

INTRODUCTION

In natural environments, our senses provide us with an abundance of incoming information, which must be analyzed and segregated in order to form veridical representations of perceptual objects. In particular, the auditory system needs to separate the information relating to concurrently active sound sources to construct a consistent and stable interpretation of the acoustic environment. The perceptual task of analyzing a sound mixture has been termed "auditory scene analysis" by Bregman (1990). In everyday situations, perception is usually multimodal; often the currently interesting sound source is also in the focus of our visual attention or can be brought into it at will. A number of studies have provided evidence that auditory perception is enhanced by integration of information from multiple sensory modalities (e.g., Sumby and Pollack, 1954; Grant and Seitz, 2000; Eramudugolla et al., 2011; for a review, see Recanzone, 2009). However, much less is known about whether and how visual cues can support sequential auditory scene analysis. Therefore, in the current study we tested the effects of congruent visual spatial cues on the segregation of two series of interleaved sounds generated by the same or two different stationary or moving sound sources.

The process of organizing sounds into coherent sequences termed "auditory streams" has been extensively studied with the help of the auditory streaming paradigm in which listeners hear a repeating triplet composed of two kinds of sounds ("ABA_ABA_") which differ from each other in some acoustic feature (for reviews, see Moore and Gockel, 2002; Cusack et al., 2004; Snyder and Alain, 2007; Winkler et al., 2009, 2012). This stimulus configuration is most commonly perceived either as a single coherent sequence with a galloping rhythm (termed the "integrated" percept) or as two concurrent streams, one consisting of the A and the other of the B sounds ("segregated" percept; Van Noorden, 1975). When listeners are presented with long unchanging sequences of this type, they report spontaneous perceptual switches between these alternative sound organizations (Gutschalk et al., 2005; Denham and Winkler, 2006; Pressnitzer and Hupé, 2006; Kondo and Kashino, 2009; Denham et al., 2010; Hill et al., 2012). These perceptual switches occur even

when the stimulus configuration strongly promotes one alternative organization over another (Denham et al., 2013). Although somewhat less commonly used, alternating (ABAB_) sequences are also typically perceived in terms of one integrated (alternating) or two segregated (A and B, separately) streams (e.g., Yabe et al., 2001; Shinozaki et al., 2003) and exposure to long alternating sequences results in perceptual bistability (Bőhm et al., 2013; Szalárdy et al., 2014).

Empirical evidence consistently shows that separation in various auditory features between the A and B sounds (such as frequency, pitch, timbre, loudness, amplitude modulation and spatial location; e.g., Vliegen and Oxenham, 1999; Grimault et al., 2002; Roberts et al., 2002; Szalárdy et al., 2013; for a review, see Moore and Gockel, 2002) acts as the main cue for auditory stream segregation. The larger the acoustic separation between the A and B sounds and the faster the presentation rate, the more likely that the ABA_ sequence is perceived in terms of two separate sound streams. Although separation in sound source location is regarded as a weak cue for auditory stream segregation (e.g., Culling and Summerfield, 1995; for a review, see Bregman, 1990), when combined with separation in tone frequency, it was shown to increase the probability of listeners reporting the segregated percept (Denham et al., 2010; Szalárdy et al., 2013). These effects, based on acoustic separation (perceptual dissimilarity) between the feature values of the A and B sounds, reflect the Gestalt principle of similarity.

Other Gestalt principles have also been shown to contribute to auditory scene analysis (e.g., "closure," see Bendixen et al., 2010). The Gestalt principle of common fate refers to similarity in perceptual trajectories, rather than in single feature values. One example is the various acoustic transformations resulting from the movement of sound sources. Bőhm et al. (2013) tested whether the potential cues resulting from auditory motion can also help to segregate two interleaved tone sequences. They varied the motion pattern (stationary location vs. two different kinds of trajectories) and the dynamics of spatial separation (standing/moving side by side or standing apart/moving on separate trajectories) of the two sound sources. Although the results showed a clear effect of spatial separation in general, auditory motion *per se* did not exert a significant influence on the proportions of time in which the segregated and integrated organizations were perceived by the listeners. Following up on these findings, we designed the present experiment to investigate whether the auditory spatial and motion cues would become more effective in supporting stream segregation when presented together with congruent visual cues.

The ability of the human brain to link information from different modalities to the same object is termed multisensory integration. Binding information from different modalities is assumed to help to reduce noise within the perceptual system (see Stein et al., 2004; Koelewijn et al., 2010). A large number of crossmodal spatial audiovisual studies have focused on the effects of presenting a congruent or conflicting unimodal cue in one modality on the detection or discrimination of targets presented in a different modality (for reviews, see Spence et al., 2004; Wright and Ward, 2008). Whereas the results of earlier cross-modal studies showed asymmetries between the effects of visual information on auditory perception and vice versa, the currently prevailing view is that crossmodal exogenous spatial cuing effects can be demonstrated for all possible combinations of successively presented auditory and visual stimuli, although the magnitude of the spatial cuing effects induced by the crossmodal cues may be different (Spence, 2010). Further, whereas some studies concluded that multisensory cues are no more effective than unimodal cues in capturing spatial attention (Ward, 1994; Spence and Driver, 2000; Vroomen et al., 2001; Santangelo et al., 2006, 2008; Van der Burg et al., 2008), the results of some more recent studies have overturned this conclusion by showing that multisensory cues do capture spatial attention more effectively than unisensory cues, at least under conditions of high perceptual load (i.e., when participants are simultaneously engaged in a concurrent perceptually demanding task; for reviews, see Spence and Santangelo, 2009; Spence, 2010) and when the unisensory components of the multisensory signals are presented from more-or-less the same spatial location (Ho et al., 2007, 2009).

Although most research on multisensory interaction focused on the integration of static objects, there has been a recent growth of interest in multisensory interactions specifically influencing the perception of motion. A number of investigations employing apparent motion have demonstrated the existence of robust interactions between sensory modalities during the extraction of motion information (for a review, see Soto-Faraco and Väljamäe, 2012). In particular, studies of the effect of directional audiovisual congruency (cross-modal dynamic capture) demonstrated that sound motion discrimination performance can be substantially enhanced when a concurrent visual motion stream is synchronously presented from the same direction as the sounds, and eliminated when the visual and auditory signals are desynchronized in time. Further, the magnitude of the dynamic capture effect far exceeded what could have been expected on the basis of simple static capture, suggesting that motion-related cross-modal interactions are probably based on motion information *per se*, rather than on other features, such as the location and timing of the stimuli (Soto-Faraco et al., 2003, 2004).

Binding processes within and across sensory modalities have also been studied using multistable stimuli. There is substantial evidence that multistable effects in one modality can be modified by stimuli in another modality (Schwartz et al., 2012). It has been shown that sounds congruent with one or the other image may affect perceptual dominance proportions in binocular rivalry, but only when the visual stimulus is consciously perceived (Kang and Blake, 2005; Conrad et al., 2010). The visual influence on auditory perception in speech was examined using a speech stimulus that generated bistable perception and a dynamic version of the Rubin vase illusion (Munhall et al., 2009). The results indicated that visual influences on auditory speech processing, at least for the McGurk illusion, necessitate conscious perception of the visual speech cues, thus supporting the hypothesis that multisensory speech integration is not completed at an early processing stage. Van Ee et al. (2009) have shown that attentional control over the perceptual organization of an ambiguous auditory stimulus can be markedly aided by matching visual stimuli. They argued that the audiovisual binding that served awareness is not fully automatic: stimuli in one modality only influence the perception of an ambiguous stimulus configuration in the other modality when

one attends to the stimulus presented in the first modality (Van Ee et al., 2009).

In the current study, we investigated whether congruent spatial visual information affects the segregation of two interleaved sounds sequences emitted by one or two stationary or moving sound sources. We hypothesized that congruent unambiguous visual spatial information would reduce the ambiguity of the auditory scene and thus bias the emerging auditory perceptual organization. That is, when participants are presented with sounds originating from one or two sound sources, congruent as compared to incongruent visual spatial information was expected to increase the proportion of time in which perception was veridical.

METHODS

PARTICIPANTS

Twenty-five young adults (10 females; 23 right-handed; 18–26 years, mean age 21.6 years) took part in the experiment for modest financial compensation. Participants were screened in advance for (1) normal hearing (thresholds below 25 dB above hearing level, measured at 250, 500, 1000, 2000, and 4000 Hz, and a maximum threshold difference between the two ears of 10 dB at 250 Hz, 5 dB at 500 and 1000 Hz, 10 dB at 2000 and 4000 Hz were required) and for (2) the ability to detect synchrony between fluctuations in auditory and visual stimuli (see in the *Procedure* section). All participants had normal or corrected-to-normal eyesight. Because our measure of the effects of the stimulus manipulations requires that listeners experience perceptual switches, the data provided by four participants were discarded from the analysis, as they did not experience/report perceptual switches in five or more stimulus conditions. (All other listeners experienced/reported perceptual switches in all but two or fewer stimulus conditions.) Thus the analyses are conducted on data from 21 listeners (10 females, 19 right-handed, 18–26 years, mean age 21.5 years). Participants gave written informed consent after the aims and procedures of the experiment were explained to them. The study was approved by the Ethical Committee of the Institute of Cognitive Neuroscience and Psychology, Research Center for Natural Sciences, Hungarian Academy of Sciences.

APPARATUS AND STIMULI

Audio recordings

The auditory stimuli were based on audio recordings made on the stage of the Ceremony Hall of the University of Cyprus. Preparatory tests showed no significant reverberations in this area. Alternating sequences of two 100 ms long (including onset and offset ramps of 10 ms long raised cosines) pure tones of equal amplitude were recorded with 25 ms inter-stimulus interval. One tone had a frequency of 400 Hz (subsequently denoted "low"), the other was three semitones higher at 475.7 Hz (subsequently denoted "high"). The tone sequences were played by Anthony Gallo Acoustics A'Diva Ti compact speakers mounted on construction helmets worn and carried around by two experimenters (TMB and GG). The speaker drivers were facing upwards to provide equal sound emission in horizontal directions. The sound signals were generated on an IBM PC and transmitted to the speakers by FM radio units. Sounds were recorded with a Head Acoustics HSU III.2 head microphones mounted within the artificial head placed at the center of the stage, and digitized with a National Instruments 4462 data acquisition card of an IBM PC at a sampling rate of 96 kHz and 24 bits resolution. In order to reduce noise inherent in on-site sound recordings, the audio signals were post-processed by removing the DC offset and bandpass filtering in the 350–526 Hz range with a third-order Butterworth filter.

In order to create 4-min-long stimulus blocks for each experimental condition, appropriate segments were selected from the audio recordings and extended to exactly 4 min by looping. We chose the longest possible segment of the signal that had good recording quality throughout and could be looped. We avoided introducing discontinuities into the spatial trajectories by selecting sections in which the initial and final estimated interaural time differences and their first derivatives roughly matched (separately for the high and the low tones). The endpoints of the sections were always placed in the middle of the silent interval before a low tone. While looping the audio segments to block length, a cross-fading procedure was applied in the 1 ms vicinity of the concatenation points to prevent audible clicks. The length of the segments chosen ranged between 20 and 64 s. Finally, the signals were re-sampled at 192 kHz and their intensity was normalized across the segments. The alternating (ABAB) tone sequence was chosen for the recordings because it provided more potential section endpoints than the repetitive triplet pattern (ABA_) typically used in the auditory streaming paradigm (Van Noorden, 1975). In a pilot study, we found that using identical stimulus parameters (pitch separation and presentation rate), the ABAB and ABA_ patterns produce similar distributions of the reported percepts.

Visual stimuli

The spatial positions of the helmet-loudspeakers were recorded by means of a Microsoft Kinect device connected to a PC and interfaced to Matlab through the Open Natural Interaction toolbox. The helmets were painted red and blue to facilitate their tracking in the video data. The Kinect's depth sensor provided a three dimensional map of its field-of-view while a color video was recorded by its camera sensor. The trajectory data was smoothed by median filtering and moving averaging, and re-sampled to 20 frames-per-second. Visual stimuli for the experiment were then created from the movement trajectories using Unity 3D computer animation software. The sound-emitting moving objects were represented by two colored balls (one red, the other blue) moving along the movement trajectories in a three dimensional room-like space. The videos had a refresh rate of 40 Hz which provided frame durations of 25 ms. Thus each sound was covered exactly by four frames in the video and the inter-stimulus interval by a single frame, which allowed the colored balls to pulsate in full synchrony with the presentation of the corresponding tones. Pulsation was created by modulating the size of the visual objects with a size ratio of 10% between the "sound-on" and "sound-off" states. The audio and video recordings were synchronized by storing timestamps for the beginning of each sound recording and for each video frame captured by the Kinect.

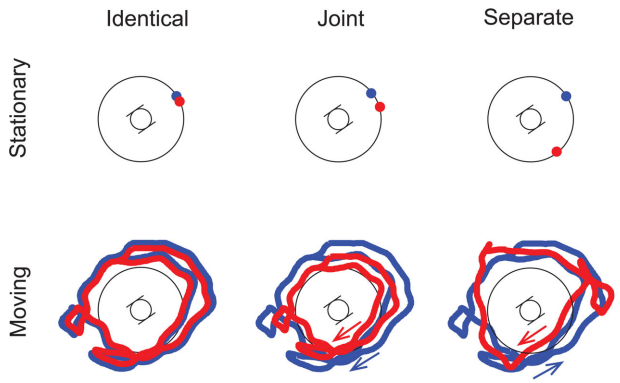


FIGURE 1 | Schematic illustration of the spatial trajectories of the sound sources. The artificial head was placed at the center of the stage, and listeners perceived the stimuli as if sitting in this position. The two trajectories are drawn in red and blue. For *Joint* and *Separate* motion, arrows indicate the direction of movement for each trajectory. Note that the trajectories are illustrations only; they do not depict the actual recordings. (Adapted by permission from Figure 1, Bőhm et al., 2013.)

Conditions

The two experimenters, each wearing a helmet with the loudspeaker on top, performed a number of different scenarios during the recording session (cf. **Figure 1**). Each experimental condition was based on the recording of a separate scenario. The two kinds of sound Motion scenarios tested were *Stationary* (based on scenarios with both experimenters standing still) and *Moving* (the experimenters were walking on a random trajectory around the artificial head). The Co-location of the two sound sources could either be characterized by *Identical* (based on recordings with one experimenter standing or moving, his helmet-loudspeaker delivering the full alternating sequence), *Joint* (the two experimenters stood or moved together hand-in-hand, i.e., the distance between their trajectories being roughly constant, with one loudspeaker delivering the high, the other the low tones), and *Separate* trajectories (when the two experimenters were standing or moving separately, again the two loudspeakers delivering different tones) (see **Figure 1**). The cross-modal Congruency of the visual and auditory stimuli was represented by the *Congruent* (the visualized movement matched the auditory movement of the objects) and the *Incongruent* conditions (the visualized movement was independent of the auditory movement). The auditory stimulation in the *Congruent* and *Incongruent* conditions was the same. The conditions were fully crossed, leading to 12 stimulus conditions in total. Samples for the *Congruent*/*Moving*/*Separate* and *Incongruent*/*Moving*/*Joint* conditions can be found at http://figshare.com/articles/Stimulus_Congruent_Moving_Separate/961766 and http://figshare.com/articles/Stimulus_Incongruent_Moving_Joint/961765.

The visual stimuli were made incongruent by applying four transformations to the spatial trajectories of the two visual objects: (1) the trajectories of the two objects were exchanged (i.e., the recorded trajectory of the "blue" object served as the basis of the incongruent trajectory of the "red" object and *vice versa*), (2) the trajectories were inverted in time, (3) the trajectories were mirrored producing a left-right flip (as rendered on the screen), and (4) the trajectories were rotated in the horizontal plane around the view-point of the observer by a random amount. The amount of rotation was drawn from a uniform distribution independently for each listener and condition, so that (a) it was between 30 and 330° and (b) the rotated trajectory fell into the field-of-view of the animation for the *Stationary* conditions. The same random amount of rotation was applied to the trajectories of both visual objects, except for the *Moving*/*Separate* conditions. In the latter case, the amounts of rotation of the two trajectories were independently drawn. As a result, the incongruent trajectories had the same statistical properties as the congruent ones (e.g., the distribution of velocities was the same), and the type of visual motion was compatible with that delivered by the audio (e.g., the two objects moved randomly, independently from each other, both in the video and the audio). However, the *Incongruent* videos differed from the *Congruent* ones sufficiently to prevent binding of the auditory and visual stimuli (i.e., the objects could be seen definitely elsewhere from where the sound came from).

PROCEDURE

Participants were preselected on the basis of the results of a synchrony test designed to check whether they were able to bind auditory and visual stimuli into a single perceptual object. To this end, before enrolling a participant on the experiment, they were presented with a demonstration of the stimuli that consisted of 7 examples of 20 s duration each. For each example, the participant was asked to report whether he/she recognized the visual objects as sound sources. If not, the example was repeated a few times. Only participants who reported experiencing the sounds as originating from the visual objects in the congruent examples were entered into the next phase of the experiment. In the first three examples, a single stationary red ball (identical to the one created for the main experiment) was shown at the center of the screen. In the first example it was pulsating at a rate of 1 Hz, in full synchrony with the tone sequence. Next an "asynchronous" example was presented which employed the same visual and auditory stimuli but the sound sequence was delivered with a 500 ms delay. In all other examples the sound sequences were pulsating synchronously with the visual stimuli. In the third example the pulsation rate of the visual object and the sound sequence was increased to 2 Hz. After these examples, two visual objects (a red ball in the center and a blue ball on the left side of the screen) were displayed, first with a pulsation rate of 2.67 Hz and then with 8 Hz, the latter being equal to the presentation rate used in the main experiment. For these demonstrations, the trajectories of the visual objects and the sound sequences were taken from the *Stationary*/*Separate* condition. To create the examples with lower pulsation rate, the necessary number of sound bursts was periodically muted, resulting in a sound sequence that could be experienced as pulsating synchronously with the visual stimuli. In the last two examples, the *Moving*/*Separate* condition was taken as the basis for creating both the auditory and the visual stimuli. Two visual objects were moving separately and they pulsated at a rate of 2.67 Hz in the first example and at 8 Hz in the second one.

After these demonstrations, synchrony tests containing 2–5 four-trial blocks were conducted. On each trial the participant

received an audiovisual stimulus of 20 s duration representing the *Stationary*/*Separate* condition (with the red and the blue ball placed at the center and on the left side of the screen, respectively). The sounds were presented at the rate of 8 Hz and on half of the trials, the visual object pulsated in full synchrony with the tone sequence. On the other half of the trials, the visual pulsation rate differed from the tone presentation rate by 30% (on half of these trials it was higher, on the other half it was lower). The test ended when the participant correctly judged the synchrony of at least 7 trials within 2 consecutive four-trial blocks, or the maximum number of trials (20 trials; 5 blocks) was reached. Participants who reached the criterion continued with the second test. The second test only differed from the first in that the visual pulsation rate could differ from the tone presentation rate by only ±20%. Participants passed this test by correctly judging synchrony for at least 6 trials within 2 consecutive blocks of trials before the maximum number of trials (20) was reached. Participants were only enrolled in the study if they passed both tests. Those who did not pass were enrolled in a different experiment.

Each condition was administered in one 4-min long stimulus block during the experimental session. The order of the 12 conditions was randomized separately for each participant. There were short breaks lasting about 30 s between successive stimulus blocks and participants could choose to have a 5-min break after any block (or they had it at the 8th block, if they did not ask for it before). The stimuli were played by an IBM PC with an Audiotrak Prodigy HD2 sound card. Sounds were amplified by a custom-made mixer-amplifier and delivered by Etymotic Research ER-2 insert earphones. The insert earphones provided at least 30 dB external noise attenuation and made sure that participants heard the sounds as if standing where the artificial head was located at the time when the sounds were recorded. Thus the binaural location cues related to head-related transfer functions were adequately reproduced, while the monaural cues (e.g., those associated with pinna) were shown not to contribute greatly to stream segregation in a similar sound configuration (Bőhm et al., 2013). The sound level was set to 50 dB above the hearing threshold of the participant, as determined immediately before the experiment in a simplified staircase measurement using a sequence alternating the two tones (400 and 475.7 Hz). The visual stimuli were presented with an NVIDIA GeForce 9500 GT video card at a resolution of 1280∗720 pixels and delivered on a Samsung LE40C530F1W LCD TV with a screen diameter of 101 cm. The participant was seated 35 cm in front of the TV screen resulting in a 103° wide viewing angle of the scene. The participant's head was supported by a chin holder to avoid head movements during the stimulus blocks. The colored balls came off-screen when their location exceeded the viewing angle of the TV screen. (This never happened in the Stationary conditions.) Given that the random trajectories of the movement of the sources provided an approximately even distribution of the horizontal viewing angle, the amount of time the balls spent off-screen was proportional to the viewing angle and it approximately corresponded to how a viewer with fixed head direction would have seen the sources in a real-life scene.

Prior to the main experiment, participants received training for the types of patterns they were asked to report using the keys. The "integrated" and "segregated" percepts were explained and illustrated with sound examples to the participants before the experiment and the experimenter made sure they understood the task. Because there is no single prototype for the "both" percept, listeners received explanations but not sound examples for these patterns.

During the main experiment, participants were comfortably seated in an anechoic chamber and were instructed to continuously report the perceived sound organization throughout the entire stimulus block using two response keys, one key held in each hand. They were to keep one key depressed as long as they perceived both high and low tones as part of a single repeating pattern (termed the "integrated" percept). The other key was to be kept depressed as long as they heard tones of the same pitch forming separate repeating patterns (the "segregated" percept). Whenever they heard both types of patterns concurrently, they were to keep both keys depressed (the "both" percept). During those times when the participant did not perceive any repeating sound pattern, he/she was instructed to release both keys (the "neither" percept). The possibility of four choices, as opposed to forcing listeners to choose between reporting integrated or segregated percepts, has been introduced after we found that asking listeners to report their perception verbally in an unconstrained manner resulted in more than two alternatives (Denham et al., 2013; for an exploration of the perception of different patterns, see Denham et al., 2014). The instructions emphasized that the appropriate key combination was to be held as long as the participant perceived the corresponding sound pattern but to be changed immediately when the perceived pattern changed to a different category. Participants were informed that there was no correct or incorrect way to perceive any of the stimulus sequences; therefore they should not try to force to hear the sounds in one or another way. Rather, they should report what they actually hear. A description of the interpretation of the percepts reported by depressing both keys at the same time can be found in Denham et al. (2013). The assignment of the two response keys was counterbalanced across participants.

Besides judgments about the perception of the tone sequences, participants were also given a secondary task the motivation of which was to help binding between the corresponding audio and visual stimuli. The blue and red balls modeling the movement trajectories blinked 5–15 times (uniform distribution) during each 4-min stimulus block, so that one ball at a time became white for 100 ms before returning to its original color (blinks occurred equiprobably and in a random order on the two visual objects). The participant was instructed to count how many times the balls blinked (all blinks, irrespective of on which object they occurred). Following each stimulus block, the participant received feedback on the reported blink count.

DATA RECORDING AND ANALYSIS

The state of the two response keys was sampled at a 40 Hz rate, and the data were analyzed similarly to the procedure used in Denham et al. (2013). For each perceptual phase (i.e., the time interval between two consecutive perceptual switches), the logarithm of its duration in milliseconds and the reported percept was extracted. Perceptual phases shorter than 300 ms were

excluded from the analysis as these presumably originate from inaccurate timing of button presses and releases, rather than from two separate perceptual switches quickly following each other (Moreno-Bote et al., 2010). Based on this data, for each participant, the proportions of the different percepts (i.e., the percent of time the listener experienced a given percept within the stimulus block) and mean log perceptual phase durations were calculated, separately for each perceptual organization (the "integrated," "segregated," and "both" percepts), and condition. The "neither" responses were not analyzed as they appeared in less than 1.4% of the stimulus block time in any of the conditions.

Repeated measures analyses of variance (Three-Way ANOVAs) were carried out separately for each of six variables (proportion and mean phase duration, separately for the "integrated," the "segregated," and the "both" percept) with the structure: Congruency [*Congruent* vs. *Incongruent*] × Motion [*Stationary* vs. *Moving*] × Co-location [*Identical* vs. *Joint* vs. *Separate*]. Where applicable, degrees of freedom were adjusted with the Greenhouse-Geisser correction factor (ε). These and the partial η^2 effect sizes are reported for all significant effects. *Post-hoc* comparisons for the 3-level factor Co-location were performed using Tukey's HSD tests. Interaction effects between Co-location and Congruency on the proportion of the segregated and the integrated percepts were followed up by separately testing the effects of Co-location for the two levels of Congruency with repeated measures ANOVAs. All analyses were carried out at the 95% confidence level.

The level of performance in the secondary task was assessed by the blink count error (the difference between the reported and the actual blink count) established separately for each condition and listener. Negative values correspond to underestimation and positive values to overestimation of the number of blinks. Counting accuracy was estimated separately for each listener and condition by dividing the unsigned blink error value by the number of the blinks in the condition. The resulting values were analyzed by repeated measures analyses of variance (Three-Way ANOVAs) with the same factors as used for the main perceptual variables (Congruency [*Congruent* vs. *Incongruent*] × Motion [*Stationary* vs. *Moving*] × Co-location [*Identical* vs. *Joint* vs. *Separate*]).

RESULTS

Figure 2 presents the group-average percentage of signed blink count errors (relative to the number of actual blinks), separately for each condition. The Three-Way ANOVA of the accuracy (unsigned) values yielded no significant main effect of Congruency. Motion had a significant main effect [$F_{(1, 20)} = 7.46$, $p < 0.05$, $\eta^2 = 0.272$]. This was due to the lower number of errors made in the *Stationary* compared with the *Moving* conditions. No other main effects or interactions were significant.

Figures 3, **4**, respectively show the group-averaged proportions and mean phase durations of the perceptual alternatives reported by the listeners ("integrated," "segregated," and "both") for all 12 stimulus conditions. Results of the ANOVA analyses are listed in **Table 1**.

The analyses yielded no significant main effect of Congruency on any of the measured variables. Co-location had a significant main effect on the proportion of the "segregated" and

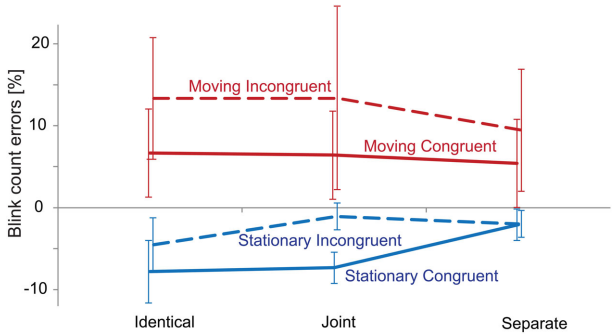


FIGURE 2 | Performance in the visual (blink counting) task. Group-averaged ($N = 21$) percentage of blink count errors obtained in the *Identical, Joint* and *Separate* Co-locations, separately for the *Stationary* (blue) and *Moving* (red) × *Congruent* (continuous line) and *Incongruent* (dashed line) conditions. Negative values correspond to underestimation and positive values to overestimation of the number of blinks. Error bars show the standard error of means.

"integrated" phases [$F_{(2, 40)} = 15.69$, $p < 0.001$, $\varepsilon = 0.624$, $\eta^2 = 0.440$ for the "segregated" and $F_{(2, 40)} = 7.69$, $p < 0.01$, $\varepsilon = 0.621$, $\eta^2 = 0.278$ for the "integrated" percept]. A significant interaction between Congruency and Co-location was obtained for the proportion of the "integrated" and "segregated" percepts [$F_{(2, 40)} = 3.70$, $p < 0.05$, $\varepsilon = 0.954$, $\eta^2 = 0.156$ for the "integrated" and $F_{(2, 40)} = 3.57$, $p < 0.05$, $\varepsilon = 0.973$, $\eta^2 = 0.152$ for the "segregated" percept]. In *post-hoc* ANOVAs the effects of Co-location were separately tested for the two levels of Congruency. In the *Congruent* condition, a significant effect of Co-location was obtained for the proportion of the "segregated" [$F_{(2, 40)} = 6.52$, $p < 0.01$, $\varepsilon = 0.806$, $\eta^2 = 0.246$] but not for the "integrated" percept. In the *Incongruent* condition, Co-location had significant effects on the proportion of both the "segregated" and the "integrated" percept [$F_{(2, 40)} = 18.32$, $p < 0.001$, $\varepsilon = 0.688$, $\eta^2 = 0.478$ and $F_{(2, 40)} = 10.06$, $p < 0.01$, $\varepsilon = 0.698$, $\eta^2 = 0.335$, respectively]. **Figures 5A,B** shows that the interaction stems from the effects of Co-location being more pronounced in the *Incongruent* than in the *Congruent* condition: (1) the proportion of the "segregated" responses was higher for the *Separate* compared with the *Identical* and the *Joint* trajectories, this difference being larger with the *Incongruent* visual cues and (2) the proportion of the "integrated" percept was influenced by spatial separation of the sound sources only with the *Incongruent* visual cues: with *Separate* trajectories the proportion of the "integrated" percept was lower compared with the *Identical* and the *Joint* trajectories.

Figure 5C illustrates the main effects of Co-location on the log-mean phase duration of all percepts. This main effect was significant only for the "segregated" phases [$F_{(2, 40)} = 12.79$, $p < 0.001$, $\varepsilon = 0.711$, $\eta^2 = 0.390$]. According to the *post-hoc* tests, the duration of the "segregated" percept was longer for the *Separate* compared with the *Identical* and the *Joint* trajectories (both $p < 0.001$).

Motion (see **Figure 5D**) had a significant main effect on the log-mean phase durations of all three percepts [$F_{(1, 20)} = 26.28$, $p < 0.001$, $\varepsilon = 1$, $\eta^2 = 0.568$ for the "segregated,"

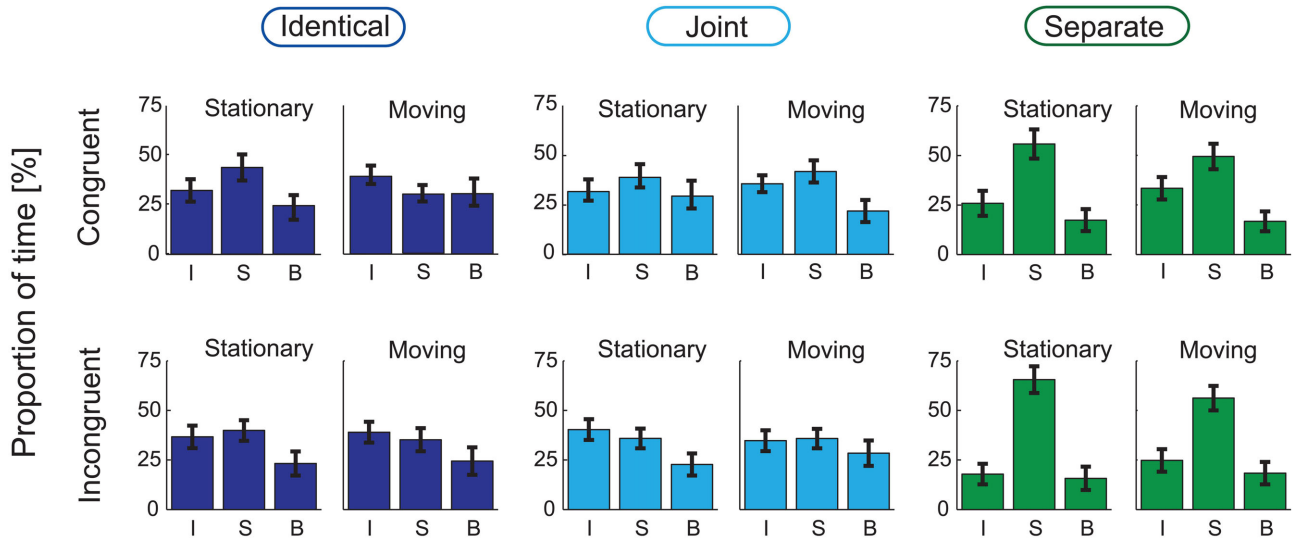


FIGURE 3 | Group-averaged ($N = 21$) percentages of time of the three response alternatives (I: "Integrated," S: "Segregated," B: "Both") are shown for the *Identical* (left panel), *Joint* (middle), and *Separate* (right) Co-locations, separately for the *Stationary* (left, separately within the three panels) and *Moving* (right) conditions and for the *Congruent* (top row) and *Incongruent* (bottom) conditions. Error bars show the standard errors of mean.

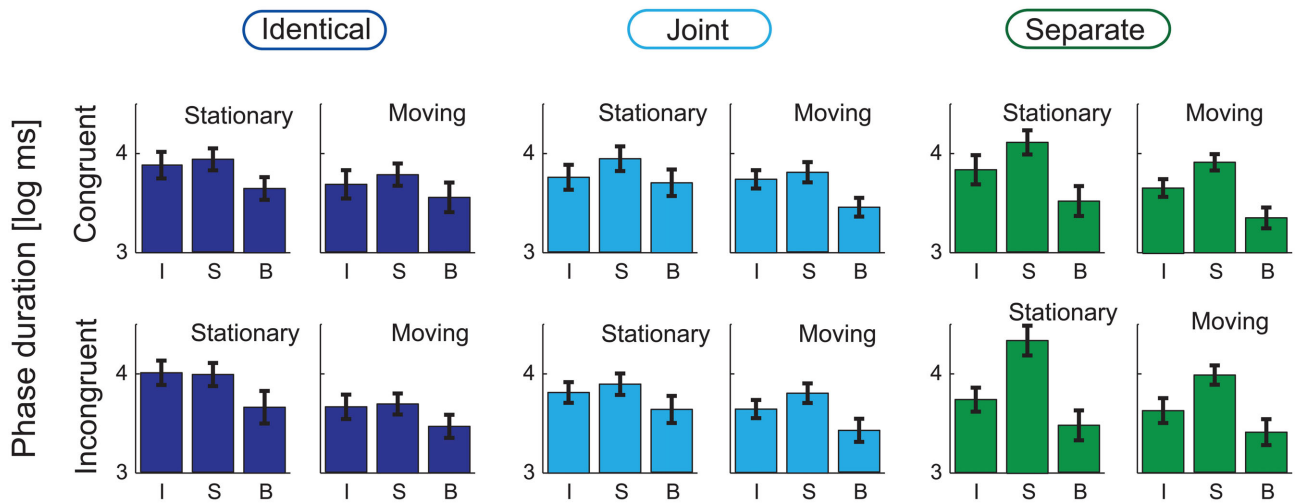


FIGURE 4 | Group-averaged ($N = 21$) average phase durations of the three response alternatives (I: "Integrated," S: "Segregated," B: "Both") are shown for the *Identical* (left panel), *Joint* (middle), and *Separate* (right) Co-locations, separately for the *Stationary* (left, separately within the three panels) and *Moving* (right) conditions and for the *Congruent* (top row) and *Incongruent* (bottom) conditions. Error bars show the standard errors of mean.

$F_{(1, 20)} = 19.27$, $p < 0.001$, $\varepsilon = 1$, $\eta^2 = 0.491$ for the "integrated," and $F_{(1, 20)} = 12.87$, $p < 0.01$, $\varepsilon = 1$, $\eta^2 = 0.392$ for the "both" responses]. For all three percepts, this was due to shorter phase durations when the sound source was moving compared with that obtained for stationary sound sources.

DISCUSSION

We studied the effects of congruent visual spatial information on the perception of stationary and moving sound sources in a bistable auditory stimulus configuration, the auditory streaming paradigm. We expected that the audio-visual binding established during the training should result in participants experiencing more "veridical" percepts when the visual display was congruent with the spatial location/trajectory of the sound sources as compared to when they received incongruent visual information. "Veridical" entails the perception of two auditory streams when the original sound sources were separate and the perception of a single auditory stream when the original sound sources were identical. Specifically, we expected to find (1) increased proportions of segregated responses in *Congruent/Separate* as compared

Table 1 | *F* values yielded by the ANOVAs for the overall phase proportions and phase durations for the factors of Congruency (*Congruent* or *Incongruent*), Motion (*Stationary* or *Moving*), and Co-location (*Identical, Joint* or *Separate*).

ANOVA factors and interactions (*df*)	IntProp	SegProp	BothProp	IntDur	SegDur	BothDur
Congruency (1, 20)	0.11	0.45	0.31	0.07	1.48	0.38
Motion (1, 20)	0.81	1.45	0.71	19.27***	26.28***	12.87**
Co-location (2, 40)	7.69**	15.69***	2.81	1.72	12.79***	2.20
Congruency × motion (1, 20)	1.56	0.05	1.44	0.73	1.24	0.02
Congruency × co-location (2, 40)	3.70*	3.57*	0.46	0.65	2.46	0.20
Motion × co-location (2, 40)	1.28	2.55	0.61	1.39	2.60	0.82
Congruency × motion × co-location (2, 40)	0.38	0.81	1.51	0.68	0.64	0.62

*The measures compared are denoted as following: IntProp, proportion of "integrated" phases; SegProp, proportion of "segregated" phases; BothProp, proportion of "both" phases; IntDur, average duration of all "integrated" phases; SegDur, average duration of all "segregated" phases; BothDur, average duration of all "both" phases. Degrees of freedom (df) are given in the row titles. Significance levels (p) are indicated by asterisks. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.*

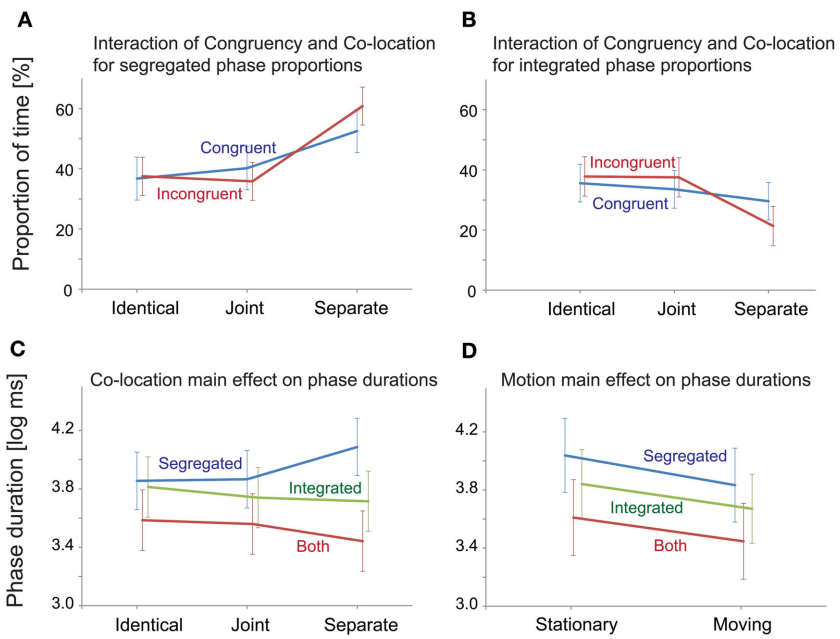


FIGURE 5 | Significant ANOVA effects for "Integrated," "Segregated," and "Both" response alternatives. (A) Spatial separation increased the proportion of the "segregated" responses, this effect being larger with the *Incongruent* visual cues. **(B)** Spatial separation decreased the proportion of the "integrated" percepts, this effect being significant only with the *Incongruent* visual cues. **(C)** Spatial separation influenced the average phase durations for all percepts, this effect being significant only for the "segregated" phases. **(D)** Motion had a significant main effect on the log-mean phase durations of all percepts. In all panels, error bars show the standard errors of mean.

to *Incongruent/Separate* conditions and (2) increased proportions of integrated responses in *Congruent/Identical* as compared to *Incongruent/Identical* conditions. However, we found no clear effects of the congruency of visual information on auditory segregation or integration. Audio-visual congruency only appeared to significantly interact with the Co-location factor, the effect indicating stronger influence of spatial separation of the sound sources on perceptual organization in the *Incongruent* than in the *Congruent* conditions. Increased amounts of spatial separation between the sound sources resulted in a lower proportion of the "integrated" percept, the effect being significant only with incongruent but not with congruent visual cues. The proportion of the "segregated" percept significantly increased with increasing spatial separation between the sound sources both with congruent and incongruent visual cues.

One explanation for this somewhat unexpected result is that participants could have had difficulties in binding the auditory and visual cues of the abstract objects employed in the present experiment. However, participants were preselected on the basis of a synchrony test conducted after the training procedure checking whether they were able to bind the auditory and visual stimuli into a single multimodal perceptual object. Further, the conditions known to be important for audio-visual cuing becoming stronger than unimodal ones are that the (1) auditory and visual stimuli should be presented from more-or-less the same direction (Ho et al., 2007, 2009) and that (2) a high perceptual load

should be used (Spence and Santangelo, 2009; Spence, 2010). In the current experiment, visual and auditory stimuli were presented from roughly the same spatial direction (at least on the horizontal plane; when the visual stimuli were on screen). The perceptual load of the procedure was rather high inasmuch as the participants also performed a secondary task. Therefore, the failure to find any benefit of congruent audio-visual stimulation over incongruent stimulation in veridical perception may rather be explained by attentional effects (see below) than by the absence of cross-modal integration. However, as the main screening procedure did not directly test audiovisual integration (only asking about the subjective experience of binding together the sounds and the visual objects), this possibility cannot be completely ruled out. Evidence suggesting that the audiovisual binding could be attention-dependent was obtained by Van Ee et al. (2009). In our experiments, one possibility is that in the *Incongruent* conditions, the participants' attention was drawn to the sounds rather than to the visual objects. This would explain why the influence of spatial separation of the sound sources on stream segregation was stronger with incongruent than with congruent visual cues. That is, the congruency between auditory and visual stimuli could take away attentional resources from evaluating spatial separation as a cue for auditory stream segregation. However, the performance in the visual task, which required participants to focus their attention on the visual objects, was not affected by audio-visual congruency. This implies that attentional effects cannot fully explain the absence of the benefit of congruent over incongruent audio-visual stimulation.

Alternatively, it is possible that the processes of auditory stream segregation do not utilize visual information. Existing data suggest that sensory integration is likely occurring at multiple processing stages (Kayser et al., 2012). Early interactions between auditory and visual modalities have been demonstrated in a number of studies (Giard and Peronnet, 1999; Fort et al., 2002; Besle et al., 2005; Cook and Van Valkenburg, 2009). Furthermore, visual effects on auditory object formation have been demonstrated in a study utilizing a bistable auditory perceptual phenomenon similar to the one employed in the current study (Rahne et al., 2007). Rahne et al. (2007) found that perceptual organization of the ambiguous auditory input was significantly influenced by synchronized presentation of visual stimuli compatible either with the within-stream (segregated) or across-stream (integrated) sound pattern. The mismatch negativity component (MMN) was elicited only when the visual input biased the sounds toward the two-stream organization, suggesting that the underlying audio-visual interaction occurred prior to the formation of the memory representation involved in the MMN deviance detection process. Similar effects have been obtained for a perceptually stable sound sequence (Rahne and Böckmann-Barthel, 2009). However in these studies only non-spatial audio-visual cues were employed. Therefore, it is possible that in contrast to temporal audio-visual congruency, spatial congruency exerts no or only weak influence on the auditory perceptual organization. In support of this conclusion, some results suggest that separate competing object representations may exist at multiple levels of processing (e.g., Blake and Logothetis, 2002; Tong et al., 2006). The question of whether audio-visual bistable perception is mediated by distributed intramodal mechanisms or is governed by a supramodal central mechanism was tested by Hupé et al. (2008) using sounds evoking auditory streaming and either visual plaids or visual stimuli evoking apparent motion. The auditory and visual stimuli had weak or strong cross-modal congruency, the latter being achieved by introducing spatial and temporal coincidence between the two modalities. The visual stimuli had no influence on the auditory switching statistics, indicating that bistability can co-occur independently in different sensory modalities. This interpretation is supported by the lack of a congruency effect on performance in the present visual task. Further, we have already noted above that the interaction between the effects of cross-modal congruency and spatial separation on sound organization probably stems from attentional differences between the congruent and incongruent conditions. Thus, the current results support models of modality-specific competition for perceptual decision and awareness.

One may wonder why the lack of advantage of congruent over incongruent audio-visual stimulation in auditory stream segregation does not apparently affect veridical perception in everyday situations. However, multistable auditory stimulus configurations (i.e., stimuli that are deprived of disambiguating cues) are very rare in everyday environments. Thus in most cases, visual information is not required to decide between alternative sound organizations. Therefore it may not have been sufficiently advantageous for the brain to utilize visual information for this purpose. This is not the case for example in identifying speech sounds, and indeed there are several results showing that e.g., lip movements affect the identification of speech sounds (see e.g., the McGurk effect; McGurk and MacDonald, 1976). Note that the previous suggestion that audio-visual integration follows auditory stream segregation does not mean that audio-visual integration does not take place at all. Further, when it comes to localizing objects, the dominance of visual over auditory information is obvious in most cases (see e.g., the ventriloquist effect; Howard and Templeton, 1966; for a review, cf. Recanzone, 2009). Thus in everyday life, we are not especially sensitive to spatial incongruence between auditory and visual stimuli. This leads us to suggest that the task requiring participants to continuously integrate between the auditory and visual objects took up more attentional resources than when they realized that the two were fully incongruent and concentrated primarily on the sounds.

We found a clear effect of sound source spatial separation on perceptual organization, which is consistent with the results of our earlier unimodal study (Bőhm et al., 2013) that also used moving sounds. It is also consistent with previous studies testing the effects of spatial separation between sound sources on auditory stream segregation (Cusack et al., 2004; Denham et al., 2010; Szalárdy et al., 2013). Sounds originating from different directions are more likely to be perceived as separate objects than sounds emitted from roughly the same direction.

Bőhm et al. (2013) found that auditory motion decreased the average perceptual phase durations of both the "integrated" and the "segregated" percepts compared to stationary sound sources without affecting the balance between the proportions of these perceptual organizations. The current results corroborated these findings: the mean durations of all percepts were lower

for moving sound sources compared to stationary ones. Shorter average phase durations correspond to faster switching between the alternative sound organizations. Thus it appears that motion exerted a rather strong destabilizing effect on the perceptual organization. The increased rate of perceptual switching may be due to the auditory system continuously re-evaluating the spatial separation between the moving sound sources. The significant effect of motion obtained in the visual task could also indicate frequent re-evaluation of the stimulus position: listeners were less accurate in the *Moving* compared with the *Stationary* conditions. Besides destabilizing, motion had no further effect on auditory perceptual organization. In that regard, the current results support the main conclusion of Bőhm et al. (2013) who suggested that cues based on the Gestalt notion of common fate, at least those stemming from moving sound sources, are not effective in auditory stream segregation.

ACKNOWLEDGMENTS

This work was supported by the European Commission's Seventh Framework Programme (ICT-FP7-231168), the Hungarian Academy of Sciences (Lendület project to István Winkler, LP2012-36/2012), the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG Cluster of Excellence 1077 "Hearing4all"), the German Academic Exchange Service (Deutscher Akademischer Austauschdienst, DAAD, Project 56265741), the Russian Foundation for Fundamental Research (Project 11-04-00008-a), the Russian Academy of Sciences—Hungarian Academy of Sciences (Exchange Cooperation Agreement 2010-2012), and the Hungarian Scholarship Board (Magyar Ösztöndíj Bizottság, MÖB, Project 39589). The authors thank Dalibor Andrijevich and Péter Szombathy for their valuable technical assistance and Zsuzsanna D'Albini for collecting the perceptual data.

REFERENCES

Bendixen, A., Denham, S. L., Gyimesi, K., and Winkler, I. (2010). Regular patterns stabilize auditory streams. *J. Acoust. Soc. Am.* 128, 3658–3666. doi: 10.1121/1.3500695

Besle, J., Fort, A., and Giard, M. H. (2005). Is the auditory sensory memory sensitive to visual information? *Exp. Brain Res.* 166, 337–344. doi: 10.1007/s00221-005-2375-x

Blake, R., and Logothetis, N. K. (2002). Visual competition. *Nat. Rev. Neurosci.* 3, 13–21. doi: 10.1038/nrn701

Bőhm, T. M., Shestopalova, L., Bendixen, A., Andreou, A. G., Georgiou, J., Garreau, G., et al. (2013). The role of perceived source location in auditory stream segregation: separation affects sound organization, common fate does not. *Learn. Percept.* 5(Suppl. 2), 55–72. doi: 10.1556/LP.5.2013.Suppl2.5

Bregman, A. S. (1990). *Auditory Scene Analysis*. Cambridge, Massachusetts: MIT Press.

Conrad, V., Bartels, A., Kleiner, M. and Noppeney, U. (2010). Audiovisual interactions in binocular rivalry. *J. Vis.* 10, 1–15. doi: 10.1167/10.10.27

Cook, L. A., and Van Valkenburg, D. L. (2009). Audio-visual organization and the temporal ventriloquism effect between grouped sequences: evidence that unimodal grouping precedes cross-modal integration. *Perception* 38, 1220–1233.

Culling, J. F., and Summerfield, Q. (1995). Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay. *J. Acoust. Soc. Am.* 98, 785–797. doi: 10.1121/1.413571

Cusack, R., Deeks, J., Aikman, G., and Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 643–656. doi: 10.1037/0096-1523.30.4.643

Denham, S., Bőhm, T. M., Bendixen, A., Szalárdy, O., Kocsis, Z., and Winkler, I. (2014). Stable individual characteristics in the perception of multiple embedded patterns in multistable auditory stimuli. *Front. Neurosci.* 8:25. doi: 10.3389/fnins.2014.00025

Denham, S. L., Gyimesi, K., Stefanics, G., and Winkler, I. (2010). "Stability of perceptual organization in auditory streaming," in *The Neurophysiological Bases of Auditory Perception*, eds E. A. Lopez-Poveda, A. R. Palmer, and R. Meddis (New York, NY: Springer), 477–487. doi: 10.1007/978-1-4419-5686-6_44

Denham, S. L., Gyimesi, K., Stefanics, G., and Winkler, I. (2013). Perceptual bistability in auditory streaming: how much do stimulus features matter? *Learn. Percept.* 5(Suppl. 2), 73–100. doi: 10.1556/LP.5.2013.Suppl2.6

Denham, S. L., and Winkler, I. (2006). The role of predictive models in the formation of auditory streams. *J. Physiol.* 100, 154–170. doi: 10.1016/j.jphysparis.2006.09.012

Eramudugolla, R., Henderson, R., and Mattingley, J. B. (2011). Effects of audio–visual integration on the detection of masked speech and non-speech sounds. *Brain Cogn.* 75, 60–66. doi: 10.1016/j.bandc.2010.09.005

Fort, A., Delpuech, C., Pernier, J., and Giard, M. H. (2002). Dynamics of cortico-subcortical cross-modal operations involved in audio-visual object detection in humans. *Cereb. Cortex* 12, 1031–1039. doi: 10.1093/cercor/12.10.1031

Giard, M. H., and Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *J. Cogn. Neurosci.* 11, 473–490. doi: 10.1162/089892999563544

Grant, K. W., and Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am.* 108, 1197–1208. doi: 10.1121/1.1288668

Grimault, N., Bacon, S. P., and Micheyl, C. (2002). Auditory stream segregation on the basis of amplitude-modulation rate. *J. Acoust. Soc. Am.* 111, 1340–1348. doi: 10.1121/1.1452740

Gutschalk, A., Micheyl, C., Melcher, J. R., Rupp, A., Scherg, M., and Oxenham, A. J. (2005). Neuromagnetic correlates of streaming in human auditory cortex. *J. Neurosci.* 25, 5382–5388. doi: 10.1523/JNEUROSCI.0347-05.2005

Hill, K. T., Bishop, C. W., and Miller, L. M. (2012). Auditory grouping mechanisms reflect a sound's relative position in a sequence. *Front. Hum. Neurosci.* 6:158. doi: 10.3389/fnhum.2012.00158

Ho, C., Reed, N., and Spence, C. (2007). Multisensory in-car warning signals for collision avoidance. *Hum. Factors* 49, 1107–1114. doi: 10.1518/001872007X249965

Ho, C., Santangelo, V., and Spence, C. (2009). Multisensory warning signals: when spatial location matters. *Exp.Brain Res.* 195, 261–272. doi: 10.1007/s00221-009-1778-5

Howard, I. P., and Templeton, W. B. (1966). *Human Spatial Orientation*. London: Wiley.

Hupé, J.-M., Joffo L.-M., and Pressnitzer, D. (2008). Bistability for audiovisual stimuli: perceptual decision is modality specific. *J. Vis.* 8, 1–15. doi: 10.1167/8.7.1

Kang, M. S., and Blake, R. (2005). Perceptual synergy between seeing and hearing revealed during binocular rivalry. *Psichologija* 32, 7–15.

Kayser, C., Petkov, C. I., Remedios, R., and Logothetis, N. K. (2012). "Multisensory influences on auditory processing," in *The Neural Bases of Multisensory processes*, ed M. M. Murray and M. T.Wallace (Boca Raton, FL: CRC press).

Koelewijn, T., Bronkhorst, A., and Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: a review of audiovisual studies. *Acta Psychologica* 134, 372–384. doi: 10.1016/j.actpsy.2010.03.010

Kondo, H. M., and Kashino, M. (2009). Involvement of the thalamocortical loop in the spontaneous switching of percepts in auditory streaming. *J. Neurosci.* 29, 12695–12701. doi: 10.1523/JNEUROSCI.1549-09.2009

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Moore, B. C. J., and Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acust. United Acust.* 88, 320–333.

Moreno-Bote, R., Shpiro, A., Rinzel, J., and Rubin, N. (2010). Alternation rate in perceptual bistability is maximal at and symmetric around equi-dominance. *J. Vis.* 10, 1. doi: 10.1167/10.11.1

Munhall, K. G., ten Hove, M. W., Brammer, M., and Paré, M. (2009). Audiovisual integration of speech in a bistable illusion. *Curr. Biol.* 19, 735–739. doi: 10.1016/j.cub.2009.03.019

Pressnitzer, D., and Hupé, J. M. (2006). Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Curr. Biol.* 16, 1351–1357. doi: 10.1016/j.cub.2006.05.054

Rahne, T., and Böckmann-Barthel, M. (2009). Visual cues release the temporal coherence of auditory objects in auditory scene analysis. *Brain Res.* 1300, 125–134. doi: 10.1016/j.brainres.2009.08.086

Rahne, T., Böckmann, M., von Specht, H., and Sussman, E. S. (2007). Visual cues can modulate integration and segregation of objects in auditory scene analysis. *Brain Res.* 1144, 127–135. doi: 10.1016/j.brainres.2007.01.074

Recanzone, G. H. (2009). Interactions of auditory and visual stimuli in space and time. *Hear. Res.* 258, 89–99. doi: 10.1016/j.heares.2009.04.009

Roberts, B., Glasberg, B. R., and Moore, B. C. (2002). Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. *J. Acoust. Soc. Am*. 112, 2074–2085. doi: 10.1121/1.1508784

Santangelo, V., Van der Lubbe, R. H., Belardinelli, M. O., and Postma, A. (2006). Spatial attention triggered by unimodal, crossmodal, and bimodal exogenous cues: a comparison of reflexive orienting mechanisms. *Exp. Brain Res.* 173, 40–48. doi: 10.1007/s00221-006-0361-6

Santangelo, V., Van der Lubbe, R. H., Belardinelli, M. O., and Postma, A. (2008). Multisensory integration affects ERP components elicited by exogenous cues. *Exp. Brain Res.* 185, 269–277. doi: 10.1007/s00221-007-1151-5

Schwartz J.-L., Grimault N., Hupé J.-M., Moore, B. C. J., and Pressnitzer D. (2012). Multistability in perception: binding sensory modalities, an overview. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 896–905. doi: 10.1098/rstb.2011.0254

Shinozaki, N., Yabe, H., Sato, Y., Hiruma, T., Sutoh, T., Matsuoka, T., et al. (2003). Spectrotemporal window of integration of auditory information in the human brain. *Cogn. Brain Res*. 17, 563–571. doi: 10.1016/S0926-6410(03)00170-8

Snyder, J. S., and Alain, C. (2007). Toward a neurophysiological theory of auditory stream segregation. *Psychol. Bull.* 133, 780–799. doi: 10.1037/0033-2909.133.5.780

Soto-Faraco, S., Kingstone, A., and Spence, C. (2003). Multisensory contributions to the perception of motion. *Neuropsychologia* 41, 1847–1862. doi: 10.1016/S0028-3932(03)00185-4

Soto-Faraco, S., Spence, C., and Kingstone, A. (2004). Crossmodal dynamic capture: congruency effects in the perception of motion across sensory modalities. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 330–345. doi: 10.1037/0096-1523.30.2.330

Soto-Faraco, S., and Väljamäe, A. (2012). "Multisensory interactions during motion perception," in *The Neural Bases of Multisensory processes*, ed M. M. Murray and M. T. Wallace (Boca Raton, FL: CRC press).

Spence, C. (2010). Crossmodal spatial attention. *Ann. N.Y. Acad. Sci.* 1192, 182–200. doi: 10.1111/j.1749-6632.2010.05440.x

Spence, C., and Driver, J. (2000). Attracting attention to the illusory location of a sound: reflexive crossmodal orienting and ventriloquism. *Neuroreport* 11, 2057–2061. doi: 10.1097/00001756-200006260-00049

Spence, C., McDonald, J., and Driver, J. (2004). "Exogenous spatial-cuing studies of human cross-modal attention and multisensory integration," in *Crossmodal Space and Crossmodal Attention*, ed C. Spence and J. Driver (Oxford: Oxford university press), 277-320.

Spence, C., and Santangelo, V. (2009). Capturing spatial attention with multisensory cues: a review. *Hear. Res.* 258, 134–142. doi: 10.1016/j.heares.2009.04.015

Stein, B. E., Stanford, T. R., Wallace, J. W. V., and Jiang, W. (2004). "Crossmodal spatial interactions in subcortical and cortical circuits," in *Crossmodal Space and Crossmodal Attention*, ed C. Spence and J. Driver (Oxford: Oxford university press), 25-50.

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am*. 26, 212–215. doi: 10.1121/1.1907309

Szalárdy, O., Bendixen, A., Bőhm, T. M., Davies, L. A., Denham, S. L., and Winkler, I. (2014). The effects of rhythm and melody on auditory stream segregation. *J. Acoust. Soc. Am.* 135, 1392–1405. doi: 10.1121/1.4865196

Szalárdy, O., Bendixen, A., Tóth, D., Denham, S. L., and Winkler, I. (2013). Modulation frequency difference acts as a primitive cue for auditory stream segregation. *Learn. Percept.* 5(Suppl. 2), 149–161. doi: 10.1556/LP.5.2013.Suppl2.9

Tong, F., Meng, M., and Blake, R. (2006). Neural bases of binocular rivalry. *Trends Cogn. Sci.* 10, 502–511. doi: 10.1016/j.tics.2006.09.003

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., and Theeuwes, J. (2008). Audiovisual events capture attention: evidence from temporal order judgments. *J. Vis.* 8, 1–10. doi: 10.1167/8.5.2

Van Ee, R., van Boxtel, J. J. A., Parker, A. L., and Alais, D. (2009). Multisensory congruency as a mechanism for attentional control over perceptual selection. *J. Neurosci.* 29, 11641–11649. doi: 10.1523/JNEUROSCI.0873-09.2009

Van Noorden, L. P. A. S. (1975). *Temporal Coherence in the Perception of Tone Sequences*. PhD, Eindhoven University of Technology, Leiden.

Vliegen, J., and Oxenham, A. J. (1999). Sequential stream segregation in the absence of spectral cues. *J. Acoust. Soc. Am*. 105, 339–346. doi: 10.1121/1.424503

Vroomen, J., Bertelson, P., and de Gelder, B. (2001). Directing spatial attention towards the illusory location of a ventriloquized sound. *Acta Psychologica* 108, 21–33. doi: 10.1016/S0001-6918(00)00068-8

Ward, L. M. (1994). Supramodal and modality-specific mechanisms for stimulus-driven shifts of auditory and visual attention. *Can. J. Exp. Psychol*. 48, 242–259. doi: 10.1037/1196-1961.48.2.242

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

Winkler, I., Denham, S., Mill R., Bőhm, T. M., and Bendixen A. (2012). Multistability in auditory stream segregation: a predictive coding view. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 1001–1012. doi: 10.1098/rstb.2011.0359

Wright, R. D., and Ward, L. M. (2008). *Orienting of Attention*. New York, NY: Oxford University Press.

Yabe, H., Winkler, I., Czigler, I., Koyama, S., Kakigi, R., Sutoh, T., et al. (2001). Organizing sound sequences in the human brain: the interplay of auditory streaming and temporal integration. *Brain Res.* 897, 222–227. doi: 10.1016/S0006-8993(01)02224-7

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

*Received: 28 January 2014; accepted: 19 March 2014; published online: 07 April 2014. Citation: Shestopalova L, Bőhm TM, Bendixen A, Andreou AG, Georgiou J, Garreau G, Hajdu B, Denham SL and Winkler I (2014) Do audio-visual motion cues promote segregation of auditory streams? Front. Neurosci. **8**:64. doi: 10.3389/fnins.2014.00064 This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.*

frontiers in
NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 29 April 2014
doi: 10.3389/fnins.2014.00093

The effect of stimulus context on the buildup to stream segregation

***Jonathan Sussman-Fort*[1]* and *Elyse Sussman*[1,2]**

[1] Department of Neuroscience, Albert Einstein College of Medicine, Yeshiva University, Bronx, NY, USA
[2] Department of Otorhinolaryngology-Head and Neck Surgery, Albert Einstein College of Medicine, Yeshiva University, Bronx, NY, USA

Edited by:
Susann Deike, Leibniz Institute for Neurobiology, Germany

Reviewed by:
Joel Snyder, University of Nevada Las Vegas, USA
Andrew R. Dykstra, University of Heidelberg, Germany

****Correspondence:***
Jonathan Sussman-Fort, Department of Neuroscience, Albert Einstein College of Medicine, Yeshiva University, 1300 Morris Park Avenue, Kennedy Center, Room 210, Bronx, NY 10461, USA
e-mail: jonathan.sussman-fort@med.einstein.yu.edu

Stream segregation is the process by which the auditory system disentangles the mixture of sound inputs into discrete sources that cohere across time. The length of time required for this to occur is termed the "buildup" period. In the current study, we used the buildup period as an index of how quickly sounds are segregated into constituent parts. Specifically, we tested the hypothesis that stimulus context impacts the timing of the buildup and, therefore, affects when stream segregation is detected. To measure the timing of the buildup we recorded the Mismatch Negativity component (MMN) of event-related brain potentials (ERPs), during passive listening, to determine when the streams were neurophysiologically segregated. In each condition, a pattern of repeating low (L) and high (H) tones (L-L-H) was presented in trains of stimuli separated by silence, with the H tones forming a simple intensity oddball paradigm and the L tones serving as distractors. To determine the timing of the buildup, probe tones occurred in two positions of the trains, early (within the buildup period) and late (past the buildup period). The context was manipulated by presenting roving vs. non-roving frequencies across trains in two conditions. MMNs were elicited by intensity probe tones in the Non-Roving condition (early and late positions) and the Roving condition (late position only) indicating that neurophysiologic segregation occurred faster in the Non-Roving condition. This suggests a shorter buildup period when frequency was repeated from train to train. Overall, our results demonstrate that the dynamics of the environment influence the way in which the auditory system extracts regularities from the input. The results support the hypothesis that the buildup to segregation is highly dependent upon stimulus context and that the auditory system works to maintain a consistent representation of the environment when no new information suggests that reanalyzing the scene is necessary.

Keywords: auditory perception, mismatch negativity, stream segregation, event-related potentials, auditory scene analysis

INTRODUCTION

When entering a noisy scene, such as a crowded sports arena, the auditory system is faced with the "problem" of parsing out the mixture of sounds to their distinct sources, a process that has been termed "auditory scene analysis" (Bregman, 1990). Bregman (1978) hypothesized that the mixture of sounds entering the ears is initially perceived as integrated. Only after some time in which information about the sound characteristics accumulates would the mixture of sounds be perceived as distinct sound streams. This was supported by Anstis and Saida (1985) who found that there was a higher probability that listeners' perceived a sound mixture as segregated as time elapsed over a 30 s trial. The time that it takes to perceive an integrated mixture of sounds as segregated streams has been called the "buildup" period. In the current study, we used the buildup period as an index of how quickly sounds are segregated into constituent parts.

Most previous studies have tested the buildup to stream segregation by manipulating the frequency distance between two sets of sounds and measuring the time it takes to indicate that two streams are perceived from the onset of the sound sequence. Using this approach, studies demonstrated that the buildup to perceiving segregated streams occurred faster the larger the frequency distance between sounds or the more rapid the stimulus presentation rate (Bregman, 1978; Cusack et al., 2004; Micheyl et al., 2005; Snyder et al., 2008; Haywood and Roberts, 2010). However, in more natural listening scenes, sound features are often dynamically changing. We hypothesize that the extent of change in the sound input contributes to how quickly we adapt to the auditory scene. That is, the buildup period should be longer when the scene is more dynamic, in that more information would be needed to make a decision about whether a sequence is segregated or not. In this study, we test the influence of changing parameters on stream segregation by measuring the timing of the buildup to detection of stream segregation when the input is stable vs. when it is changing. In this first step of addressing our hypothesis, we took the influence of attention out of the equation, and measured the timing to stream segregation using a neurophysiologic index that we have used previously (Sussman et al., 2007).

There has been some debate about the role of attention in the formation of auditory streams (Jones et al., 1999; Carlyon et al., 2001; Cusack et al., 2004; Sussman et al., 2007) but certainly

there is evidence that attention modulates the perceptual decision about stream segregation (Jones et al., 1999; Carlyon et al., 2001; Cusack et al., 2004; Snyder et al., 2006). One issue with attention is whether the behavioral response is measuring the buildup, or the time to perceptual decision (Deike et al., 2012). Additionally, performing a task or attending to a subset of sounds may influence threshold levels for detecting segregation (Sussman et al., 1998a, 2005; Sussman, 2007; Sussman and Steinschneider, 2009). In the current study, we used an index of change detection to assess the timing of when sounds were neurophysiologically segregated without the influence of task performance. This allowed us to assess the influence of stimulus context on the timing of the buildup, if it exists, without attentional manipulation.

To test our hypothesis, we recorded event-related brain potentials (ERPs), and measured the mismatch negativity (MMN) component. The MMN is an appropriate tool because it can index sound change detection irrespective of the direction of attention (Sussman et al., 2003; Winkler et al., 2005). MMN is generated bilaterally within auditory cortices and the negative waveform is observed with a fronto-central scalp distribution that inverts in polarity at the mastoid electrodes (Vaughan and Ritter, 1970; Giard et al., 2014). MMN is elicited by detected infrequent deviant tones presented among a series of more frequently presented standard tones. The standard-to-deviant relationship can be set up by repetition of any tone feature (e.g., frequency, intensity, or duration). Thus, in the simplest case, the MMN component is the end result of a comparison process between the frequently presented standard tone and detection of a deviant feature. In the current study, we used intensity deviants to elicit MMN in an oddball stream that emerges only when the sounds are neurophysiologically segregated (Sussman et al., 2007; Sussman and Steinschneider, 2009). Thus, the ability to detect the intensity deviants (herein we will call them "probe tones" because they are not always detected as deviants), and the elicitation of the MMN in the current study, indicates neurophysiologic segregation of sounds (see Methods for details).

Sussman et al. (2007) found evidence that there was a buildup to stream segregation even when no task was performed with the sounds, when participants were watching a movie and reading closed-captions. In that study, trains of low (L) and high (H) tones were presented with probe tones randomly embedded during the supposed buildup period (in the beginning of the train of tones), as well as randomly placed after the buildup period was surpassed (in the later part of the train). The MMN was elicited by intensity deviants in an oddball stream by probe tones only in the later part of the train. The absence of MMN to probe tones embedded in the beginning of the trains indicated that the probes occurred before the L and H tones were neurophysiologically segregated (i.e., during the buildup period). No intensity regularity was detected during that time period. However, the study design could be considered as a factor in the results, in that the trains of tones roved in frequency across 10 different frequency levels. There was no repetition of tone frequency from train-to-train even though the semitone (ST) distance between trains was the same for all trains (8 ST). This roving frequency leaves open the question of whether the occurrence of each new train of an unpredicted frequency value may have signaled a new "event." That is, it is possible that the roving frequency may have initiated the analysis of stream segregation for each train, and precluded any carryover effects that may have occurred if each train had the same tone frequencies as the previous ones. If, on the other hand, one could tell at the beginning of the train that it was the same as the one that occurred before it (i.e., the same event occurred again after the silence), we hypothesized for the current study that the predictability or repetition could then affect the timing of the buildup (i.e., faster to segregation when each train could be anticipated). Thus, we predicted that the timing of segregation, as indexed by elicitation of the MMN to probe tones randomly occurring in the beginning of the tone trains, would be significantly altered by the stimulus context (whether the train-to-train frequency was repeated). To test this, we replicated a condition from Sussman et al. (2007) that roved trains by frequency (Roving condition) and compared it to a condition in which the tone frequency of the trains was repeated (Non-Roving condition). If the buildup could be modulated by stimulus context, whether the tone trains were repeated or roved in frequency, we predicted that it would be faster when there was repetition of tone frequency than when there was not, and thus elicitation of MMN to probe tones occurring in the beginning part of the tone trains, during the buildup period, should differ between the two conditions.

MATERIALS AND METHODS

SUBJECTS

Fifteen adults, 6 males (22–40 years) ($M = 30$ years, $SD = 4.8$ years) were paid for their participation in the study. All participants passed a hearing screen (20 dB HL at 500, 1000, 2000, and 4000 Hz), and had no reported history of neurological disorders. One participant's data were excluded due to excessive eye artifacts. The data from the remaining 14 participants were included in the analysis and are reported. All procedures were approved by the Albert Einstein College of Medicine Internal Review Board and were conducted in accordance with the Declaration of Helsinki. Informed consent was obtained from all participants after the procedures were explained to them.

STIMULI AND PROCEDURES

Complex tones (five harmonics above the fundamental frequency) were created using Adobe Audition 1.0 (Adobe Systems Inc., San Jose, CA). Tone duration was 30 ms (including a 5 ms rise/fall time). **Table 1** displays the five different tone pairs used in the study. Each column represents one tone pair and is labeled "A"–"E." Hereforth, individual tones will be referred to by their fundamental frequency shown in **Table 1**. Tones were presented bilaterally via insert earphones (E-A-RTone 3A; Indianapolis, IN) using Neuroscan STIM hardware and software (Compumedics Inc., Charlotte, NC). Sounds were calibrated using a Bruel and Kjaer 2209 (Denmark) impulse precision sound pressure level meter with an artificial ear by presenting each complex tone to the meter and adjusting for any difference in the actual sound level measured.

Figure 1 displays the stimulus paradigm. Tones were presented in blocks of 50 trains. Each train consisted of 36 tones, with 24 lower frequency tones (L) and 12 higher frequency tones (H) presented in a repeated three-tone sequence of L-L-H (**Figure 1A**).

Stimulus onset asynchrony (SOA) was 70 ms, with an intertrain interval (ITI) of 3.75 s (**Figure 1B**). The Δf of each frequency pair was eight semitones (ST). This frequency distance was chosen because it has previously been shown to induce streaming in passive listening conditions (Sussman et al., 2007; Sussman and Steinschneider, 2009). Tone intensity was used to elicit the MMN component within the high stream oddball sequence (**Figure 1A**), which occurs only when the tones neurophysiologically segregate (Sussman et al., 2001). To do this, tone intensity of the H tones was 59 dB SPL, with randomly and infrequently (17%) occurring "probe" tones that had a louder intensity value (68 dB SPL) (**Figure 1A**). The L tones were randomly assigned with four different stimulus intensities (56, 62, 65, and 71 dB SPL) that spanned above and below the intensity values of the H tones. This was done so that the oddball standard and deviant intensity tones had neither the loudest or softest intensity values within the overall tone train. Therefore, only when the tones neurophysiologically segregated would the within-stream oddball intensity relationship be detected, and have the possibility to elicit MMN. In contrast, if the tones did not neurophysiologically segregate,

Table 1 | Fundamental frequencies of the tones.

Tone pair	A	B	C	D	E
Low (Hz)	329.83	440.00	587.33	783.99	1046.50
High (Hz)	523.35	698.25	932.33	1244.50	1661.20

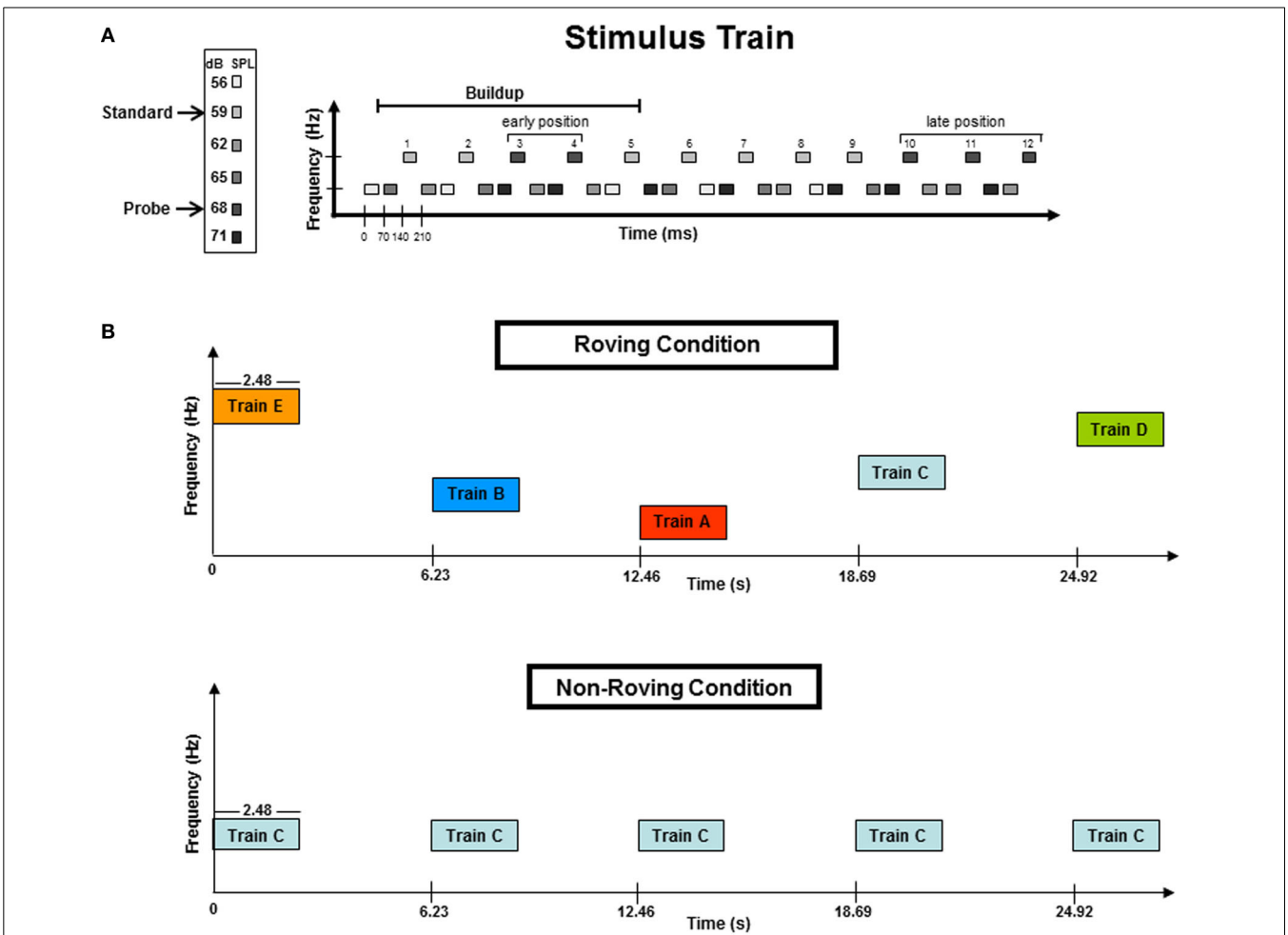


FIGURE 1 | (A) A stimulus train consisted of 36 tones: 24 lower (L) frequency tones and 12 higher (H) frequency tones presented in a three-tone repeating sequence of L-L-H. The frequency separation between the H and L tones was eight semitones (ST). The H tones formed a simple intensity oddball with early position probe tones randomized across 3rd or 4th tones in the train. The buildup period, designated by previous studies, is delineated with a horizontal line. Late position probe tones were randomized in the 10th, 11th, or 12th tones of the trains. The late position tones occurred outside (after) the buildup interval. Frequency (Hz) is represented on the vertical axis while the horizontal axis represents time (ms). **(B)** Schematic of the Roving and Non-Roving conditions. The Roving condition (top panel) depicts the five frequency levels (A–E, see Methods for further details) randomly distributed across trains within a stimulus block. The ST distance was 8-ST for all five frequency levels. In the Non-Roving condition (bottom panel), only one frequency level was presented throughout each stimulus block (here represented by Train C). Frequency levels were randomized across stimulus blocks (not trains). All five frequency levels (A–E) were used in both conditions. The vertical axis represents frequency (Hz) and the horizontal axis represents time (in seconds). Train duration was 2.48 s. Intertrain interval (ITI) was 3.75 s of silence.

there was no standard intensity value to be detected, from which any tone could be detected as deviant. Therefore, the presence of the MMN component indicated segregation.

To test the hypothesis that the dynamics of the tone context can alter the timing of the buildup to segregation, probe tones were placed in both "early" and "late" positions within the trains (**Figure 1A**). Probe tones occurred randomly in either the third or the fourth tone position within the H tone stream (these were "early" probe tones), and randomly in the 10th, 11th, or 12th tone position within the H stream (these were "late" probe tones). Probe tone position was randomized across trains so that the position within the train could not be anticipated for either the early or late positions. The early position probe tones occurred within the buildup period, whereas the late position tones occurred after the buildup period was surpassed based on results of previous psychophysical and ERP studies (Cusack et al., 2004; Sussman et al., 2007). Thus, responses to the early position probe tones would indicate the timing of the buildup, as they were expected to elicit MMN only when the H and L tones were neurophysiologically segregated in memory at the time the probe tones occurred. All late position probe tones were expected to elicit MMN, and these served as a control, so that if no MMN were elicited in the beginning of the trains, MMN elicited by probe tones at the end of the train would provide evidence that the buildup had been surpassed.

PROCEDURES

Participants were seated in a comfortable chair in a sound-attenuated booth [Industrial Acoustics Company (IAC), Bronx, NY]. Two conditions were presented: Roving and Non-Roving (**Figure 1B**). In the Roving condition, the frequency values of the tones roved on five levels, randomly across trains within the stimulus block (**Figure 1B**, top panel). This was done so that there was no expectation of frequency value, no carryover from one train to the next. In the Non-Roving condition, all five tone pairs were used (i.e., A–E, as in the Roving condition) except that each block contained only one tone pair throughout. Thus, in the Non-roving condition, the tone frequency that occurred from train to train was fully expected. The 3.75 s ITI was chosen because it has been shown to reset the segregation process following that length of silence (Bregman, 1978; Cusack et al., 2004; Sussman et al., 2007); thus, each train served as a new trial for observing the timing of the buildup. Half of participants were presented with the Roving condition first, and the other half with the Non-roving condition first. The order of stimulus blocks within each condition was randomized across participants using a Latin Squares design. Additionally, control blocks were conducted to obtain comparison stimuli for the probe tones. In the control blocks, the intensity value of the standard and the probe tones in the high tone stream were reversed so that the standard tone intensity level was 68 dB and the probe tone was the softer 59 dB, without any other changes in the stimulus parameters. Thus, to delineate the MMN, we subtracted the ERP elicited by the standard tone obtained in the control block from the ERP elicited by the probe tone in the experimental blocks, contrasting two tones that had the same physical properties but differed only in the status within the train. (i.e., 68 dB probe-minus-68 dB standard) (Sussman et al., 2007). Twelve total blocks were presented, which includes the two control blocks, and each block was 5 min in length. Participants were instructed to watch a closed-captioned video and had no task with the sounds. The duration of the session was approximately 2.5 h, which included electrode cap placement and breaks.

DATA ACQUISITION

Electroencephalogram (EEG) was recorded using a 32-electrode cap including a subset of the International 10–20 System (Jasper, 1958) (FPz, Fz, Cz, Pz, Oz, FP1, FP2, F3, F4, F7, F8, FC5, FC6, FC1, FC2, T7, T8, C3, C4, CP5, CP6, CP1, CP2, P7, P8, P3, P4, O1, O2) and left and right mastoids (LM and RM, respectively). The reference electrode was positioned on the tip of the nose. Data was digitized at a 500 Hz sampling rate (bandpass 0.05–100 Hz) (Neuroscan Synamps, Compumedics Inc., Charlotte, NC). Eye blinks were recorded using a bipolar configuration (VEOG) with an external electrode positioned below the left eye (EOG) and the FP1 electrode. Eye saccades were monitored using a bipolar configuration (HEOG) between the F7 and F8 electrodes. Impedances for each electrode were kept below 5 kOhms. Throughout the experiment the EEG was monitored for excessive movement and for eye saccades to ensure the participant was reading the captions.

DATA ANALYSIS

Post-processing of the data included bandpass filtering on the continuous recordings of the EEG offline from 1 to 15 Hz using a zero phase shift Butterworth filter and 24 dB/octave rolloff. Epochs were then created, 600 ms long, including a 100 ms prestimulus onset period. Each epoch was baseline corrected across this entire epoch before artifact rejection ($\pm 75\,\mu V$), then baseline corrected again using the prestimulus period.

ERP responses to the early 3rd and 4th position probe tones were averaged separately for both Roving and Non-Roving conditions. These were the main dependent measures used to evaluate the timing of the buildup to stream segregation. The ERP responses to the 10th, 11th, and 12th position probe tones were averaged together, and served as the "late position" deviant response separately in each condition. The 68 dB high tones were averaged together from the control blocks and served as the standard comparison, separately in each condition. Separately in both Roving and Non-Roving conditions, the averaged ERP responses to the control and to the probe tones were collapsed across all five frequency levels (train types A–E).

The MMN component was delineated in the grand averaged difference waveforms (average ERP elicited by the probe tone minus average ERP elicited by the control). To statistically evaluate the presence of the MMN, a 30-ms interval centered on the peak of the MMN was used to obtain the mean amplitudes evoked by the control and probe tones. The peak was determined in the difference waveforms in the Non-Roving condition from the 4th position and late position, where MMN peaks were observed. These peaks defined the intervals used to measure the MMN in the early and late positions in both conditions (**Table 2**). A two-way repeated measures analysis of variance (ANOVA) with factors of position (early 3rd, early 4th, late) and stimulus type (probe,

Table 2 | Mean amplitudes and intervals used to measure the MMN in each condition and position at the Fz electrode.

Condition	Position	Interval (ms)	Mean (std) μV
Roving	3rd	115–145	0.66 (0.04)
	4th		−0.29 (0.16)
	Late	130–160	−0.58 (0.02)*
Non-Roving	3rd	115–145	0.06 (0.04)
	4th		−1.09 (0.10)*
	Late	130–160	−1.03 (0.08)*

*Indicates significant MMN.

control) was calculated (Statistica 10, Statsoft Inc., Tulsa, OK) to statistically compare the mean amplitudes in each condition. Huynh-Feldt corrections were applied for violations of sphericity and corrected *p*-values reported. *Post-hoc* tests were calculated using the Tukey HSD. Voltage maps of the scalp distribution were created from the peak latency used to measure the MMN in each condition and position, using BESA Research 6.0 software (BESA GmbH, Gräfelfing, Germany).

To compare amplitude of the significant MMNs (Roving late, Non-Roving early 4th, and Non-Roving late position), the mean amplitudes of the difference waveforms were used in a one-way repeated measures ANOVA (**Table 2**).

To compare peak latencies of the significant MMNs (early 4th vs. late) in the Non-Roving condition, latencies were first determined by identifying the local minima [within a 90 ms window (85–175 ms for early probe tones, and 100–190 ms for late probe tones)] on each difference waveform, in each individual. Latency was then compared using a Students *t*-test for dependent measures.

RESULTS

Figure 2 displays the grand-averaged ERPs elicited by the probe and control stimuli. The P1 and the N1 components are small (peak latencies at 50 and 100 ms, respectively) due to the rapid 70 ms SOA presentation rate used in this study (Näätänen and Picton, 1987; Sussman et al., 1998b), but are nonetheless distinctly visible in the standard and deviant ERPs. The 600 ms epoch displays the responses to multiple successive stimuli that are visible and overlap in the waveform after tone onset. The Roving condition, due to the stochastic nature of the frequencies across trains within stimulus blocks shows the greatest variability throughout the epoch. This can be seen as larger differences in the responses to probe vs. control tones early in the epoch before the expected latency (100–250 ms) for MMN (Näätänen, 1995). For this reason MMN was not considered to be present prior to 100 ms.

In the *Roving condition*, there was no main effect of stimulus type [$F_{(1,\ 13)} < 1$, $p = 0.68$]. There was a main effect of position [$F_{(2,\ 26)} = 3.65$, $\varepsilon = 0.87$, $p = 0.047$]. *Post-hoc* analysis showed that the response to the late position probe tone was more negative than the responses to the early 3rd and early 4th position probe tones, with no difference between the 3rd and the 4th position. There was a significant interaction between position and stimulus type [$F_{(2,\ 26)} = 3.92$, $\varepsilon = 0.90$, $p = 0.038$]. *Post-hoc* analysis revealed that the response to the probe tone was more negative than the response to the control tone only for the late position stimuli, but not between the control and probe tone responses in either of the early position stimuli. There was no difference in response to the control tones by position. Thus, the difference was solely due to the response to the late position probe tone. These results show that MMN was significantly elicited (significant difference between probe and control tone) by the late position, but not either early position probe tones.

In the *Non-Roving condition*, there was a main effect of stimulus type [$F_{(1,\ 13)} = 11.28$, $p = 0.005$]. Overall, the ERP response to the probe tone was more negative than the ERP response to the control. There was also a main effect of position [$F_{(2,\ 26)} = 9.34$, $\varepsilon = 0.87$, $p = 0.002$]. *Post-hoc* analysis showed that the responses to the early 4th and late position probe tones were both more negative than the response to the early 3rd position probe tone, with no significant difference between the early 4th and late position probe tones. There was a significant interaction between position and stimulus type [$F_{(2,\ 26)} = 9.59$, $\varepsilon = 0.94$, $p = 0.001$]. *Post-hoc* analysis showed that the probe tone was significantly more negative than the control tone for the early 4th and late position, but not for the early 3rd position. There were no significant differences in the response to the control tones. These results show that MMN was significantly elicited by the early 4th and late position probe tones, but not by the early 3rd position probe tone.

Table 2 displays the mean amplitudes of the MMNs. The amplitudes of the significant MMNs did not significantly differ from each other [$F_{(2,\ 26)} = 1.66$, $p = 0.21$]. Peak latencies of the significant MMNs in the Non-Roving condition also did not significantly differ from each other [$t_{(13)} = 0.62$, $p = 0.55$].

Figure 3 displays the grand averaged difference waveforms. The voltage maps are shown from the peak of the measuring interval (**Table 2**) positioned above each graph to display the scalp topography. The scalp topography of the significant MMNs are consistent with typical fronto-central pattern observed for MMN, including the inversion in polarity at the mastoid electrodes (Giard et al., 2014). As there was no a priori expectation for differences in the cortical generators of the MMNs elicited by Roving and Non-Roving conditions, and no apparent differences observed in the voltage maps, source analyses were not performed.

DISCUSSION

The goal of the current study was to determine whether stimulus context modulated the timing to stream segregation. The key finding was that the regularity of stimulus input led to more rapid stream segregation, as indexed by MMN to intensity probe tones occurring earlier in the tone trains for the Non-Roving (4th position) than the Roving condition (late position). MMNs were elicited by intensity probe tones only in the later part of the tone trains in the Roving condition, suggesting a longer buildup period to neurophysiologic stream segregation for those trains compared to when frequency was repeated from train to train.

These results can be interpreted on the basis of a difference in the timing of the buildup between conditions due to a change

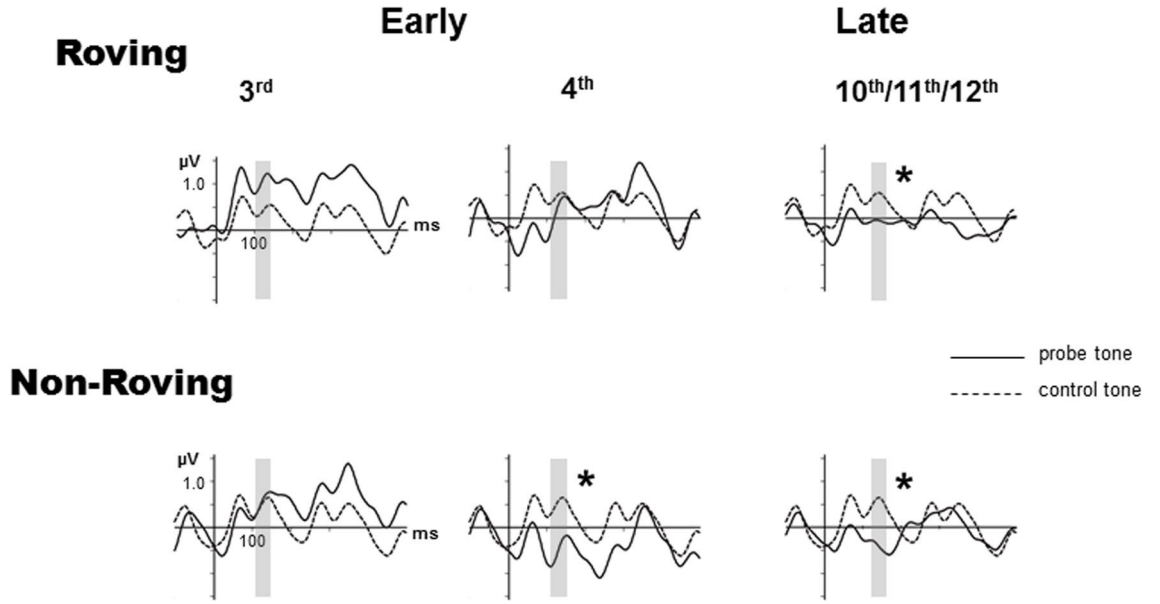


FIGURE 2 | Grand-averaged event-related potentials (ERPs) evoked by probe tones (solid line) and control tones (dashed line) are overlain at the Fz electrode for the early 3rd position tone (left column), early 4th position tone (middle column), and the late position tone, which is the average of the 10–12th tones (right column), separately for the Roving (top row) and Non-Roving (bottom row) conditions. The vertical axis is positioned at stimulus onset, with a 100 ms prestimulus baseline and 500 ms post-stimulus epoch. The interval used to statistically measure the MMN is shaded in gray. Significant MMNs are marked with an asterisk.

in the speed of the segregation process. That is, the elicitation of MMN by the 4th position probe tones occurring only in the Non-Roving condition suggests that the tones were neurophysiologically segregated when the 4th position probe tone occurred. The absence of the MMN to the 3rd position probe tones in the Non-Roving condition suggests that the buildup period was still ongoing at that time. That is, the absence of MMN to the 3rd position probe tone indicates that the silence reset the system but that the speed was faster when the tone frequencies of the trains were repeated. An alternative possibility is that there was no resetting across trains but that the MMN generating system was limited. Although there is evidence that at least two standard tones are required before MMN can be elicited in a single oddball sequence that roves frequencies across tone trains (Cowan et al., 1993), there is no evidence about whether that is also true for oddball sequences that first require segregation to be detected. That is, if the process of detecting the standards occurs only after the segregation occurs, then the MMN system would not have had enough time in our fast paradigm. There may have been a confound between the process of segregation and the standard formation process (Sussman, 2007) in this fast paradigm. We have evidence that the stream segregation process occurs prior to the regularity detection process (Yabe et al., 2001; Sussman, 2005). Thus, if one assumes that the standard formation process begins only once the stream segregation process is neurophysiologically represented, then the standard formation process would begin only after that. These data cannot resolve that issue, and therefore, from these results, we can only conclude that the representation of stream segregation occurred more quickly in the Non-Roving than the Roving condition. Adaptation to the constancy of the frequency presentation could be allowing the segregation process to occur more quickly.

Previous behavioral studies support the explanation of resetting perceived segregation under various conditions (Bregman, 1978; Anstis and Saida, 1985; Beauvois and Meddis, 1997; Carlyon et al., 2001; Cusack et al., 2004; Snyder et al., 2008; Haywood and Roberts, 2010). Although previous behavioral studies have overall suggested that 4 s of silence resets the segregation process (Bregman, 1978; Cusack et al., 2004), Bregman (1978) also demonstrated that shortening the length of the silent period biased the system toward a quicker recovery of segregation. Beauvois and Meddis (1997) showed that varying the silent interval led to an exponential decay of segregation responses as the silence was increased. They presented a 10 s constant frequency induction sequence of a repeating tone at 90 ms SOA and then a test sequence of eight repetitions of two alternating tones with a six semitone frequency separation and the same 90 ms SOA. Between these sequences they varied the silent period from 0 to 8 s. Subjects were instructed to respond after hearing the test sequence and indicate whether they heard an integrated or segregated percept. There was an exponential decrease in the average number of segregation responses as the silent period was increased. This suggested that the time between trains of sounds in their paradigm determined the degree of resetting. However, Beauvois and Meddis (1997) used a constant frequency induction sequence with no silence between the induction and test sequence. Haywood and Roberts (2013), in contrast, showed that an alternating frequency induction sequence had less of an effect on the perception of streaming. Even if the effect was smaller, our neurophysiologic results may be partially explained by their

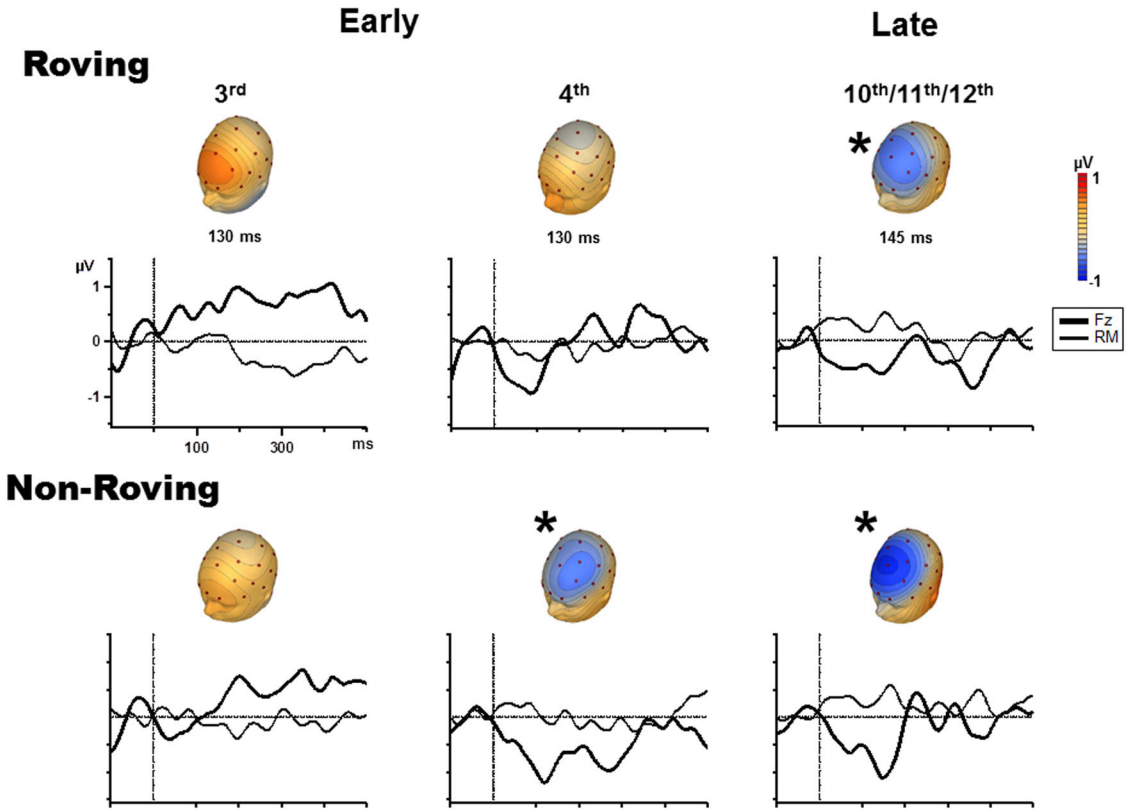


FIGURE 3 | Difference waveforms (ERP response to the probe minus ERP response to the control) are displayed for the Fz (solid, thick line) and mastoid (solid, thin line) electrode sites for the 3rd position (left column), 4th position (middle column), and late position (right column) in the Roving condition (top row) and Non-Roving condition (bottom row). Voltage maps obtained from the difference waveforms at the peak of the measuring interval are presented above each waveform. Significant MMNs are marked with an asterisk, showing typical fronto-central MMN scalp topography.

results. The constancy of the Non-Roving condition may have had a similar biasing effect on the buildup as did shortening of the silence between trains. That is, the silence between trains may have shifted the exponential recovery of neurophysiologic segregation. This suggests a possible rectifying explanation, whereby carryover effects allow the auditory system to maintain a neural representation of segregation that shortens the corresponding buildup period by allowing a quicker recovery of segregation in succeeding trains when the environment is Non-Roving for any number of factors that allow for constancy of perceptual organization.

Overall, our results demonstrate that the dynamics of the environment influences the way in which the auditory system extracts regularities from the input. In dynamically changing situations, it would seem that there would be a propensity to maintain the model of the auditory environment that has been ongoing (Bregman, 1990), and therefore, more information would need to be accumulated before the model was altered. Fully resetting the stream segregation process every time a silence occurs may not be advantageous to maintaining a consistent representation of the auditory scene. In contrast, with multiple overlapping properties of the stimulus input, further analysis may be needed to determine whether a putative new piece of information was an adjustment to the current model, or a signaling to the onset of a new "event." This could alter the timing of segregation for the neurophysiologic model of the auditory scene. To put it into a more realistic context, when walking into a new auditory environment, and successfully organizing the scene (identifying the distinct streams within it), it would be disadvantageous if the scene integrated back every time there was a slight change or silent moment. In contrast, relatively large jumps in frequency from what has been previously ongoing may suggest the onset of a new event, such as a new speaker. Overall, these mechanisms provide the auditory system with an elegant solution to the problem of maintaining stable neural representations in the face of consistency or change, each of which are often encountered in a noisy auditory scene.

ACKNOWLEDGMENTS

This research was supported by the National Institutes of Health (R01 DC004263, Elyse Sussman) and the MSTP training grant, (T32-GM007288, Jonathan Sussman-Fort). We thank Jean DeMarco, Sufen Chen, Grace Sadia and Wei-Wei Lee for assistance with subject preparation and data collection.

REFERENCES

Anstis, S. M., and Saida, S. (1985). Adaptation to auditory streaming of frequency-modulated tones. *J. Exp. Psychol. Hum. Percept. Perform.* 11, 257–271. doi: 10.1037/0096-1523.11.3.257

Beauvois, M. W., and Meddis, R. (1997). Time decay of auditory stream biasing. *Percept. Psychophys.* 59, 81–86. doi: 10.3758/BF03206850

Bregman, A. S. (1978). Auditory streaming is cumulative. *J. Exp. Psychol. Hum. Percept. Perform.* 4, 380–387. doi: 10.1037/0096-1523.4.3.380

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: The MIT Press.

Carlyon, R. P., Cusack, R., Foxton, J. M., and Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *J. Exp. Psychol. Hum. Percept. Perform.* 27, 115–127. doi: 10.1037/0096-1523.27.1.115

Cowan, N., Winkler, I., Teder, W., and Näätänen, R. (1993). Memory prerequisites of mismatch negativity in the auditory event-related potential (ERP). *J. Exp. Psychol. Learn. Mem. Cogn.* 19, 909–921. doi: 10.1037/0278-7393.19.4.909

Cusack, R., Deeks, J., Aikman, G., and Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 643–656. doi: 10.1037/0096-1523.30.4.643

Deike, S., Heil, P., Böckmann-Barthel, M., and Brechmann, A. (2012). The build-up of auditory stream segregation: a different perspective. *Front. Psychol.* 3:461. doi: 10.3389/fpsyg.2012.00461

Giard, M.-H., Besle, J., Aguera, P.-E., Gomot, M., and Bertrand, O. (2014). Scalp current density mapping in the analysis of mismatch negativity paradigms. *Brain Topogr.* (in press). doi: 10.1007/s10548-013-0324-8

Haywood, N. R., and Roberts, B. (2010). Build-up of the tendency to segregate auditory streams: resetting effects evoked by a single deviant tone. *J. Acoust. Soc. Am.* 128, 3019–3031. doi: 10.1121/1.3488675

Haywood, N. R., and Roberts, B. (2013). Build-up of auditory stream segregation induced by tone sequences of constant or alternating frequency and the resetting effects of single deviants. *J. Exp. Psychol. Hum. Percept. Perform.* 39, 1652–1666. doi: 10.1037/a0032562

Jasper, H. H. (1958). The ten–twenty electrode system of the international federation. *Electroencephalogr. Clin. Neurophysiol.* 10, 367–380.

Jones, D., Alford, D., Bridges, A., Tremblay, S., and Macken, B. (1999). Organizational factors in selective attention: the interplay of acoustic distinctiveness and auditory streaming in the irrelevant sound effect. *J. Exp. Psychol. Learn. Mem. Cogn.* 25, 464–473. doi: 10.1037//0278-7393.25.2.464

Micheyl, C., Tian, B., Carlyon, R. P., and Rauschecker, J. P. (2005). Perceptual organization of tone sequences in the auditory cortex of awake macaques. *Neuron* 48, 139–148. doi: 10.1016/j.neuron.2005.08.039

Näätänen, R. (1995). The mismatch negativity: a powerful tool for cognitive neuroscience. *Ear Hear.* 16, 6–18. doi: 10.1097/00003446-199502000-00002

Näätänen, R., and Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology* 24, 375–425. doi: 10.1111/j.1469-8986.1987.tb00311.x

Snyder, J. S., Alain, C., and Picton, T. W. (2006). Effects of attention on neuroelectric correlates of auditory stream segregation. *J. Cogn. Neurosci.* 18, 1–13. doi: 10.1162/089892906775250021

Snyder, J. S., Carter, O. L., Lee, S.-K., Hannon, E. E., and Alain, C. (2008). Effects of context on auditory stream segregation. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 1007–1016. doi: 10.1037/0096-1523.34.4.1007

Sussman, E., Ceponiene, R., Shestakova, A., Näätänen, R., and Winkler, I. (2001). Auditory stream segregation processes operate similarly in school-aged children and adults. *Hear. Res.* 153, 108–114. doi: 10.1016/S0378-5955(00)00261-6

Sussman, E., Ritter, W., and Vaughan, H. G. Jr. (1998a). Attention affects the organization of auditory input associated with the mismatch negativity system. *Brain Res.* 789, 130–138. doi: 10.1016/S0006-8993(97)01443-1

Sussman, E., Ritter, W., and Vaughan, H. G. Jr. (1998b). Predictability of stimulus deviance and the mismatch negativity. *Neuroreport* 9, 4167–4170. doi: 10.1097/00001756-199812210-00031

Sussman, E., and Steinschneider, M. (2009). Attention effects on auditory scene analysis in children. *Neuropsychologia* 47, 771–785. doi: 10.1016/j.neuropsychologia.2008.12.007

Sussman, E., Winkler, I., and Wang, W. (2003). MMN and attention: competition for deviance detection. *Psychophysiology* 40, 430–435. doi: 10.1111/1469-8986.00045

Sussman, E. S. (2005). Integration and segregation in auditory scene analysis. *J. Acoust. Soc. Am.* 117, 1285–1298. doi: 10.1121/1.1854312

Sussman, E. S. (2007). A new view on the MMN and attention debate: the role of context in processing auditory events. *J. Psychophysiol.* 21, 164–175. doi: 10.1027/0269-8803.21.34.164

Sussman, E. S., Bregman, A. S., Wang, W. J., and Khan, F. J. (2005). Attentional modulation of electrophysiological activity in auditory cortex for unattended sounds within multistream auditory environments. *Cogn. Affect. Behav. Neurosci.* 5, 93–110. doi: 10.3758/CABN.5.1.93

Sussman, E. S., Horváth, J., Winkler, I., and Orr, M. (2007). The role of attention in the formation of auditory streams. *Percept. Psychophys.* 69, 136–152. doi: 10.3758/BF03194460

Vaughan, H. G. Jr., and Ritter, W. (1970). The sources of auditory evoked responses recorded from the human scalp. *Electroencephalogr. Clin. Neurophysiol.* 28, 360–367. doi: 10.1016/0013-4694(70)90228-2

Winkler, I., Czigler, I., Sussman, E., Horváth, J., and Balázs, L. (2005). Preattentive binding of auditory and visual stimulus features. *J. Cogn. Neurosci.* 17, 320–339. doi: 10.1162/0898929053124866

Yabe, H., Winkler, I., Czigler, I., Koyama, S., Kakigi, R., Sutoh, T., et al. (2001). Organizing sound sequences in the human brain: the interplay of auditory streaming and temporal integration. *Brain Res.* 897, 222–227. doi: 10.1016/S0006-8993(01)02224-7

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

Received: 16 January 2014; accepted: 11 April 2014; published online: 29 April 2014.
*Citation: Sussman-Fort J and Sussman E (2014) The effect of stimulus context on the buildup to stream segregation. Front. Neurosci. **8**:93. doi: 10.3389/fnins.2014.00093*
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 23 May 2014
doi: 10.3389/fnins.2014.00116

Online tracking of the contents of conscious perception using real-time fMRI

Christoph Reichert[1,2,3], Robert Fendrich[1,4], Johannes Bernarding[5], Claus Tempelmann[1], Hermann Hinrichs[1,3,6,7,8] and Jochem W. Rieger[9,10]*

[1] Department of Neurology, University Medical Center A.ö.R., Magdeburg, Germany
[2] Department of Knowledge and Language Processing, Otto-von-Guericke University, Magdeburg, Germany
[3] Forschungscampus STIMULATE, Magdeburg, Germany
[4] Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA
[5] Institute for Biometry and Medical Informatics, Medical Faculty, Otto-von-Guericke University, Magdeburg, Germany
[6] Department of Behavioral Neurology, Leibniz Institute for Neurobiology, Magdeburg, Germany
[7] German Center for Neurodegenerative Diseases (DZNE), Magdeburg, Germany
[8] Center for Behavioral Brain Sciences, Magdeburg, Germany
[9] Department of Applied Neurocognitive Psychology, Carl-von-Ossietzky University, Oldenburg, Germany
[10] Research Center for Neurosensory Sciences, Carl-von-Ossietzky University, Oldenburg, Germany

Edited by:
Susann Deike, Leibniz Institute for Neurobiology, Germany

Reviewed by:
Peter Christiaan Klink, Royal Netherlands Academy of Arts and Sciences, Netherlands
Hirohito M. Kondo, NTT Corporation, Japan

***Correspondence:**
Jochem W. Rieger, Neurocognitive Psychology Lab, Department of Psychology, Faculty VI, Carl-von-Ossietzky University, D-26111 Oldenburg, Germany
e-mail: jochem.rieger@uni-oldenburg.de

Perception is an active process that interprets and structures the stimulus input based on assumptions about its possible causes. We use real-time functional magnetic resonance imaging (rtfMRI) to investigate a particularly powerful demonstration of dynamic object integration in which the same physical stimulus intermittently elicits categorically different conscious object percepts. In this study, we simulated an outline object that is moving behind a narrow slit. With such displays, the physically identical stimulus can elicit categorically different percepts that either correspond closely to the physical stimulus (vertically moving line segments) or represent a hypothesis about the underlying cause of the physical stimulus (a horizontally moving object that is partly occluded). In the latter case, the brain must construct an object from the input sequence. Combining rtfMRI with machine learning techniques we show that it is possible to determine online the momentary state of a subject's conscious percept from time resolved BOLD-activity. In addition, we found that feedback about the currently decoded percept increased the decoding rates compared to prior fMRI recordings of the same stimulus without feedback presentation. The analysis of the trained classifier revealed a brain network that discriminates contents of conscious perception with antagonistic interactions between early sensory areas that represent physical stimulus properties and higher-tier brain areas. During integrated object percepts, brain activity decreases in early sensory areas and increases in higher-tier areas. We conclude that it is possible to use BOLD responses to reliably track the contents of conscious visual perception with a relatively high temporal resolution. We suggest that our approach can also be used to investigate the neural basis of auditory object formation and discuss the results in the context of predictive coding theory.

Keywords: anorthoscopic, bistable perception, ambiguous stimulus, real-time fMRI, object integration, slit viewing

INTRODUCTION

Human cognitive neuroscience makes the strong assumption that all subjective human experience is tightly linked to spatiotemporal patterns of a neuronal activation. For sensory perception this assumption predicts that patterns of brain activity should not only reflect the physical stimulus but also the content of perceptual awareness derived from the physical input (von Helmholtz, 1867). Perceptually ambiguous stimuli, in which the same physical stimulus elicits categorically different and mutually exclusive percepts, provide an excellent opportunity to investigate constructive brain processes underlying the formation and the maintenance of subjective perceptual experiences. The neural correlates of different ambiguous visual stimuli have been studied using fMRI, including binocular rivalry (Tong et al., 1998; Haynes and Rees, 2005), ambiguous static figures (Kleinschmidt et al., 1998), structure-from-motion (Brouwer and van Ee, 2007; Freeman et al., 2012), occluded moving drawings (Murray et al., 2002; Fang et al., 2008), and moving plaids (Castelo-Branco et al., 2002). It has been suggested that a strong link between dynamically changing brain activation patterns and subjective perceptual states can be established by predicting different states of perception from brain activation (Cox and Savoy, 2003; Rieger et al., 2008; Chang et al., 2010). However, none of the previous fMRI-studies on perceptual organization of ambiguous stimuli made a serious attempt to quantify the relevance of the observed

BOLD-modulations for the tracking of spontaneously changing states of awareness.

In vision, occlusion of a moving object, a situation common in everyday vision, requires integration of sequentially visible object fragments into the percept of a coherent object. The stimulus we employ in this study, mimics this occlusion problem (**Figure 1**). It simulates an outline figure moving horizontally back and forth behind an occluder with a narrow vertical aperture, a situation similar to looking through a door that is only open a slit. Importantly, with our stimulus an observer's conscious percept intermittently switches between the actually presented sequence of object parts, short line segments moving vertically, and a horizontally moving integrated object that is located behind an occluder. In previous studies (Fendrich et al., 2005; Rieger et al., 2007) we have shown that under free viewing conditions the integrated object percept is constructed post-retinally in the brain and that eye movements do not contribute to the construction of the integrated object. Murray et al. (2002) investigated the effects of dynamic bistable figure integration on the BOLD-activation level in human V1. The subjects observed a diamond shaped outline figure moving behind apertures. The authors reported a decrease of BOLD-activation in V1 while subjects perceived the integrated figure and conclude that this reduction is in concordance with predictive coding theory. Predictive coding theory states that representation of simple stimulus features in early sensory cortices is explained away by feedback from higher areas involved in the creation of derived percepts, such as occluded objects. Fang et al. (2008) and de-Wit et al. (2012) reported activation changes in the lateral occipital complex (LOC) with opposite sign than those in V1. Unfortunately, all these authors reported only BOLD-modulations in a small pre-selected set of visual brain areas, and therefore do not address the extent and functional characteristics of the brain networks that carry information about the content of visual perception and are thus likely to be involved in establishing the current percept of the ambiguous stimulus.

In the auditory domain, object formation from sequentially presented tones requires temporal integration, similar to the occluded-object stimulus we are using in the current study. Whole head fMRI suggests roles for intraparietal sulcus in modality independent structuring of sensory input Cusack (2005) and feedforward-feedback interactions between cortical and thalamic regions during percept switching (Kondo and Kashino, 2009). In addition, neuromagnetic measurements indicate that non-primary auditory areas maintain the representation of auditory streams (Gutschalk et al., 2005). Furthermore, left auditory cortex was found to play a dominant role in active auditory stream segregation (Deike et al., 2010). Interestingly, Pressnitzer and Hupé (2006) suggest that although in vision and audition object formation and perceptual bistablity may have different underling neuronal correlates, both modalities may share similar functional mechanisms with similar dynamics. Thus, a method suitable to establish a tight link between BOLD-activity and the state of awareness in bistable visual object formation will likely be useful to investigate bistable auditory object formation.

Our first goal in this study is to determine if the brain activation changes linked to the two conscious percepts of our ambiguous display are robust enough to allow a reliable determination of the current content of conscious visual perception. It should be noted that activation changes that occur exclusively in the visual system will not necessarily be the most informative indicators of the actual conscious visual percept. Therefore, we use whole-brain fMRI scans and multivariate machine learning techniques to determine brain activation patterns that discriminate between subjective perceptual states (Norman et al., 2006). As an ultimate proof of concept our main goal was to perform the discrimination online. In our asynchronous online prediction scheme no prior knowledge about the time of percept switches is available which makes tracking of percepts more challenging than prediction of synchronous perceptual events (e.g., Hollmann et al., 2011). To date, real-time fMRI has been employed in several tasks to predict brain states (LaConte et al., 2007; Hollmann et al., 2011) and to provide online feedback (Weiskopf et al., 2007; Weiskopf, 2012). However, we are unaware of a study that shows the feasibility of rtfMRI discrimination of the contents of visual perception with ambiguous stimuli. Furthermore, we propose a method to test the generalizability of the multivariately assessed brain patterns over subjects. Our third goal was to characterize the brain networks that allow for discrimination between the perception of integrated objects vs. object parts and to establish whether the relative activation levels in this network are concordant with predictive coding theory. The methods we apply may also be applicable for discriminating bistable auditory percepts, provided that effect sizes and temporal dynamics are similar.

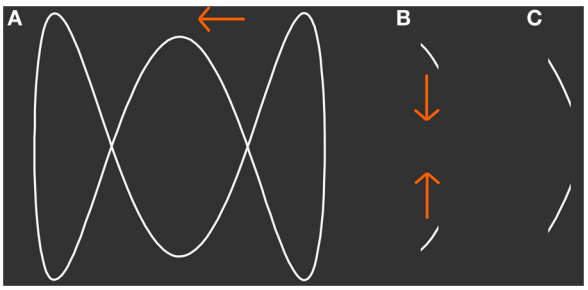


FIGURE 1 | Presented stimulus. (A) The shape of a 3-loop-figure that was simulated to move horizontally back and forth behind a narrow aperture. **(B,C)** Show snapshots of the figure that can be seen through the simulated aperture at two different points in time. Although only line segments are visible at each point in time the subject is able to perceive the figure as a whole. At a specific width of the aperture the participant's conscious perception switches spontaneously between vertically jumping line segments and the horizontally moving figure. From **(B)** to **(C)** the figure moves to the left.

MATERIALS AND METHODS

STIMULI AND STIMULUS PRESENTATION

The stimulus was the simulation of an outline loop-figure moving horizontally behind a narrow vertical aperture with invisible borders (**Figure 1**). It was back projected onto a translucent screen with a JVC DLA-G150CL projector and viewed by the subject via a mirror. The eye-screen distance was 59 cm.

The width of the aperture was individually adjusted such that observers had intermittent percepts that spontaneously switched

between the physical stimulus, short line segments with changing orientation moving vertically up and down, or an integrated but occluded object moving horizontally back and forth behind an aperture (Fendrich et al., 2005). The height and width of the outline figure was 6.8 and 8.7° visual angle, respectively. The individual aperture widths ranged between 3.7 and 7.0% of the outline figure width. We will denote the percept of the physical stimulus as the lines percept and the percept of the integrated figure as the object percept.

Subjects freely viewed the display and reported their percept by holding one of two buttons pressed throughout the whole interval the percept lasted. We found in earlier investigations that under free viewing conditions eye movements tracking the horizontal object movements were negligibly small and did not influence the quantity or quality of object percepts (Fendrich et al., 2005; Rieger et al., 2007). The assignment of the button presses to the line and object percepts was switched after each run to avoid a correlation of any purely motor BOLD-activations produced by the button presses with the activations produced by changes in the conscious percepts. In real-time mode, we presented visual feedback by switching line color slightly and auditory feedback by presenting the decoded percept as spoken word from an audio file in the moment of switch detection, respectively.

SUBJECTS AND EXPERIMENTS

Data were acquired in two experiments which we will refer to as offline experiment and online experiment. In the offline experiment, we used conventional fMRI to record the BOLD-activations while subjects viewed the stimulus and reported their subjective percepts via button presses. Fifteen subjects participated in this experiment (8 female, 7 male, mean age = 26.3 years). The Data of this experiment we analyzed only in an offline mode as it is conventionally performed in fMRI. In order to prove the reliability of our approach, we conducted a second experiment. The online experiment was designed to track the current content of conscious perception using real-time fMRI BOLD-measurements, and we performed both an online and offline data analysis. Ten subjects participated in this experiment (6 female, 4 male, mean age = 25.7 years).

The experiments were approved by the ethics committee of the Medical Faculty of the Otto-von-Guericke University. All subjects gave their informed consent prior to the start of the experiment. They had normal or corrected to normal vision, and were paid for participation.

MR-SCANNING

A Siemens-Trio scanner equipped with an 8-channel phased array head-coil was used for anatomical and functional MRI. Full head T1-weighted anatomical images were obtained with an MPRAGE sequence (80 axial slices, slice thickness = 2 mm, field of view = 256 by 192 mm; inplane matrix = 256 by 192). Functional images were obtained from the whole head with a gradient recalled echoplanar imaging (EPI) sequence [32 axial slices, slice thickness = 4 mm, field of view = 200 mm; inplane matrix = 64 by 64; $TR = 2.5$ s (offline experiment) and 2.0 s (online experiment), echo time = 30 ms (offline experiment) and 27 ms (online experiment)]. Each run lasted 420 s, and each subject completed between 7 and 8 runs in the offline experiment and 8 runs in the online experiment.

SUPPORT VECTOR MACHINE LEARNING

Support vector machines (SVMs) represent a class of machine learning algorithms characterized by good generalization performance even in high dimensional feature spaces (Vapnik, 1998). This is achieved because the internal regularization avoids overfitting so that the complexity of the classifier does not depend on the complexity of the feature space (Cherkassky and Mulier, 1998). Therefore, SVMs are a suitable approach for analyzing fMRI signals, particularly when a whole head analysis is preferred. The central idea of the algorithm is to maximize the distances from the training data to the separating hyperplane which is characterized by its normal vector $\vec{w}$, also referred to as the weight vector, and a bias parameter b. With these parameters estimated from the training data, labels y_i of new data $\vec{x}_i$ can be predicted by a simple inner product with the function:

$$y_i = sign\,(\vec{w} \cdot \vec{x}_i + b) \tag{1}$$

Note that the inner product between the classifier weights and the measured data can be interpreted as a weighted "voting" of each voxel for one or the other class. Due to the properties of the inner product, votes for positive and negative classes can be generated by multiple combinations of positive or negative weights with positive or negative voxel values. It is thus difficult to interpret the sign of the weight. A comprehensive introduction to support vector machine learning is provided in Burges (1998). In our study we pre-selected a relatively high number of voxels and assigned the signal change of each voxel as input data for a linear SVM. Labels for both training and testing of the SVMs were derived from the button presses shifted 5 s forward in time to account for the delay of the hemodynamic response function (HRF).

DATA PROCESSING

Preprocessing

In the offline discrimination of line vs. object percepts we employed several preprocessing steps using the SPM5 package (Wellcome Department of Cognitive Neurology, University College London, UK). We corrected slice acquisition time and head movements, and we spatially normalized the individual brains to the MNI standard brain. Finally, we spatially smoothed the functional data with an 8 mm FWHM Gaussian kernel. In the temporal domain we band-pass filtered the voxel time series. We set the high-pass cut-off frequency to 1/128 Hz and determined the low-pass cut-off frequency between 1/27 and 1/8 Hz individually as described below. We discarded voxels exceeding 10% signal change from the analysis because at 3T they are likely to stem from large vessel contributions.

Offline-classification

In the offline analysis which was applied to both experiments, we performed a leave-one-run-out cross-validation rather than using a leave-one-volume-out or random-selection of volumes approach to test for generalizability. This avoids information transfer between training and test set due to temporal correlations in the slowly varying BOLD response. In this procedure, all

runs but one are used for feature selection, classifier training, and filter optimization. The decoding accuracy of the trained classifier is then tested on the reserved test set.

We used a univariate statistical approach for feature selection to reduce computational costs. Therefore, we calculated model BOLD-responses from the reported object and line percepts, regressed them on to the voxel BOLD-time courses and calculated univariate *F*-tests to assess statistical difference between the BOLD-amplitudes estimated for the two percepts. Finally, the 5000 voxels with the lowest *p*-value were selected for the classification step.

In addition to feature selection, the cutoff frequency of the low-pass filter was optimized in the cross-validation loop to minimize high frequency noise and to account for individual variations in the duration of the percepts. We tested the classification performance on the training set with four low-pass cutoff frequencies 1/27, 1/18, 1/12 and 1/8 Hz and selected the cut-off with the smallest classification error in cross validation performed on the training set. Although minimal loss on the training set does not necessarily imply optimal performance on the test data, we found this approach appropriate.

The trained classifier was then applied to each EPI-volume in the left out test run to determine the subject's current conscious percept from the BOLD-activation patterns. For evaluation of the classifier's discrimination performance we compared the time course of the subject's button presses shifted by 5 s to take into account the HRF delay. For classification and cross-validation we combined the Princeton MVPA toolbox (http://code.google.com/p/princeton-mvpa-toolbox/) with the Spider machine learning toolbox (http://www.kyb.tuebingen.mpg.de/bs/people/spider).

Online-classification

Online analysis is the ultimate test of the ability to track the phenomenal content of ambiguous perception time resolved from BOLD-activations. We implemented a real-time fMRI-classification experiment based on the experimental description language EDL (Hollmann et al., 2008). Data acquisition was performed with the same parameters as in the offline experiment except for shorter TRs. The MR scanners' EPI scanning protocol was modified to immediately export the acquired and motion corrected volumes (Hollmann et al., 2008). On the classification host, transformation matrices for spatial normalization were calculated from the first acquired EPI-volume and applied to all subsequently acquired volumes. After normalization, the EPI-volumes were spatially smoothed (8 mm FWHM Gaussian kernel). Stimulus presentation and classification were performed on two separate computers that communicated via RS-232 connection.

The first two runs of each scanning session were used for feature selection and for training of the initial classifier. First, the BOLD time series were zero centered and the baseline values were kept for subsequent subtraction to approximate a zero centered signal online. Then the time series were temporally filtered applying a digital fourth order butterworth band-pass filter with cutoff frequencies 1/128 and 1/16 Hz. Initial univariate feature selection was performed similar to the procedure described in the offline experiment except that we applied an ANOVA as statistical test and retained the voxels with the smallest 10,000 *p*-values for a first pass of classifier training. We used the initial classifier for a second, multivariate feature selection in which we retained only voxels with feature weights exceeding a threshold w_{th} which was defined at $w_{th} = 10^{-1} max\,|w_k|\,,\ k = 1 \ldots 10,000$. The resulting feature space was used for further analysis including retraining on the current data set. Online discrimination of the content of perception started with the third run. Auditory feedback was provided by playing the respective word (female voice speaking) when a switch in the state of the percept was detected. Furthermore, the according state was continuously indicated by slightly changing the color of line segments to greenish for decoded object percepts and to reddish for decoded line percepts. Analogous to the offline experiment, subjects reported their perceptual state by button presses. Moreover, we instructed the participants to mentally assess the correctness of the decoded percept, but to consider a delay of approximately 8 s, expectable due to the hemodynamic response delay and data processing time. Eight runs were acquired for each subject. The classifier was updated after every run starting from the third.

Cross-subject generalization

In analogy to the classical statistical approach, we aimed to use multivariate classification to derive predictive patterns of BOLD-activation that would generalize across individual subjects. Multivariate pattern analysis often focuses on individual discriminative patterns within a small circumscribed brain area (e.g., Kay et al., 2008). Therefore, this approach is often not commensurate to the attempt in classical statistical analysis to generalize experimental results to a population. Our approach aims to derive patterns of BOLD-activation from a group of subjects that will generalize to new subjects. The generalization is tested by showing that the patterns found can discriminate conscious perceptual contents with better than chance performance in data from new subjects. This last test of generalization performance is typically omitted in classical statistics but critical for the evaluation of the relevance of brain activation patterns at the population level, as it is performed with standard parametric univariate testing.

In our approach, we used a leave-one-subject-out cross-validation. Because the total number of EPI-volumes (19,188 in the offline experiment and 16,480 in the online experiment, respectively) is prohibitively large for our whole head analysis approach and considering that classification appears to be more reliable distant from perceptual switches, we reduced the amount of training data by averaging over three EPI-volumes around the center of each interval of a sustained percept. This reduced the data set to 2526 EPI-volumes in the offline experiment and 1412 EPI-volumes in the online experiment, respectively.

For feature selection we trained an SVM on all available volumes of a single subject involving 30,000 voxels with lowest *p*-values derived from an *F*-statistic. Then we calculated a weighted combination of each training subject's weight vector

$$\overline{w}^k = \frac{1}{n} \sum_{i=1}^{n} 2 \left(\hat{p}_i - \frac{1}{2} \right) w_i^k \qquad (2)$$

In this equation, w_i^k is the individual weighting of the k^{th} feature in the normal vector $\vec{w}_i$ and $\hat{p}_i$ is the rate of correct classifications achieved with the i^{th} of n subjects. The result is a map of voxels weighted by both the importance of the voxel for the individual subjects and the relevance of the whole individual voxel map for the discrimination of the perceptual content. To perform the group analysis, we generated these maps in cross-validations and selected the 30,000 highest weighted voxels from the combined map.

Permutation testing

We evaluated the reliability of the classifier's discrimination rate with respect to the empirical guessing level and its 95% confidence obtained with a permutation test. Only discrimination rates exceeding the 95% confidence interval for guessing were considered reliable (Good, 2005). Note that the mean empirical guessing levels can substantially deviate from the expected level (50% in the two class case), and may have wide confidence intervals (see e.g., Rieger et al., 2008). For permutation testing, the available labels, line, or object percept, were randomly reassigned among the measured EPI-volumes. Ideally, this procedure should prevent the classifier from learning information useful for separating the classes. However, even though labels are randomly assigned to the EPI-volumes in most cases the classifier will discriminate line and object percepts with a rate different from 50% correct. The critical question is, whether the discrimination rate obtained with the actually measured data set is within the confidence interval for discrimination rates that can be obtained with randomly assigned labels. However, since we discriminate events in a continuous time series the permutation scheme should consider temporal correlations due to the distribution of durations of the line and object percepts. We retained these durations in the permutation samples by permuting full blocks of consecutively equal sample labels. The empirical guessing levels were estimated within the above described cross-validation framework in a total of 500 random permutations.

A second application of permutation testing is to generate a distribution of classifier weights that can be obtained by separating the data without knowledge of actual class labels. We used this method to determine the significance that a feature weight deviates from a randomly obtained weight. According to Mourão-Miranda et al. (2005) we calculated the p-value, denoted p_w, as the ratio of number of features exceeding the weight of the classifier in the distribution of randomly obtained weights for this feature to the number of permutations. A low value for p_w indicates that the voxel contributes discriminative information to the classifier.

RESULTS

BEHAVIORAL RESULTS

The distribution of the duration of the perceptual interval lengths in bistable percepts can be approximated by a gamma probability density function (pdf) (Kleinschmidt et al., 1998). **Figure 2** shows the histograms of the interval durations in the two experiments and the fitted gamma pdf. The median duration in the online experiment was 21.4 s (**Figure 2B**). The maximum of the fitted gamma pdf is at 18.3 s (goodness of fit: $R^2 = 0.8416$). In the offline experiment the median interval duration was 12.1 s (**Figure 2A**) and the maximum of the gamma pdf is at 6.9 s (goodness of fit: $R^2 = 0.9851$). The longer perceptual intervals in the online experiment were presumably due to familiarizing the subjects with the stimulus before the beginning of the experiment. The average proportion of object percepts does not differ between the online and the offline experiment [offline: 53.9%, online: 51.7%, two-sample t-test: $t_{(23)} = 0.9$, $p = 0.37$].

ACCURACY OF THE DECODED PERCEPTS

Averaged over all subjects, the phenomenal content of perception (lines or object) was correctly determined 79.2% of the time (EPI-volumes) in the offline experiment, with a standard error of the mean (SE) of 2.6%. **Figure 3A** shows the individual results for all 15 subjects including the empirical 95% confidence intervals for guessing. In the best subject, the classifier tracked the phenomenal content nearly perfectly, at 94.4%. Most errors in this subject occurred in EPI-volumes around the time of a perceptual switch (**Figure 3B**). This observation was also found in other subjects and is supported by the observation that without the three volumes around a perceptual switch the decoding accuracy over all subjects increases to 84.8% and that the decoding accuracy dropped to 72.4% when we considered only the EPI-volumes around the switches for accuracy calculation. This suggests that the limited temporal sampling of fMRI (here 0.4 Hz sampling rate) may contribute to the decoding error. Importantly, in every subject the classification rate clearly exceeds the 95% confidence interval for guessing. The theoretical 50% guessing level is always included in the 95% confidence interval determined by the permutation test. This indicates that there is no bias for one class in the data sets (Rieger et al., 2008).

ONLINE TRACKING OF CONSCIOUS PERCEPTION

Online, time resolved discrimination of the phenomenal content of the subjects' percepts was possible with great precision. Importantly, the online prediction started after only 14 min of training data acquisition (412 samples). On average 82.8% (SE: 2.4%) of the time the classifier correctly determined whether the subject currently perceived lines or an integrated object. Classification rates ranged from nearly perfect (94.2%) for the best subject to 65.9% for the worst. Although the classifier was retrained after each run, decoding accuracies did not significantly change over runs (average linear regression slope −0.24%, $p > 0.05$). In concordance with the offline experiment, most discrimination errors appeared around the perceptual switch (average decoding accuracy increases to 87.0% correct when ignoring three volumes around switch and decreases to 70.4% correct when taking only switch volumes into account), indicating that a considerable proportion of the errors might occur due to slow hemodynamic response and a relative low sampling rate of fMRI.

The average temporal delay from the beginning of the EPI-volume acquisition to feedback was 4.3 s (SE: 0.05 s). About half of this delay was due to the EPI-acquisition ($TR = 2$ s). Most of the remaining delay was required for data transfer and signal processing. Assuming a hemodynamic delay of 5 s, feedback was delayed by 9.3 s. The median duration of a phenomenal percept was 21.4 s and only 4.5% of the percept durations were shorter than this delay. Subjects reported that they could readily relate the

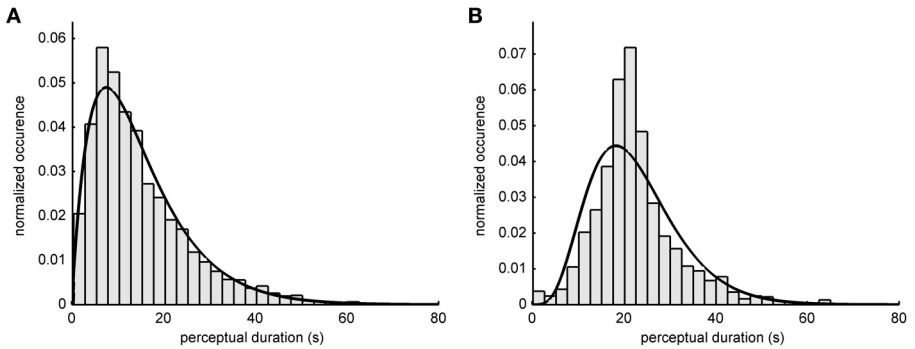


FIGURE 2 | Perceptual switch intervals. Histograms show the relative occurrence how long a perceptual state lasted for **(A)** 15 participants of the offline experiment and **(B)** 10 participants of the online experiment. A gamma probability density function (pdf) was fitted to the distributions. The median duration of one percept in the offline experiment was 12.1 s (pdf shape: 2.00; scale: 7.51). In the online experiment the median duration of the percept was 21.4 s (pdf shape: 5.29; scale: 4.26).

feedback to their phenomenal percepts despite the delay. The feature selection scheme described in section Online-classification revealed an average feature set size of 3997 voxels (SE: 215 voxels).

The classification rates obtained in an online experiment are most likely sub-optimal. One reason is that we stopped feature selection after the second run. To test if even more reliable information about the momentary conscious percept can be revealed by optimization of the signal processing chain we performed an additional offline leave-one-run-out cross-validation analysis in which we optimized preprocessing parameters as described in section Offline-classification. On average the decoding accuracy increased by 6% to an accuracy of 88.8% (SE: 1.9%; 74.5–94.2%) (**Figure 4**). All classification rates clearly exceeded the individually determined 95% confidence intervals for guessing. Importantly, the decoding accuracy is higher when subjects received feedback compared to the experiment without feedback presentation. Our results clearly show that fMRI in combination with advanced machine learning approaches makes it possible to track the phenomenal content of a subject's percept online with substantial time resolved accuracy.

GENERALIZATION OVER SUBJECTS

In a final analysis we investigated whether a classifier derived from a population of subjects, would be capable of predicting the phenomenal content of a new subject's percept. This is of particular interest for brain computer interface related research. We trained the classifier in a leave-one-subject-out cross-validation scheme as described in section Cross-subject generalization. When we excluded the EPI-volumes around perceptual switches for training and evaluation, using only the EPI-volumes with the most reliable labels, we obtained 78.1% correct classifications for the data from the offline experiment and 79.4% accuracy for the data from the online experiment. Including the EPI-volumes acquired around switches in the evaluation yielded 68% accuracy with the data from the offline and 73.11% with the data of the online experiment. In this analysis accuracy exceeded the guessing level for all but one of the subjects. The result suggests that a classifier derived from a population can indeed transfer to new subjects, although information loss occurred compared to the single subject analysis.

BRAIN AREAS INVOLVED IN PERCEPT DISCRIMINATION

An essential question is what brain networks discriminate between the contents of conscious percepts in dynamic object integration. We analyzed the trained classifiers to address this question. In case the input features are scaled comparably, it is possible to interpret the feature weights learned with linear SVM as informative voxels in a brain network that discriminates the contents of the percepts.

To determine if a voxel contributed significantly to discrimination of percepts we used a combined multivariate and univariate approach. First we calculated a multivariate p-value p_w for each voxel by non-parametric permutation testing involving all subjects from both experiments using the group data sets described in Cross-subject generalization and labels "line" or "object" percept permuted among the BOLD-data. This procedure provides a null distribution of weights for each voxel. We trained 1000 classifiers for both group data sets. In the next step we used the weights from the classifiers trained with the actually measured labels and the permutation distributions of weights to calculate the probability p_w that the weights obtained with the "correct" labels were observed by chance. In **Figure 5** we show a map of the voxels that reached a weight-threshold of $p_w < 0.1$ in the conjunction of both experiments. Second, in order to show BOLD change direction and univariate effect sizes we calculated t-values for each voxel. Red to yellow indicates that BOLD-activity in the voxel is higher during object percepts and blue to cyan indicates lower BOLD-activity during object percepts. The two sided t-value for a conventional univariate threshold of $p < 0.001$ would correspond to $t_{(3936)} < -3.2$ or $t_{(3936)} > 3.2$, which applies to 88.6% of the voxels shown in **Figure 5** as well as to all of the clusters shown. Note that the maps are smooth and unvariate effects are uniform in each cluster. This smoothness indicates systematic effects over voxel in each cluster. Moreover, the combination of multivariate classifier training with univariate analysis suggests that the classifier

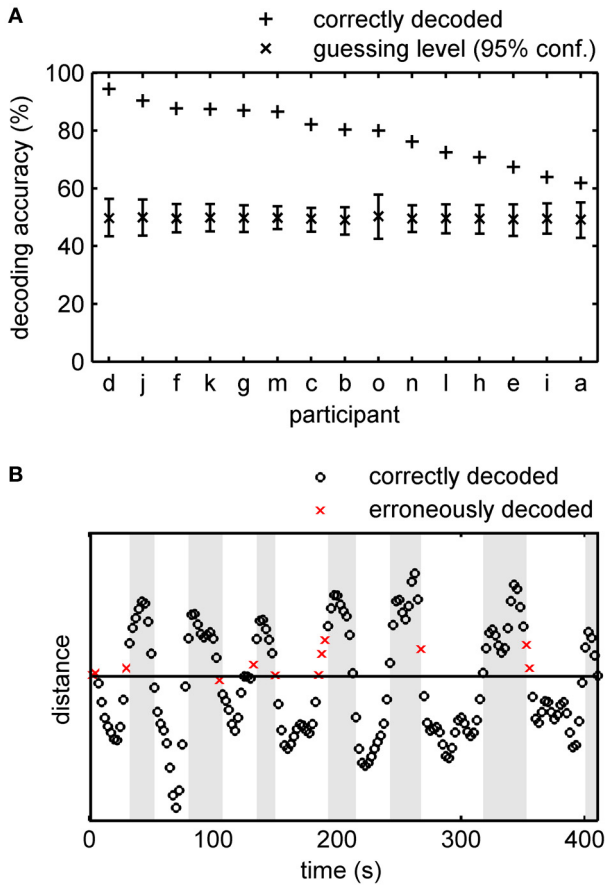


FIGURE 3 | Accuracy of offline tracking. (A) Single subject decoding accuracy obtained by a leave-one-run-out cross-validation. On average 79.2% (SE: 2.6%) of the acquisition time the decoded percept was correct. For each participant the mean guessing level as determined by permutation testing is approximately 50% and the fraction of correctly decoded perceptual states exceeds the 95% confidence interval of the guessing level. Participants are sorted by performance. **(B)** Time course of the distance (arbitrary unit) of the data points in classification space to the separating hyperplane constructed by a SVM for participant *d*, Run 5 (92.7% accuracy). White background signifies intervals in which the participant indicated perceiving the integrated object and gray background signifies intervals with line percepts. EPI-volumes for which the conscious percept was correctly identified are shown as circles, incorrectly identified percepts as crosses.

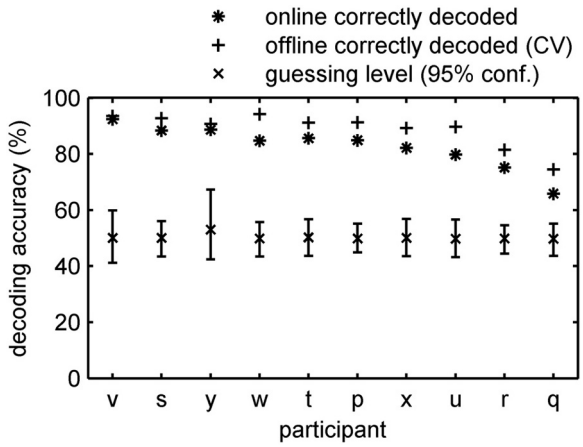


FIGURE 4 | Accuracy of online tracking. Single subject decoding accuracy obtained during the online decoding experiment with feedback. On average 82.8% (SE: 2.4%) of the time the content of the conscious percept was correctly decoded. An additional offline cross-validation procedure applied to the same data revealed 88.8% (SE: 1.9%) accuracy on average. Participants are sorted by performance.

does not only cancel out correlated noise in voxels with reliable weights.

Brain areas that contributed significantly to the classification include relatively early retinotopic visual cortex [MNI coordinates (5, −77, 8), $t_{(3936)} = -23.0$] and the middle temporal MT+ [MNI coordinates (47, −70, 13) and (−39, −65, 12), $t_{(3936)} = -14.5$ and $t_{(3936)} = -13.7$] where BOLD activity is lower during the object perception, and ventral visual stream areas such as putative lateral occipital complex [LOC, MNI coordinates (32, −88, 13) and (−34, −88, 8), $t_{(3936)} = 4.6$ and $t_{(3936)} = 7.4$] where BOLD activity is higher during the object percept. However, the whole head classification approach suggests that the discriminative network in most subjects includes brain areas that are not primarily visual such as the lateral intraparietal sulcus [IPS, MNI coordinates (27, −57, 63) and (−22, −58, 58), $t_{(3936)} = 15.9$ and $t_{(3936)} = 14.3$], the frontal eye fields [FEF, MNI coordinates (30, −9, 51) and (−28, −7, 53), $t_{(3936)} = 7.5$ and $t_{(3936)} = 13.2$], and sporadic small clusters of voxels in the frontal lobe.

DISCUSSION

In this work we show that it is possible to use rtfMRI to reliably track online the current content of conscious visual perception while observers view ambiguous stimuli. The reliability of the approach is indicated by permutation tests and the observation that errors occurred primarily close to the time of perceptual switches. Importantly, we found that with our approach the trained classifiers generalize over subjects. Our approach to investigating the neural networks that underlie perceptual representation is data driven as well as hypothesis generating. This distinguishes it from earlier approaches that have focused on hypothesis testing (Kleinschmidt et al., 1998; Murray et al., 2002; Yin et al., 2002; Haynes and Rees, 2005). Our approach revealed novel components of a network that is informative about the current content of visual perception and allowed the relevance of these components to be quantified in terms of their decoding accuracy.

LINKING SUBJECTIVE PERCEPTS WITH OBJECTIVE MEASUREMENTS OF BRAIN ACTIVATION

There is a long standing debate as to whether subjectively perceived qualities can be related to objectively observed neural activations. Several decoding studies have successfully shown that certain stimulus features can be decoded from BOLD-activity (for reviews see Rees et al., 2002; Tong et al., 2006; Tong and Pratte, 2012). However, it is not always clear in these studies if informative changes in brain activation relate to encoding of visual

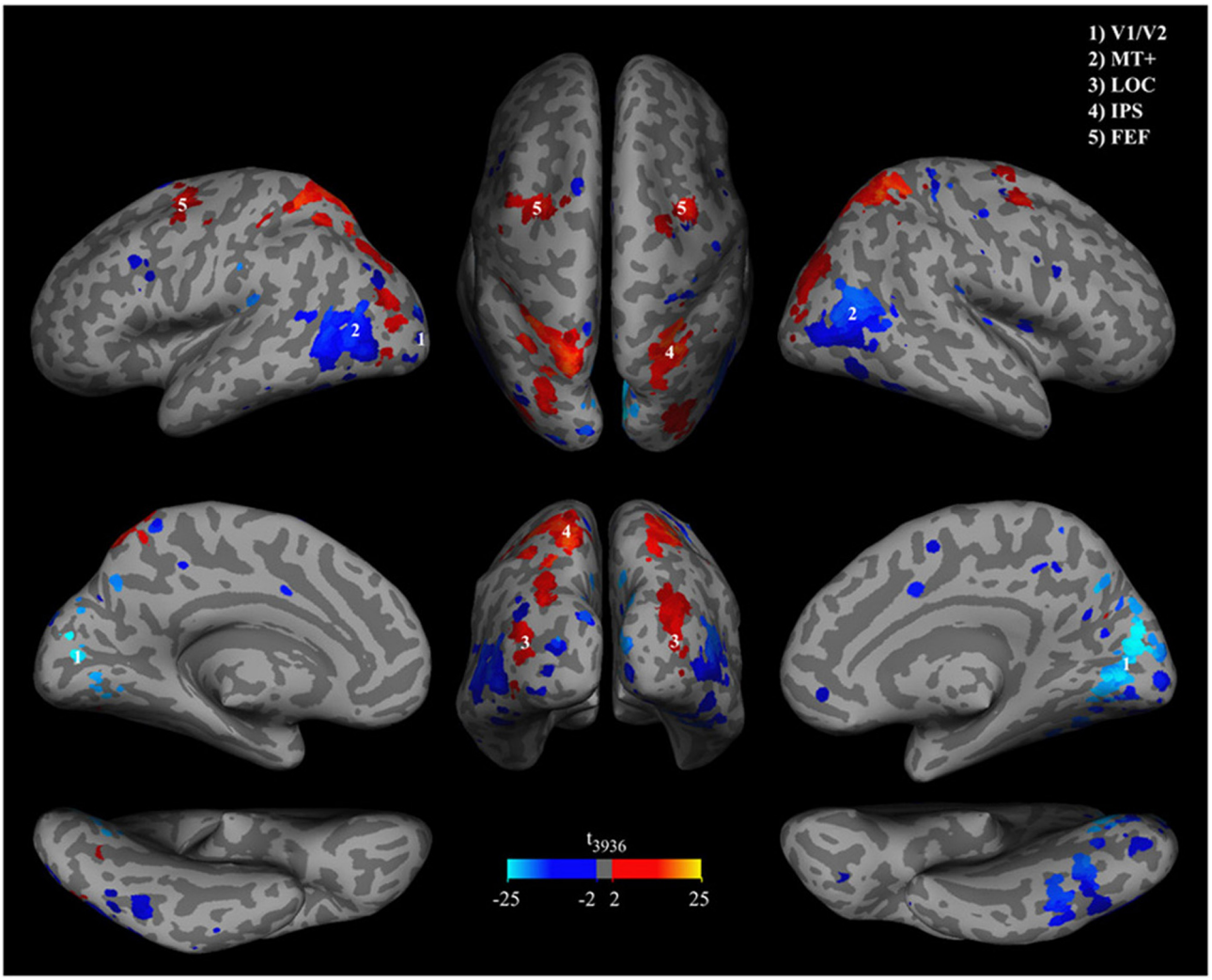


FIGURE 5 | Significance of classifier weights and activation differences. Brain patterns were extracted from the group classifiers of both offline and online experiments and overlaid on an MNI template. Brain regions were masked by *p*-values of classifier weights, thresholded at $\mathbf{p_w} < 0.1$. *P*-values were obtained in a permutation test. Hot colors indicate higher BOLD signal during object percepts and cool colors indicate lower BOLD signals during object percepts.

features or subjective percepts. A particularly strong case is made for decoding of subjective awareness when changes in the subjective state are uncorrelated with the physical stimulus (e.g., Grill-Spector et al., 2000; Chang et al., 2010). Here we employed a visual stimulus that is either perceived as a pair of vertically moving lines or as an occluded object moving behind a narrow slit. Importantly, earlier studies (Fendrich et al., 2005; Rieger et al., 2007) demonstrated that eye movements did not contribute to the formation of the object percept. This implies that the changes in brain activation we tracked reflect brain processes related to changes in subjective experiences rather than changes in the physical stimulus. We show to our knowledge for the first time that spontaneous changes in subjective, stimulus independent experiences can be tracked online with high accuracy, good temporal resolution (0.5 Hz sampling rate), and over an extended time period (>1 h) using real time fMRI. The high decoding accuracies we obtained, with up to 94% correct predictions, accentuate the strength of the link between neural network states and states of subjective perception.

The informative patterns of brain activations we found with our approach generalized well over subjects. This is not generally the case and may depend on the spatial resolution of the attempted analysis (e.g., Haxby et al., 2001) or the measurement technique employed (Fazli et al., 2009). We found that functional MRI combined with normalization to a standard brain and slight smoothing is a successful strategy for revealing informative networks at the scale of the full brain that transfer well between subjects. Such networks likely reflect general mechanisms underlying perceptual representation. The relative importance of individual vs. general (between-subject) features of brain organization and brain processing strategies can be estimated by assuming that the accuracy obtained with an individual

classifier reflects maximum obtainable performance, given a particular classification approach. The relative reduction in accuracy produced by the between-subject generalization then indicates the importance of individual features. In our investigation this reduction was statistically significant but almost all of the decoding accuracies achieved with left-out subjects were well above chance level. This is important because it indicates that patterns generalizing over subjects have a meaningful physiological interpretation. Together these results indicate that our data driven approach can be used in combination with ambiguous stimuli to reveal brain networks that are tightly linked to subjective percepts.

BRAIN NETWORKS WITH DECREASED BOLD-ACTIVITY DURING INTEGRATED OBJECT PERCEPTS

Analyzing the importance of individual brain voxels for the discrimination of the line vs. object percepts revealed an extended posterior brain network in which activation in lower tier sensory brain areas decreases during object percepts and activation in higher tier brain areas tends to increase. The results from our data driven approach confirm and considerably extend previous studies.

In concordance with earlier studies using similar ambiguous stimuli (Murray et al., 2002; Fang et al., 2008; de-Wit et al., 2012) we find a reduction of brain activity in early retinotopic visual areas, that represent the physical stimulus features such as line length and orientation. Concordant with de-Wit et al. (2012) we observe that these activation reductions extend around calcarine sulcus into the depth of the hemispheric cleft. In our study we also find reduced activation during object percepts in an area on the lateral surface, in the ascending limb of the inferior temporal sulcus, a reliable landmark for human motion sensitive area MT+ (Dumoulin et al., 2000) and in the vicinity of the fusiform gyrus, considered to be involved in higher level visual object representations (Ishai et al., 2000). Murray et al. (2002) argue that the reduction of activation in V1 during integrated object percepts is in accord with predictive coding theory (Mumford, 1992; Rao and Ballard, 1999). In this theory, higher tier areas convey a prediction of the features of the object represented to the lower tier areas which signal the difference between the prediction and the actual stimulus. This error signal is fed forward to the higher tier areas to adjust the object representation. The smaller the error between the prediction from higher tier areas and the stimulus representation in early sensory areas, the lower is their activation. Murray et al. (2002) suggested that during integrated object percepts higher tier brain areas predict the representation of the sensory stimulus in earlier areas and, as a consequence, the activation drops in lower tier areas. The reduced activation in hMT+ during object percepts fits with this interpretation. In earlier studies (Fendrich et al., 2005) we could show that during object percepts, the brain derives knowledge of horizontal object motion and could predict the movement of the integrated moving object from the movement of the segments. Similar to our results Castelo-Branco et al. (2002) found increased MT+ activity while the movement of two plaid stimuli was perceived as belonging to different objects and decreased activity when they were perceived as one integrated object. However, more detailed analysis reveals two effects that may require further elaboration of the predictive coding explanations of the neural deactivation patterns. First, we find reduction of brain activation during object percepts in ventral brain areas which are considered object selective and located relatively high in the processing hierarchy, according to standard theories of visual processing (Ungerleider and Mishkin, 1982). Conversely, within the predictive coding framework, the reduction of activation during object percepts suggests that these brain areas are located relatively low in the processing hierarchy. Second, as de-Wit et al., 2012 already pointed out, the reduction of activation in early visual cortex may extend beyond the retinotopic representation of the stimulus.

BRAIN NETWORKS WITH INCREASED BOLD-ACTIVITY DURING INTEGRATED OBJECT PERCEPTS

In concordance with Fang et al. (2008) we find increased brain activity bilaterally on the lateral surface of the occipital lobe during object percepts. This brain area is localized in the anatomical region of the LOC, which has been linked to the construction of object shape (Kourtzi and Kanwisher, 2000). The other areas with increased information were located toward and in parietal cortex. The function of the parietal areas is less clear. Recent studies suggest a functional role of parietal areas in the grouping of sequentially presented elements into coherent object percepts. For example, Zaretskaya et al. (2013) reported increased activation in anterior intraparietal sulcus (aIPS) during grouping of local elements into objects, and Peltier et al. (2007) reported LOC and IPS activation during haptic shape perception in which the sequential haptic information about shape has to be integrated into a coherent object percept. These parietal activations are not unique to the visual modality. Cusack (2005), using an auditory streaming paradigm, reported increased activity in IPS during object integration and hypothesized that the intraparietal area is generally involved in structuring sensory information. Moreover, we consider it highly unlikely that the parietal activation patterns can be explained by eye movements alone. In two earlier studies (Fendrich et al., 2005; Rieger et al., 2008) we characterized eye movements during the line vs. object percepts using highly accurate eye tracking methods and found only a small difference in the eye movement patterns between lines and object percepts.

Together, the results from our data driven approach suggest a working hypothesis: the LOC together with parietal brain areas are involved in constructing the percept of the integrated object. The reduction of activation in early visual areas during object percepts might be an indication of an interaction between higher and lower tier brain areas in which the former predict the representations in the latter. However, the spatial extent of the activation reduction in early retinotopic cortex and the function of ventral object selective cortices in the visual hierarchy require further evaluation.

REAL TIME vs. OFFLINE fMRI-ANALYSIS

Real-time analysis allowed us to feed the decoded percepts back to the subject. For such biofeedback experiments it is necessary that the classification method provides high generalization performance and that it requires little training data. We found that the average accuracies in the online and offline experiments were comparable, with a slight tendency for higher accuracies

in the online experiment. Theoretically, the online classification should perform worse than the cross-validation method because it uses a smaller training set. Subsequent cross-validation analysis of the online data confirmed that this offline approach can further increase the accuracy obtained with the online data set. We speculate that the reason of the accuracy advantage in the online experiment is due to the feedback which may have increased the subject's sense of agency. Receiving a feedback of the current conscious perceptual state could reduce mental fatigue and thereby reduce noise, which is essential for good classification performance. Another factor that may have contributed to the increased accuracy is the longer durations of the perceptual intervals which may be better reflected in the slow BOLD-response. The observation that most classification errors occurred around the time of perceptual switches supports this assumption. Studies indicate that subjects can volitionally control the switching of the percepts (van Ee et al., 2005; Kornmeier et al., 2009). Due to the delay of the BOLD-response it takes several seconds until the feedback signals a switch in perception. Although the subject's informal reports indicated that they could deal well with the delay, too short perceptual intervals could have led to more discordant feedback. This may have motivated our subjects to volitionally control and extend the duration of the percepts in order to increase the synchronization of the percept with the delayed feedback. Thus, the feedback may have increased perceptual inertia or cognitive bias leading to prolonged percepts in the online experiment. Further studies are necessary to characterize the neural basis of this biofeedback effect.

AMBIGUITY AND STREAM SEGREGATION

Ambiguous stimuli can induce bistable perceptual organization that switches between categorically different object percepts despite constant visual input. This property makes them ideal to investigate processes of perceptual organization independent of potential confounds by changing physical stimuli. Here we outlined a data driven classification approach to reveal brain networks underlying perceptual organization. The classification accuracy provides an intuitive measure of the relevance of the observed effects. We would like to point out, that the approach is not restricted to the visual domain. In the auditory domain, multistable stimuli, such as two tone sequences (Bregman, 1990; Gutschalk et al., 2005; Deike et al., 2010), play an important role in the characterization of processes underlying stream organization. We suggest that the approach presented here, can be used to investigate the neural basis of auditory object formation with ambiguous auditory stimuli. Support for the assumption that this is possible comes from studies showing that the dynamics of auditory and visual perceptual switches are similar. In both modalities percept durations are gamma distributed and have similar lengths (Pressnitzer and Hupé, 2006; Denham et al., 2012). Furthermore, the different modalities seem to share common brain regions for perceptual organization (Cusack, 2005).

ACKNOWLEDGMENTS

This work was supported by the EU project ECHORD 231143, Land Sachsen-Anhalt Grant FKZ MK48-2009/003, Niedersächsisches Vorab of the Volkswagen Foundation and the Ministry of Science and Culture of Lower Saxony as part of the Interdisciplinary Research Center on Critical Systems Engineering for Socio-Technical Systems, DFG SFB-TR31, and BMBF Forschungscampus *STIMULATE*.

REFERENCES

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Brouwer, G. J., and van Ee, R. (2007). Visual cortex allows prediction of perceptual states during ambiguous structure-from-motion. *J. Neurosci.* 27, 1015–1023. doi: 10.1523/JNEUROSCI.4593-06.2007

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* 2, 121–167. doi: 10.1023/A:1009715923555

Castelo-Branco, M., Formisano, E., Backes, W., Zanella, F., Neuenschwander, S., Singer, W., et al. (2002). Activity patterns in human motion-sensitive areas depend on the interpretation of global motion. *Proc. Natl. Acad. Sci. U.S.A.* 99, 13914–13919. doi: 10.1073/pnas.202049999

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nat. Neurosci.* 13, 1428–1432. doi: 10.1038/nn.2641

Cherkassky, V., and Mulier, F. M. (1998). *Learning from Data: Concepts, Theory, and Methods*. New York, NY: John Wiley & Sons.

Cox, D. D., and Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) brain reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* 19, 261–270. doi: 10.1016/S1053-8119(03)00049-1

Cusack, R. (2005). The intraparietal sulcus and perceptual organization. *J. Cogn. Neurosci.* 17, 641–651. doi: 10.1162/0898929053467541

Deike, S., Scheich, H., and Brechmann, A. (2010). Active stream segregation specifically involves the left human auditory cortex. *Hear. Res.* 265, 30–37. doi: 10.1016/j.heares.2010.03.005

Denham, S., Bendixen, A., Mill, R., Toth, D., Wennekers, T., Coath, M., et al. (2012). Characterising switching behaviour in perceptual multi-stability. *J. Neurosci. Methods* 210, 79–92. doi: 10.1016/j.jneumeth.2012.04.004

de-Wit, L. H., Kubilius, J., Wagemans, J., and Op de Beeck, H. P. (2012). Bistable Gestalts reduce activity in the whole of V1, not just the retinotopically predicted parts. *J. Vis.* 12:12. doi: 10.1167/12.11.12

Dumoulin, S. O., Bittar, R. G., Kabani, N. J., Baker, C. L., Le Goualher, G., Pike, G. B., et al. (2000). A new anatomical landmark for reliable identification of human area V5/MT: a quantitative analysis of sulcal patterning. *Cereb. Cortex* 10, 454–463. doi: 10.1093/cercor/10.5.454

Fang, F., Kersten, D., and Murray, S. O. (2008). Perceptual grouping and inverse fMRI activity patterns in human visual cortex. *J. Vis.* 8:2. doi: 10.1167/8.7.2

Fazli, S., Popescu, F., Danószy, M., Blankertz, B., Müller, K. R., and Grozea, C. (2009). Subject-independent mental state classification in single trials. *Neural Netw.* 22, 1305–1312. doi: 10.1016/j.neunet.2009.06.003

Fendrich, R., Rieger, J. W., and Heinze, H. J. (2005). The effect of retinal stabilization on anorthoscopic percepts under free-viewing conditions. *Vision Res.* 45, 567–582. doi: 10.1016/j.visres.2004.09.025

Freeman, E. D., Sterzer, P., and Driver, J. (2012). fMRI correlates of subjective reversals in ambiguous structure-from-motion. *J. Vis.* 12:6. doi: 10.1167/12.6.35

Good, P. I. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. New York, NY: Springer.

Grill-Spector, K., Kushnir, T., Hendler, T., and Malach, R. (2000). The dynamics of object-selective activation correlate with recognition performance in humans. *Nat. Neurosci.* 3, 837–843. doi: 10.1038/77754

Gutschalk, A., Micheyl, C., Melcher, J. R., Rupp, A., Scherg, M., and Oxenham, A. J. (2005). Neuromagnetic correlates of streaming in human auditory cortex. *J. Neurosci.* 25, 5382–5388. doi: 10.1523/JNEUROSCI.0347-05.2005

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430. doi: 10.1126/science.1063736

Haynes, J. D., and Rees, G. (2005). Predicting the stream of consciousness from activity in human visual cortex. *Curr. Biol.* 15, 1301–1307. doi: 10.1016/j.cub.2005.06.026

Hollmann, M., Mönch, T., Mulla-Osman, S., Tempelmann, C., Stadler, J., and Bernarding, J. (2008). A new concept of a unified parameter management, experiment control, and data analysis in fMRI: application to

real-time fMRI at 3T and 7T. *J. Neurosci. Methods* 175, 154–162. doi: 10.1016/j.jneumeth.2008.08.013

Hollmann, M., Rieger, J. W., Baecke, S., Lützkendorf, R., Müller, C., Adolf, D., et al. (2011). Predicting decisions in human social interactions using real-time fMRI and pattern classification. *PLoS ONE* 6:e25304. doi: 10.1371/journal.pone.0025304

Ishai, A., Ungerleider, L. G., Martin, A., and Haxby, J. V. (2000). The representation of objects in the human occipital and temporal cortex. *J. Cogn. Neurosci.* 12, 35–51. doi: 10.1162/089892900564055

Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature* 452, 352–355. doi: 10.1038/nature06713

Kleinschmidt, A., Büchel, C., Zeki, S., and Frackowiak, R. S. J. (1998). Human brain activity during spontaneously reversing perception of ambiguous figures. *Proc. Biol. Sci.* 265, 2427–2433. doi: 10.1098/rspb.1998.0594

Kondo, H. M., and Kashino, M. (2009). Involvement of the thalamocortical loop in the spontaneous switching of percepts in auditory streaming. *J. Neurosci.* 29, 12695–12701. doi: 10.1523/JNEUROSCI.1549-09.2009

Kornmeier, J., Hein, C. M., and Bach, M. (2009). Multistable perception: when bottom-up and top-down coincide. *Brain Cogn.* 69, 138–147. doi: 10.1016/j.bandc.2008.06.005

Kourtzi, Z., and Kanwisher, N. (2000). Cortical regions involved in perceiving object shape. *J. Neurosci.* 20, 3310–3318.

LaConte, S. M., Peltier, S. J., and Hu, X. P. P. (2007). Real-time fMRI using brain-state classification. *Hum. Brain Mapp.* 28, 1033–1044. doi: 10.1002/hbm.20326

Mourão-Miranda, J., Bokde, A. L. W., Born, C., Hampel, H., and Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. *Neuroimage* 28, 980–995. doi: 10.1016/j.neuroimage.2005.06.070

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol. Cybern.* 66, 241–251. doi: 10.1007/BF00198477

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P., and Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15164–15169. doi: 10.1073/pnas.192579399

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* 10, 424–430. doi: 10.1016/j.tics.2006.07.005

Peltier, S., Stilla, R., Mariola, E., LaConte, S., Hu, X., and Sathian, K. (2007). Activity and effective connectivity of parietal and occipital cortical regions during haptic shape perception. *Neuropsychologia* 45, 476–483. doi: 10.1016/j.neuropsychologia.2006.03.003

Pressnitzer, D., and Hupé, J. M. (2006). Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Curr. Biol.* 16, 1351–1357. doi: 10.1016/j.cub.2006.05.054

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

Rees, G., Kreiman, G., and Koch, C. (2002). Neural correlates of consciousness in humans. *Nat. Rev. Neurosci.* 3, 261–270. doi: 10.1038/nrn783

Rieger, J. W., Grüschow, M., Heinze, H. J., and Fendrich, R. (2007). The appearance of figures seen through a narrow aperture under free viewing conditions: effects of spontaneous eye motions. *J. Vis.* 7:10. doi: 10.1167/7.6.10

Rieger, J. W., Reichert, C., Gegenfurtner, K. R., Noesselt, T., Braun, C., Heinze, H. J., et al. (2008). Predicting the recognition of natural scenes from single trial MEG recordings of brain activity. *Neuroimage* 42, 1056–1068. doi: 10.1016/j.neuroimage.2008.06.014

Tong, F., Meng, M., and Blake, R. (2006). Neural bases of binocular rivalry. *Trends Cogn. Sci.* 10, 502–511. doi: 10.1016/j.tics.2006.09.003

Tong, F., Nakayama, K., Vaughan, J. T., and Kanwisher, N. (1998). Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron* 21, 753–759. doi: 10.1016/S0896-6273(00)80592-9

Tong, F., and Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annu. Rev. Psychol.* 63, 483–509. doi: 10.1146/annurev-psych-120710-100412

Ungerleider, L. G., and Mishkin, M. (1982). "Two cortical visual systems," in *Analysis of Visual Behavior*, eds D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield (Cambridge: MIT Press), 549–586.

van Ee, R., van Dam, L. C. J., and Brouwer, G. J. (2005). Voluntary control and the dynamics of perceptual bi-stability. *Vision Res.* 45, 41–55. doi: 10.1016/j.visres.2004.07.030

Vapnik, V. N. (1998). *Statistical Learning Theory.* New York, NY: John Wiley & Sons.

von Helmholtz, H. (1867). *Handbuch der Physiologischen Optik.* Hamburg: Voss.

Weiskopf, N. (2012). Real-time fMRI and its application to neurofeedback. *Neuroimage* 62, 682–692. doi: 10.1016/j.neuroimage.2011.10.009

Weiskopf, N., Sitaram, R., Josephs, O., Veit, R., Scharnowski, F., Goebel, R., et al. (2007). Real-time functional magnetic resonance imaging: methods and applications. *Magn. Reson. Imaging* 25, 989–1003. doi: 10.1016/j.mri.2007.02.007

Yin, C., Shimojo, S., Moore, C., and Engel, S. A. (2002). Dynamic shape integration in extrastriate cortex. *Curr. Biol.* 12, 1379–1385. doi: 10.1016/S0960-9822(02)01071-0

Zaretskaya, N., Anstis, S., and Bartels, A. (2013). Parietal cortex mediates conscious perception of illusory Gestalt. *J. Neurosci.* 33, 523–531. doi: 10.1523/JNEUROSCI.2905-12.2013

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

Received: 28 February 2014; accepted: 02 May 2014; published online: 23 May 2014.
*Citation: Reichert C, Fendrich R, Bernarding J, Tempelmann C, Hinrichs H and Rieger JW (2014) Online tracking of the contents of conscious perception using real-time fMRI. Front. Neurosci. **8**:116. doi: 10.3389/fnins.2014.00116*
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 06 June 2014
doi: 10.3389/fnins.2014.00119

Evaluating auditory stream segregation of SAM tone sequences by subjective and objective psychoacoustical tasks, and brain activity

Lena-Vanessa Dolležal[1], André Brechmann[2], Georg M. Klump[1]* and Susann Deike[2]

[1] Animal Physiology and Behavior Group, Department for Neuroscience, School for Medicine and Health Sciences, Center of Excellence "Hearing4all," Carl von Ossietzky University Oldenburg, Oldenburg, Germany
[2] Special Lab Non-invasive Brain Imaging, Leibniz Institute for Neurobiology, Magdeburg, Germany

Edited by:
Elyse S. Sussman, Albert Einstein College of Medicine, USA

Reviewed by:
Andrew R. Dykstra, University of Heidelberg, Germany
Pierre Divenyi, Veterans Affairs Northern California Health Care System, USA
Makio Kashino, NTT Corporation, Japan

***Correspondence:**
Georg M. Klump, Animal Physiology and Behavior Group, Department for Neuroscience, School for Medicine and Health Sciences, Center of Excellence "Hearing4all," Carl von Ossietzky University Oldenburg, Carl von Ossietzky Str. 9-11, D-26129 Oldenburg, Germany
e-mail: georg.klump@uni-oldenburg.de

Auditory stream segregation refers to a segregated percept of signal streams with different acoustic features. Different approaches have been pursued in studies of stream segregation. In psychoacoustics, stream segregation has mostly been investigated with a subjective task asking the subjects to report their percept. Few studies have applied an objective task in which stream segregation is evaluated indirectly by determining thresholds for a percept that depends on whether auditory streams are segregated or not. Furthermore, both perceptual measures and physiological measures of brain activity have been employed but only little is known about their relation. How the results from different tasks and measures are related is evaluated in the present study using examples relying on the ABA- stimulation paradigm that apply the same stimuli. We presented A and B signals that were sinusoidally amplitude modulated (SAM) tones providing purely temporal, spectral or both types of cues to evaluate perceptual stream segregation and its physiological correlate. Which types of cues are most prominent was determined by the choice of carrier and modulation frequencies (f_{mod}) of the signals. In the subjective task subjects reported their percept and in the objective task we measured their sensitivity for detecting time-shifts of B signals in an ABA- sequence. As a further measure of processes underlying stream segregation we employed functional magnetic resonance imaging (fMRI). SAM tone parameters were chosen to evoke an integrated (1-stream), a segregated (2-stream), or an ambiguous percept by adjusting the f_{mod} difference between A and B tones (Δf_{mod}). The results of both psychoacoustical tasks are significantly correlated. BOLD responses in fMRI depend on Δf_{mod} between A and B SAM tones. The effect of Δf_{mod}, however, differs between auditory cortex and frontal regions suggesting differences in representation related to the degree of perceptual ambiguity of the sequences.

Keywords: auditory scene analysis, amplitude modulation, temporal and spectral cues, time shift detection, BOLD response, fMRI

INTRODUCTION

In everyday life, the auditory system organizes acoustic signals based on similarities and differences in their sound features (Bregman, 1990). Especially in complex acoustic scenes, spectral or temporal stimulus parameters affect the grouping and segregation processes in assigning sounds to different sources. Sounds from one source are perceived as one coherent auditory stream. Studies of auditory stream segregation commonly applied the ABA- paradigm (e.g., Van Noorden, 1975; Moore and Gockel, 2002, 2012). The amount of stream segregation depends on the differences in sound features, i.e., the physical differences between A and B signals. It has been proposed that these differences will lead to the representation of the signals assigned to the separate streams by separate populations of neurons being differentially activated in time (e.g., Fishman et al., 2001; Elhilali et al., 2009). Furthermore, the separate representation of the A and B signals can be observed already at the first stages of auditory processing (i.e., the cochlear nucleus) as Pressnitzer et al., 2008 demonstrated. Previous studies evaluated the auditory streaming percept and the neural mechanisms underlying the perceptual organization of sounds elicited by spectral (e.g., Van Noorden, 1975; Fishman et al., 2001, 2004; Bee and Klump, 2004, 2005; Deike et al., 2004, 2010; Micheyl et al., 2005; Micheyl and Oxenham, 2010) and temporal differences (Grimault et al., 2002; Vliegen and Oxenham, 1999; Roberts et al., 2002; Gutschalk et al., 2007; Itatani and Klump, 2011; Dolležal et al., 2012b) between A and B signals. These studies have provided evidence how auditory streaming is affected by a variety of simple features. Although applying a common paradigm, the outcome of these studies may also depend on the psychoacoustical task that was employed or on the measure that was used for evaluating the segregation of the streams. The majority of

the psychoacoustical tasks relied on a subjective perceptual judgment (e.g., Van Noorden, 1975; Vliegen and Oxenham, 1999; Grimault et al., 2002; Roberts et al., 2002; Micheyl et al., 2005; Gutschalk et al., 2007; Micheyl and Oxenham, 2010; Dolležal et al., 2012b). In the subjective task, subjects simply report their streaming percept. Few studies employed objective tasks (Van Noorden, 1975; Neff et al., 1982; Vliegen et al., 1999; Cusack and Roberts, 2000; Roberts et al., 2002; Micheyl and Oxenham, 2010; Thompson et al., 2011). In the objective task, the subject's perceptual threshold is determined using stimulus conditions in which threshold sensitivity is enhanced by one perceptual organization and hampered by the other (i.e., 1- and the 2-stream percept). Thus, in the objective task the streaming percept is inferred from the measured perceptual sensitivity. The different measures that were used range from the evaluation of perception to the assessment of the brain activity by applying invasive or non-invasive measurement techniques. Since few studies compared results obtained with different tasks (Micheyl and Oxenham, 2010) and measures (Gutschalk et al., 2007; Wilson et al., 2007), we have only little evidence how well these results are correlated.

Here, we investigate the correlation of the extent of stream segregation for sinusoidally amplitude modulated (SAM) A and B signals across two different psychoacoustical tasks and across two different measures (i.e., subjective psychoacoustical task and fMRI) providing a comprehensive approach to auditory stream segregation. The comparison across different psychoacoustical tasks involves an objective and a subjective task presenting signals with identical sound features to the same subjects. We propose that the thresholds obtained in the objective task are correlated with the subjective percept of stream segregation indicating that either task allows measuring the amount of perceptual stream segregation. Since any salient difference between sequential signals may elicit stream segregation (Moore and Gockel, 2002, 2012), we expect that a correlation will be found irrespective whether temporal or spectral cues can be utilized to differentiate between A and B signals. As is outlined in the methods below, SAM signals offer temporal, spectral, or both types of cues for stream segregation dependent on the modulation frequency (f_{mod}) and carrier frequency (f_c), respectively. By an appropriate choice of f_c and f_{mod} for the SAM tone stimulus sequences the amount of stream segregation between A and B SAM tones elicited by spectral cues and temporal cues can be varied (Dolležal et al., 2012a).

The comparison across measures includes the combination of the subjective task with fMRI. Previous human fMRI and MEG studies using either spectral (Deike et al., 2004, 2010; Gutschalk et al., 2005; Snyder et al., 2006; Wilson et al., 2007) or temporal (Gutschalk et al., 2007) differences between A and B stimuli consistently showed an increase of activity throughout the auditory cortex combined with a change of the dominant percept from 1-stream to 2-stream with increasing difference between stimuli. Based on these results we propose that A and B SAM tones show the same Δf_{mod} dependent activity in auditory cortex irrespective of the type of cue (i.e. spectral vs. temporal). The second goal pursued in obtaining fMRI activity measurements was to find further evidence of the specific involvement of regions outside the auditory cortex in stream segregation which is still an open question in auditory streaming research. Cusack (2005), e.g., found that the intraparietal sulcus was differentially involved depending on the perceptual organization of physically identical stimuli in perceptual ambiguous sequences with a stronger BOLD activation for the segregated two-stream compared to the integrated one-stream percept. Such a segregation specific activation in the absence of physical differences was not found by Dykstra et al. (2011) when using intracranial EEG in neurosurgical patients with epilepsy. However, he observed a Δf dependent activity in middle temporal gyrus, pre- and post-central gyri, inferior and middle frontal gyri, and the supra-marginal gyrus. The human fMRI study by Kondo and Kashino (2009) found neural correlates of perceptual switching in the posterior insula, thalamus, and supra-marginal gyrus. Finally, some other studies did not describe evidence for the involvement of regions outside auditory areas in stream segregation (e.g., Wilson et al., 2007).

MATERIALS AND METHODS

SUBJECTS

Psychoacoustical measurements

Six human subjects (age 25–44 years, mean age 30 years, five females, including the first author) participated in two main experiments (Experiment 1 and 2; ABA- sequences) comparing subjective and objective psychoacoustical measures of stream segregation. All subjects had normal audiograms, with absolute pure tone thresholds <20 dB hearing level in the range from 0.25–10 kHz. Four of the subjects had previous experience with psychoacoustic experiments. In a control experiment (applying the conditions of Experiment 1 in B-only sequences) four of the subjects (age 25–42 years, mean age 30 years, three females, including the first author) participated. All experiments were undertaken with the understanding and written informed consent of each subject, following the Code of Ethics of the World Medical Association (Declaration of Helsinki). The experiments were approved by the local ethics committee of the University of Oldenburg. In addition to these psychoacoustical measurements in quiet, the subjective streaming percept was determined during fMRI (see fMRI measurements).

fMRI measurements

In Experiment 1, 13 subjects (age 20–31 years, mean age 26 years, five female) and in Experiment 2, 10 subjects (age 20–33 years, mean age 26 years, four female) participated. One subject participated in both experiments. Due to technical problems, psychophysical data of one subject in Experiment 2 are missing and only the data of 9 subjects were analyzed. All but one subject were right-handed (Edinburgh Handedness Inventory; laterality quotient $\geq +45$) and this one subject was ambidextrous (laterality quotient: 18). All showed a language laterality toward the left hemisphere tested as described in Bethmann et al. (2007). The subjects gave written informed consent to the study that was approved by the Ethics Committee of the University of Magdeburg. Five additional participants were excluded from the final analysis: one because of more than five missing responses and four because of head movements during

the fMRI-measurement that were stronger than 2.3 mm translation and/or 2.3° rotation.

APPARATUS, STIMULI, AND PROCEDURE

Stimuli

In the present study ABA- sequences (the dash indicates a silent interval of the same duration as the signal duration) were presented that consisted of fully sinusoidally amplitude modulated tones (SAM). ABA- sequences are commonly applied to determine the amount of stream segregation by varying the physical difference between A and B signals (Van Noorden, 1975). A and B signals with small physical differences are perceptually grouped into a single sequence (i.e., 1-stream percept) with a galloping rhythm (i.e., ABA-ABA-ABA-. . .), whereas A and B signals with large physical differences are perceptually segregated to two streams (i.e., 2-stream percept) with different isochronous rhythms (i.e., A-A-A-A-A-A-. . . and -B---B---B--. . .). For A and B signals with intermediate physical differences subjects may have an ambiguous percept, that is characterized by a switching between the 1- and the 2-stream percept (e.g., Moore and Gockel, 2002, 2012).

The SAM tones were digitally synthesized in Matlab (Version 7.1) at a sampling frequency of 44.1 kHz and produced by a Hammerfall DSP (Multiface II, RME). These signals (10 ms raised cosine rise/fall) had a duration of 125 ms and were presented at an overall presentation level of 70 dB SPL with a tone repetition time (TRT) of 250 ms. SAM tones have the advantage that the carrier frequency (f_c) and the modulation frequency (f_{mod}) can be adjusted in such a way that, depending on the parameter values and the auditory filter bandwidth (Kohlrausch et al., 2000), they provide either temporal, spectral, or both types of cues for stream segregation (Dolležal et al., 2012a). Dolležal et al. (2012a) used a computational model of the auditory periphery to calculate excitation pattern differences of A and B SAM tones and estimate spectral stream segregation thresholds based on these differences. If the observed thresholds were below the prediction based on spectral cues alone (i.e., could not be explained by spectral cues), they concluded that only temporal cues were relevant for the segregated percept. If the observed thresholds were similar or higher than the thresholds predicted on the basis of spectral cues, it was concluded spectral cues could provide a basis for perceptual stream segregation (for more details see Dolležal et al., 2012a). **Table 1** summarizes the different parameter settings and highlights conditions in which spectral cues alone could explain stream segregation. For the remaining parameter settings, spectral cues are unlikely to explain stream segregation.

Here, ABA- sequences consisted of SAM tones that had the same carrier frequency (f_c) but different modulation frequencies (f_{mod}). Note the f_{mod} of the A SAM tones ($f_{\mathrm{mod\ A}}$) was always lower than the f_{mod} of the B SAM tones ($f_{\mathrm{mod\ B}}$). For the psychoacoustical tasks and the fMRI measurements, in the two different experiments the effect of $f_{\mathrm{mod\ A}}$ (Experiment 1) or the effect of the f_c (Experiment 2) on stream segregation was analyzed. In Experiment 1 SAM tones had an f_c of 1 kHz and an $f_{\mathrm{mod\ A}}$ of either 100 or of 300 Hz and in Experiment 2 SAM tones had an f_c of either 1 or of 4 kHz and an $f_{\mathrm{mod\ A}}$ of 100 Hz. For each condition three f_{mod} differences between A and B SAM tones (Δf_{mod}) were chosen to evoke a 1-stream, a 2-stream and an ambiguous percept for the tested conditions, respectively (see **Table 1**). The value of Δf_{mod} was adjusted based on the study by Dolležal et al. (2012a). Dolležal et al. (2012a) also presented ABA SAM tone sequences, but they used either a TRT of 125 ms or a TRT of 375 ms for SAM tones of 125 ms duration. Based on their results the preset study chose for both experiments Δf_{mod} stimulus conditions that enable a comparison of stream segregation elicited by spectral and non-spectral cues. In Experiment 1 such a comparison can be made at the medium Δf_{mod} condition and in Experiment 2 at the large Δf_{mod} condition (**Table 1**; see Dolležal et al., 2012a). The results obtained in the present study were compared across two different psychoacoustical tasks and across two different measures (i.e., subjective psychoacoustical task and fMRI) for all Δf_{mod} stimulus conditions.

Psychoacoustical measurements

Psychoacoustical data were obtained in two different locations (Oldenburg and Magdeburg). In Oldenburg all subjects participated in the subjective and in the objective task in quiet in a sound-attenuating chamber (IAC, Industrial Acoustics Company, Mini 250). The stimuli were presented diotically with calibrated headphones (Sennheiser HDA 200). In Magdeburg, subjects participated in the subjective task during fMRI. Written instructions and additional verbal explanations, if necessary, were given to the subjects before the beginning of the tasks.

Table 1 | Stimulus conditions for Experiment 1 and Experiment 2.

f_c [kHz]	$f_{\mathrm{mod\ A}}$ [Hz]	Small Δf_{mod}[Hz]	SmallΔ f_{mod}[oct]	Medium Δf_{mod} [Hz]	Medium Δf_{mod}[oct]	Large Δf_{mod} [Hz]	Large Δf_{mod}[oct]
EXPERIMENT 1, EFFECTS OF $f_{\mathrm{mod\ A}}$							
1	100	33	0.415	100	1.0	300	2.0
1	300	29	0.135	100	0.415	300	1.0
EXPERIMENT 2, EFFECTS OF f_c							
1	100	33	0.415	100	1.0	300	2.0
4	100	33	0.415	100	1.0	300	2.0

For each experiment, the carrier frequency (f_c), the modulation frequency of the A SAM tone ($f_{mod\ A}$) and the modulation frequency difference between A and B SAM tones (Δf_{mod}) are indicated. Δf_{mod} conditions that are printed in bold typeface show conditions in which stream segregation was can be explained on the basis of spectral cues (Dolležal et al., 2012a) estimated from a model calculating excitation pattern differences between the A and the B SAM tones in the inner ear. Spectral cues provided by Δf_{mod} conditions that are printed in regular typeface are unlikely to be sufficient for stream segregation.

Objective task in quiet. Subjects started the experiment with the objective task in quiet. To measure objectively the perceptual segregation of the A and B SAM tones, subjects performed a shift detection task (**Figure 1**) in a Go/NoGo experiment determining the detection of a time shifted B SAM tone in the ABA- sequence. Thresholds obtained with the shift detection task should be smaller for ABA- sequences that are perceptually integrated into one stream than for ABA- sequences that are perceptually segregated to two streams (e.g., Van Noorden, 1975). In the present study, subjects listened to the presentation of a repeated ABA- triplet without a time shifted B SAM tone. Within 1 to 7 s (randomized time interval) after subjects started a trial by pressing a button on the touch screen either a forward shifted B SAM tone (Go-stimulus) replaced the regular B SAM tone or no replacement took place and a regular B SAM tone was presented (NoGo-stimulus, 30% of trials). If subjects detected the Go-stimulus in time (response latency < 1 s) by pushing a button on a touch screen, a correct response (hit) was registered and a green light flashed. If the subjects missed the Go-stimulus, a miss was recorded and the next trial was automatically initiated. The Go-response in this complex time-shift detection task could be based on the evaluation of the time interval between the A SAM tone and the successive B SAM tone or on the time interval between two sequential B SAM tones. Responses to NoGo-stimuli (false alarms) were registered too. Hit and false alarm rates were used to calculate the sensitivity measure d′ (Green and Swets, 1966; see data analysis, psychoacoustics). For threshold estimation in each stimulus condition (**Table 1**) subjects had to complete a minimum of three sessions consisting of one obligatory training session and two subsequent test sessions (within each session a specific Go-stimulus was presented 10 times). A session lasted for about 20 min and consisted of eleven blocks of ten trials each. The first block of each session served as a warm-up block in which only the most salient Go-stimuli were presented. Each of the remaining ten blocks consisted of seven different Go-stimuli and three NoGo-stimuli that were presented in a random order. The Go-stimuli with a time shifted B SAM tone (step size 6.25 or 12.5 ms; i.e., 5 or 10% of the SAM tone duration) were chosen according to the method of constant stimuli. The range of the time shifts imposed on the B SAM tone was individually adjusted before each session to provide both sub-threshold and supra-threshold Go-stimuli. After each session a psychometric function was constructed relating the hits and misses of seven different Go-stimuli (different amounts of a time shifted B SAM tone) to d'-values (a measure of sensitivity for detecting the shift; see **Figure 2**). Between threshold sessions presenting different stimulus conditions a minimum pause of 5 min occurred. Within the objective task in quiet, the threshold estimation for the different stimulus conditions was randomized.

In addition to the objective task in quiet presenting ABA- sequences with time shifted B SAM tones, a control experiment

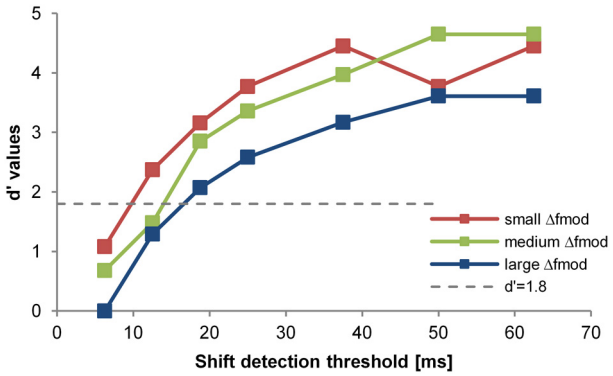


FIGURE 2 | Psychometric function of one subject for one stimulus condition (i.e., exp. 2, $f_c = 4$ kHz). The d'-value is plotted in relation to the shift of the B-signal in ms (x-axes). The differently colored lines and symbols show the different Δf_{mod} conditions tested (see legend).The threshold criterion of $d' = 1.8$ is indicated by the dotted gray line. The shift detection threshold ($d' = 1.8$) was interpolated between data points lying above and below that d'-value. The slight differences in largest d' values are due to different false alarm rates for the different Δf_{mod} conditions.

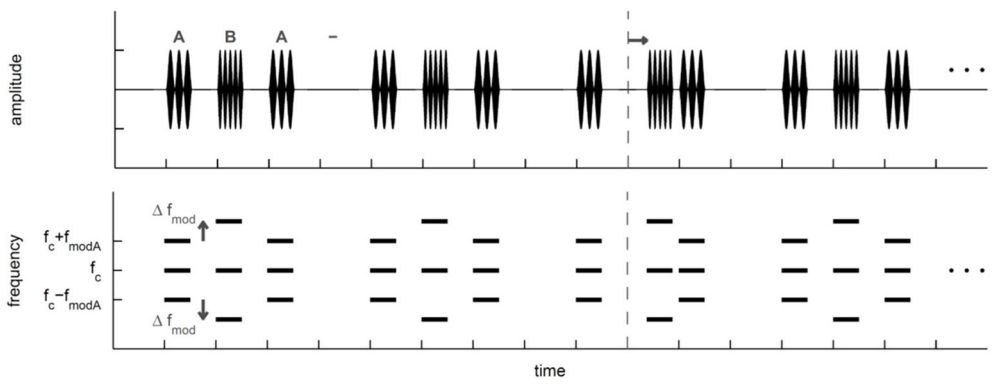


FIGURE 1 | Schematic view of the ABA-triplets presented in the objective task that relied on the detection of a time shifted B signal. In the third ABA- triplet a black arrow indicates the shift of the B signal, whereas the dashed line indicates the former position the un-shifted B signal. Top: Schematic temporal view of the ABA- triplets that were sinusoidally amplitude modulated (SAM) tones. A and B SAM tones had the same carrier frequency (f_c) but different modulation frequencies (f_{mod}). The f_{mod} of the B SAM tone was always larger than the f_{mod} of the A SAM tone ($f_{mod\ A}$). Here the f_{mod} difference between A and B SAM tones (Δf_{mod}) is schematically shown (see **Table 1** for exact values). Bottom: Schematic spectral view of the SAM ABA- triplets.

(for all stimulus conditions of Experiment 1) was conducted presenting B-only sequences. In B-only sequences only B SAM tones (omitting the A SAM tones) were presented (-B---B---B--) resulting in a TRT of 1000 ms. This experiment mimics a condition with a completely segregated percept, in which subjects solely rely on the stream of B SAM tones for the shift detection.

Subjective task in quiet. After performing in the objective task in quiet, subjects participated in the subjective task in quiet. Here, the same stimulus conditions as in the objective task were applied. ABA- sequences (15 s duration) of each stimulus condition were presented six times in randomized order. A pause of 45 s was introduced between the presentation of ABA-sequences of different f_c and $f_{\mathrm{mod\,A}}$. After the presentation of each ABA- sequence subjects were instructed to indicate their percept (e.g., 1- or 2-stream percept) on a touch screen (Elo, 1542L, 15″, Rear-Mount Touch-monitor). Then, the next ABA-sequence with another randomly chosen stimulus condition was initiated. Before starting the experiments in the subjective task in quiet, subjects attended a training session to familiarize with the task.

Subjective task during fMRI. During fMRI measurements, the subjects were presented with the same stimuli as in the subjective task in quiet. The duration of the stimulus sequences was increased to 16 s to adapt to the repetition time ($TR =$ 2000 ms) of the functional echo planar imaging (EPI) sequence. Each of the three conditions (small, medium and large Δf_{mod}) were presented 10 times for each $f_{\mathrm{mod\,A}}$ (100, 300 Hz) in Experiment 1 and for each f_c (1, 4 kHz) in Experiment 2, respectively, resulting in the presentation of 60 sequences per experiment. For each experiment, the order of the 60 sequences was pseudo-randomized with silence blocks of 16 s duration in between, which served as baseline condition. The stimuli were presented diotically via fMRI compatible headphones (Baumgart et al., 1998) at an individually adjusted, comfortable sound level, using Presentation (Neurobehavioral Systems Inc., San Francisco, USA). During the fMRI measurements, the subjects' heads were fixed with a cushion with attached earmuffs containing the headphones. Additionally, the subjects wore earplugs.

Prior to the fMRI measurements, the subjects received written instructions and additional verbal explanations if necessary. The subjects were asked to listen to the sound sequences and to indicate their percept at the end of each sequence by pressing the left button on a response panel with their right index finger when they perceived the SAM tones as one coherent stream, and the right button with their right middle finger when they perceived them as two separate streams. All button presses were recorded using Presentation (Neurobehavioral Systems Inc., San Francisco, USA) to test the perception of the SAM tone sequences under background scanner noise conditions. To familiarize the subjects with the sound sequences and the task, prior to the actual measurements, they were exposed to sequences, which most likely promote one or the other perceptual alternative, i.e., the 1-stream and the 2-stream percept, respectively.

fMRI measurements and data acquisition

The study was carried out on a 3 Tesla scanner (Siemens Trio; Erlangen, Germany) equipped with an eight channel head coil. A three-dimensional anatomical data set of the subject's brain (192 slices of 1 mm each) was obtained before the functional measurement. Additionally, before each functional run an Inversion-Recovery-Echo-Planar-Imaging (IR-EPI) with the identical geometry as in the functional measurement was acquired. Functional volumes were collected using a continuous EPI sequence (echo time $TE = 30$ ms; $TR = 2000$ ms; flip angle $= 80°$; 32 slices; matrix size $= 64 \times 64$; field of view (FOV) $= 19.2\,\mathrm{cm}^2$, 3 mm isotropic resolution). The total experiment comprised 968 volumes scanned in 32 min 16 s.

DATA ANALYSIS

Psychoacoustical measurements

Psychoacoustical data were analyzed with repeated-measures analyses of variance (rmANOVAs, IBM SPSS Statistics Version 21.0). In all rmANOVAs, we report the *F*-values, the *p*-values and the partial η^2, a non-additive value representing the "variance accounted-for" measure of the effect size, which can vary from 0 to 1 for the main effects. *Post-hoc* Tukey tests were Bonferroni corrected.

Objective task. For the threshold estimation of a stimulus condition data from two consecutive valid sessions in which thresholds differed by no more than 6.25 ms (i.e., 5% of the SAM tone duration) from each other were combined. A session was accepted as being valid based on two criteria: (1) Subject had a mean hit rate of 80% of the two easiest Go-stimuli (largest time shifts of the B SAM tone) and (2) their false alarm rate (NoGo-stimuli) was below 20%. Based on the rates of hits and misses, a psychometric function was constructed relating d'-values to each of the time shifts. By linearly interpolating between adjacent values of the psychometric function a shift detection threshold was determined as the time shift resulting in a d'-value of 1.8 (Green and Swets, 1966: **Figure 2**). To exclude training effects, the stimulus conditions of each experiment were randomized. Furthermore, after thresholds for all stimulus conditions were obtained subjects had to repeat the threshold measurement for the first condition of the series. If the new shift detection threshold differed by more than 6.25 ms from the shift detection threshold obtained in the first run subjects had to repeat measurements until the new threshold matched the threshold obtained in the first measurement (threshold difference ≤ 6 ms). In these cases the repeated shift detection threshold was taken for further analysis, discarding the previously measured threshold. In the rmANOVA, the shift detection thresholds were analyzed in relation to the stimulus condition (Δf_{mod}) and $f_{\mathrm{mod\,A}}$ (Experiment 1) or f_c (Experiment 2).

Subjective task. For each subject and each condition the mean proportion of a 2-stream percept was calculated from the presentations of 6 (in quiet) or 10 sequences (during fMRI), respectively, per condition and then averaged across subjects. The proportion of a 2-stream percept in relation to the stimulus condition (Δf_{mod}) and $f_{\mathrm{mod\,A}}$ (Experiment 1) or f_c (Experiment 2) was analyzed in a rmANOVA. The effect of the condition

of presentation (in quiet or during fMRI) on the proportion of a 2-stream percept was tested as between-subjects factor.

fMRI measurements

The functional data were analyzed using BrainVoyager™ QX (Brain Innovation, Maastricht, Netherlands). A standard sequence of pre-processing steps, such as 3D-motion correction, linear trend removal, and filtering with a high-pass of three cycles per scan was performed. The functional data sets were projected to the IR-EPI-images, co-registered with the 3D-data sets, and then transformed to Talairach space.

For each experiment separately, a conjunction analysis using a multi-subject random-effects general linear model (RFX-GLM) was performed to identify brain regions which showed positive deflections of the BOLD signal in at least one of the 3 conditions compared to the baseline ($t \geq 4.5$, $p < 0.002$ (uncorrected for multiple comparisons), cluster threshold: 108 mm^3) for each of the two stimulus variants:

Experiment 1: $f_{mod\ A}$ 100 Hz > baseline AND $f_{mod\ A}$ 300 Hz > baseline,
Experiment 2: f_c 1 kHz > baseline AND f_c 4 kHz > baseline.

The analysis included %-transformed functional data of all subjects and used the standard 2-gamma hemodynamic response function implemented in BrainVoyager™ QX. From the resulting clusters volumes-of-interest (VOIs) were defined. The BOLD responses of each VOI were subjected to repeated-measures analyses of variance (rmANOVAs) testing for the within factors condition (Experiment 1 and 2: small, medium and large Δf_{mod}), $f_{mod\ A}$-variant (Experiment 1: 100, 300 Hz) and f_c-variant (Experiment 2: 1, 4 kHz). *Post-hoc* pair wise comparisons were performed using RFX-GLM analyses.

RESULTS

PSYCHOACOUSTICAL MEASUREMENTS IN QUIET AND DURING fMRI

The perceptual segregation of SAM tones was evaluated using either the subjective task (subjects directly reported their perceptual state in quiet or during fMRI) or the objective task that relied on the detection of a forward shifted B SAM tone within the ABA- sequence. In the first experiment the effect of $f_{mod\ A}$ was evaluated, whereas in the second experiment the effect of f_c was evaluated. Both, a variation of $f_{mod\ A}$ as well as a variation of f_c affects the representation of the SAM tones by temporal and/or spectral cues.

EXPERIMENT 1—THE EFFECT OF THE MODULATION FREQUENCY OF THE A SAM TONE ($f_{mod\ A}$)

Subjective task

The proportion of a 2-stream percept depended significantly on the stimulus condition Δf_{mod} [$F_{(2,\ 34)} = 31.755$; $p < 0.001$, $\eta^2 = 0.651$]. The $f_{mod\ A}$ and the condition of presentation (in quiet and during fMRI) did not have a significant effect on the proportion of a 2-stream percept (**Figure 3**). Pair-wise comparisons showed a significant difference in the proportion of a 2-stream percept between all tested Δf_{mod} stimulus conditions (all $p \leq 0.003$). The mean proportion of a 2-stream percept increased significantly with increasing Δf_{mod} condition showing the least mean proportion of a 2-stream percept of 16.2% for ABA- sequences presented with the small Δf_{mod} condition. For ABA- sequences presented with the medium Δf_{mod} condition a mean proportion of a 2-stream percept of 47.7% was observed. The largest mean proportion of a 2-stream percept of 77.7% was observed for the large Δf_{mod} condition. No significant interaction was found.

Objective task in quiet

The shift detection threshold of the B signal was significantly affected by the Δf_{mod} stimulus condition [$F_{(2,\ 10)} = 38.795$; $p < 0.001$, $\eta^2 = 0.886$, **Figure 4**]. No significant main effect of $f_{mod\ A}$

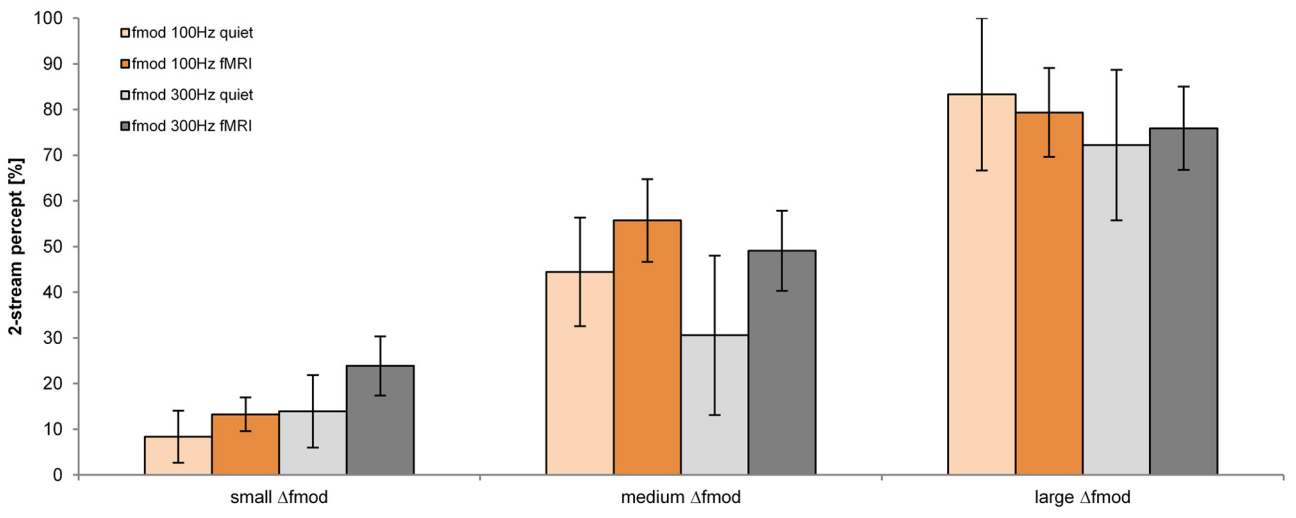


FIGURE 3 | Proportions of a 2-stream percept (mean and s.e.m.) are shown for the $f_{mod\ A}$ of 100 Hz (orange) and 300 Hz (gray) for the measurements in quiet (lighter coloring: $n = 6$) and during fMRI (darker coloring: $n = 13$) for all Δf_{mod} conditions.

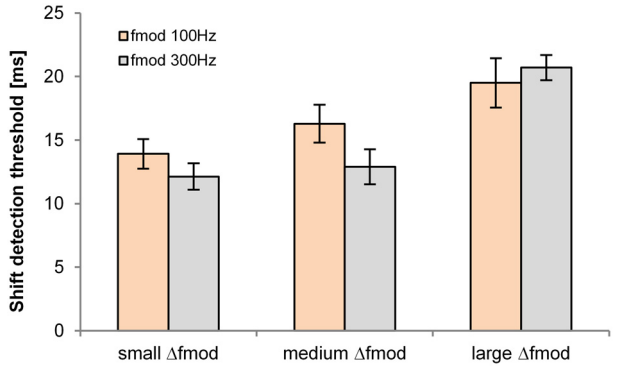


FIGURE 4 | Shift detection thresholds of the B SAM tone ($n = 6$; mean and SEM) are shown for the $\Delta f_{\text{mod A}}$ of 100 Hz (orange) and 300 Hz (gray) for all Δf_{mod} conditions.

on the shift detection threshold was observed. Pair-wise comparisons showed significantly higher shift detection threshold for the large (mean = 20.1 ms) than for the small (mean = 13 ms; $p = 0.001$) and medium Δf_{mod} condition (mean = 14.6 ms; $p = 0.001$). No significant difference between the shift detection threshold of the small and the medium Δf_{mod} condition was observed and no significant interaction was found.

Whether B SAM tones were presented by themselves (control experiment presentation of B-only sequences) or together with A SAM tones (only large Δf_{mod} condition of the main experiment, presentation of ABA- sequences) had a significant effect on the shift detection thresholds [$F_{(1,\ 3)} = 34.272$; $p = 0.01$, $\eta^2 = 0.920$]. Pair-wise comparisons showed significant higher mean shift detection thresholds for the control experiment (48.3 ± 2.4 ms) than observed in the large Δf_{mod} condition of the main experiment (mean $= 20.1 \pm 1.1$ ms).

fMRI MEASUREMENTS

Table 2 lists all brain regions which were commonly activated or deactivated ($t = 4.5$, $p < 0.002$), respectively, by both f_{modA} in at least one of the three conditions (small, medium, and large Δf_{mod}) compared to the baseline condition.

In Experiment 1, the ANOVAs of BOLD responses within the respective VOIs revealed a main effect of Δf_{mod} condition in left Heschl's gyrus (HG) [$F_{(2,\ 24)} = 4.840$, $p = 0.017$] and left posterior cingulated gyrus (PCG) [$F_{(2,\ 24)} = 3.515$, $p = 0.045$]. In the left HG the BOLD response amplitude increased with increasing Δf_{mod} (see **Figure 5**), The *post-hoc* tests showed a significant difference between the small and the large Δf_{mod} condition ($t = 2.892$, $p = 0.013$) and a trend between the medium and the large Δf_{mod} condition ($t = 2.102$, $p = 0.057$). In left PCG, the *post-hoc* tests showed a significantly stronger negative deflection of the BOLD signal of the medium compared to the small Δf_{mod} condition ($t = 3.465$, $p = 0.005$).

In addition, in left and right HG [$F_{(1,\ 12)} = 22.800$, $p < 0.001$; $F_{(1,\ 12)} = 47.735$, $p < 0.001$], left and right insula [$F_{(1,\ 12)} = 6.685$, $p = 0.024$; $F_{(1,\ 12)} = 5.227$, $p = 0.041$], and the left posterior medial frontal cortex (pMFC) [$F_{(1,\ 12)} = 11.655$, $p = 0.005$] a main effect of $f_{\text{mod A}}$ was found with higher BOLD response amplitudes during $f_{\text{mod A}}$ 300 Hz compared to $f_{\text{mod A}}$ 100 Hz stimulation (see **Figure 5**). There was no significant interaction of the factors Δf_{mod} condition and $f_{\text{mod A}}$.

EXPERIMENT 2—THE EFFECT OF THE CARRIER FREQUENCY (f_C)

Subjective task

The proportion of a 2-stream percept depended significantly on the Δf_{mod} condition [$F_{(2,\ 26)} = 51.595$; $p < 0.001$, $\eta^2 = 0.799$], on the f_c of the SAM tones [$F_{(1,\ 13)} = 11.623$; $p = 0.005$, $\eta^2 = 0.472$] and on the condition of presentation [in quiet or during fMRI; $F_{(1,\ 13)} = 8.168$; $p = 0.013$, $\eta = 0.386$; **Figure 6**]. Pair-wise comparisons showed a significant difference in the proportion of a 2-stream percept between all tested Δf_{mod} conditions (all $p \leq 0.009$). The proportion of a 2-stream percept increased significantly with increasing Δf_{mod} (mean percentage of a 2-stream percept for the small $\Delta f_{\text{mod}} = 11.9\%$, medium $\Delta f_{\text{mod}} = 44.0\%$ and large $\Delta f_{\text{mod}} = 86.9\%$). ABA- SAM tone sequences presented with the lower f_c of 1 kHz showed a significantly higher proportion of a 2-stream percept (50.5%) than SAM tones of the higher f_c of 4 kHz (44.7%). The proportion of a 2-stream percept measured in quiet was significantly smaller (mean = 35.2%) than the proportion of a 2-stream percept measured during fMRI (mean = 55.8%). The Two-Way interaction of the factors f_c and condition of presentation was significant ($p < 0.001$), showing a significant higher proportion of a 2-stream percept for the lower f_c of 1 kHz in quiet (mean = 45.4%) than for the higher f_c of 4 kHz in quiet (mean = 25.0%; $p = 0.006$), whereas the proportion of a 2-stream percept during fMRI was not affected by the f_c. No other interaction was significant.

Objective task in quiet

The detection threshold of the time shifted B SAM tone of the ABA- sequence was significantly dependent on the stimulus condition Δf_{mod} [$F_{(2,\ 10)} = 10.018$; $p = 0.004$, $\eta^2 = 0.667$, **Figure 7**]. No significant main effect of f_c on the shift detection threshold was observed. Pair-wise comparisons showed a significantly smaller shift detection threshold for the small (mean = 14.2 ms) than for the large Δf_{mod} stimulus condition (mean = 19.2 ms; $p = 0.01$). No significant difference between the shift detection threshold of the medium (mean = 15.9 ms) and the small and large Δf_{mod} was observed. No significant interaction was found.

fMRI MEASUREMENTS

Table 3 lists all brain regions which were commonly activated ($t = 4.5$, $p < 0.002$) by both f_c in at least one of the three conditions (small, medium, and large Δf_{mod}) compared to the baseline condition.

In Experiment 2, a main effect of condition was found in the left HG, the right superior temporal gyrus (STG), the left MedFG, and the right inferior parietal lobe (IPL) [$F_{(2,\ 18)} = 8.667$, $p = 0.002$; $F_{(2,\ 8)} = 19.634$, $p < 0.001$; $F_{(2,\ 18)} = 3.598$, $p = 0.048$; $F_{(2,\ 18)} = 3.501$, $p = 0.052$]. In left HG and right STG the same gradual increase in BOLD response amplitude with increasing Δf_{mod} was observed as in the left AC in Experiment 1 (see **Figures 5**, **8**). *Post-hoc* tests in left HG and right STG revealed significant differences in BOLD responses between the small

Table 2 | Brain regions (BA-Brodmann area; *x,y,z*-Talairach coordinates) showing positive or negative deflections of the BOLD signal in at least one of the three Δf_{mod} conditions compared to the baseline ($t \geq 4.5$, $p < 0.002$) for each of the two $f_{mod\ A}$ (100, 300 Hz) tested in Experiment 1 and the results of ANOVAs within the resulting VOIs.

Brain region		Talairach coordinates			VOI analysis (ANOVA)		
					Main effect		
	BA	*x*	*y*	*Z*	Δf_{mod} condition	$f_{mod\ A}$	Volume (mm^3)
EXPERIMENT 1							
REGIONS OF ACTIVATION							
Left Hemisphere							
Heschl's gyrus	41	−48	−20	11	$F = 4.840$; $p = 0.017$	$F = 22.800$; $p < 0.001$	1307
Insula	13	−29	23	4	$F = 0.591$; $p = 0.562$	$F = 6.685$; $p = 0.024$	614
Putamen		−21	−1	8	$F = 0.101$; $p = 0.904$	$F = 0.972$; $p = 0.344$	150
Precentral gyrus	6	−38	2	30	$F = 0.091$; $p = 0.913$	$F = 2.591$; $p = 0.133$	207
Inferior parietal lobe	40	−36	−50	41	$F = 0.429$; $p = 0.656$	$F = 2.607$; $p = 0.132$	1047
Posterior medial frontal cortex	6	−5	8	46	$F = 0.048$; $p = 0.953$	$F = 11.655$; $p = 0.005$	389
Right Hemisphere							
Heschl's gyrus	41	53	−17	9	$F = 2.173$; $p = 0.136$	$F = 47.735$; $p < 0.001$	345
Superior temporal gyrus	22	62	−13	7	$F = 2.048$; $p = 0.151$	$F = 2.422$; $p = 0.146$	2357
Insula	13	32	21	2	$F = 1.804$; $p = 0.186$	$F = 5.227$; $p = 0.041$	390
Inferior parietal lobe	40	36	−48	40	$F = 0.375$; $p = 0.6$	$F = 1.017$; $p = 0.333$	1205
Middle frontal gyrus	9	47	15	26	$F = 0.215$; $p = 0.808$	$F = 0.863$; $p = 0.371$	738
Posterior medial frontal cortex	6	3	17	44	$F = 0.313$; $p = 0.734$	$F = 2.026$; $p = 0.180$	309
REGIONS OF DEACTIVATION							
Left Hemisphere							
Posterior cingulated gyrus	31	−48	−20	11	$F = 3.515$; $p = 0.045$	n.s.	138

Significant effects are highlighted in gray.

and the large Δf_{mod} condition ($t = 3.710$, $p = 0.005$; $t = 5.318$, $p < 0.001$) and between the small and the medium Δf_{mod} condition ($t = 5.727$, $p < 0.001$; $t = 5.929$, $p < 0.001$). In right STG, the large Δf_{mod} condition also resulted in a significantly stronger BOLD response than the medium Δf_{mod} condition ($t = 2.698$, $p = 0.024$). In left pMFC no gradual increase in BOLD response amplitude with increasing Δf_{mod} was observed. In contrast, the BOLD response of the medium Δf_{mod} condition was stronger than those of the small and the large Δf_{mod} condition (see **Figure 8**) with a significant difference between the medium and the small Δf_{mod} condition ($t = 2.258$, $p = 0.050$). The BOLD responses of the small and the large Δf_{mod} condition were very similar ($t = 0.643$, $p = 0.536$). *Post-hoc* testing in right IPL did not reach significance. No significant main effect of the f_c and no significant interaction of the factors condition and f_c were found.

CORRELATION BETWEEN TASKS AND MEASURES OF STREAM SEGREGATION

Correlation between tasks

For all subjects and the two main experiments the mean proportion of a 2-stream percept (subjective task in quiet) and the mean shift detection threshold (objective tasks in quiet) for the tested stimulus conditions were significantly correlated (Spearman's $\rho = 0.683$, $p = 0.042$, **Figure 9A**). The Spearman's non–parametric correlation coefficients for the single subject analyses were rather large ($\rho = 0.527$) for all but one subjects. Only for one subject the correlation reached a significant value ($p = 0.001$).

Correlation between measures

For the purpose of comparison and as an example, the proportion of a 2-stream percept for all tested Δf_{mod} stimulus conditions obtained during fMRI was related to the strength of BOLD responses (beta weights) in left Heschl's gyrus. Spearman's correlation of averaged group data did not reach significance ($\rho = 0.450$, $p = 0.224$, **Figure 9B**).

DISCUSSION

PSYCHOACOUSTICAL EVALUATION OF STREAM SEGREGATION BY SAM

The psychoacoustical results of both experiments and both the subjective and objective task show that an increasing Δf_{mod} between A and B SAM tones promotes stream segregation, being in agreement with the results of other psychoacoustical studies that evaluated stream segregation by either different SAM tones (e.g., Dolležal et al., 2012a; Szalárdy et al., 2013) or SAM noise bursts (Grimault et al., 2002).

In the present study in Experiment 1 the subjective perception of stream segregation is not affected by the $f_{mod\ A}$ (100 and 300 Hz, respectively) of the SAM tones. Dolležal et al. (2012a), who presented sequences of SAM tones differing in multiple parameters in addition to $f_{mod\ A}$ and f_c [e.g., tone pattern (combinations of TRT and tone duration), modulation depth and presentation time] and used more steps of Δf_{mod}, however, observed an increasing proportion of a 2-stream percept for increasing $f_{mod\ A}$ (30, 100, and 300 Hz). This difference between the two

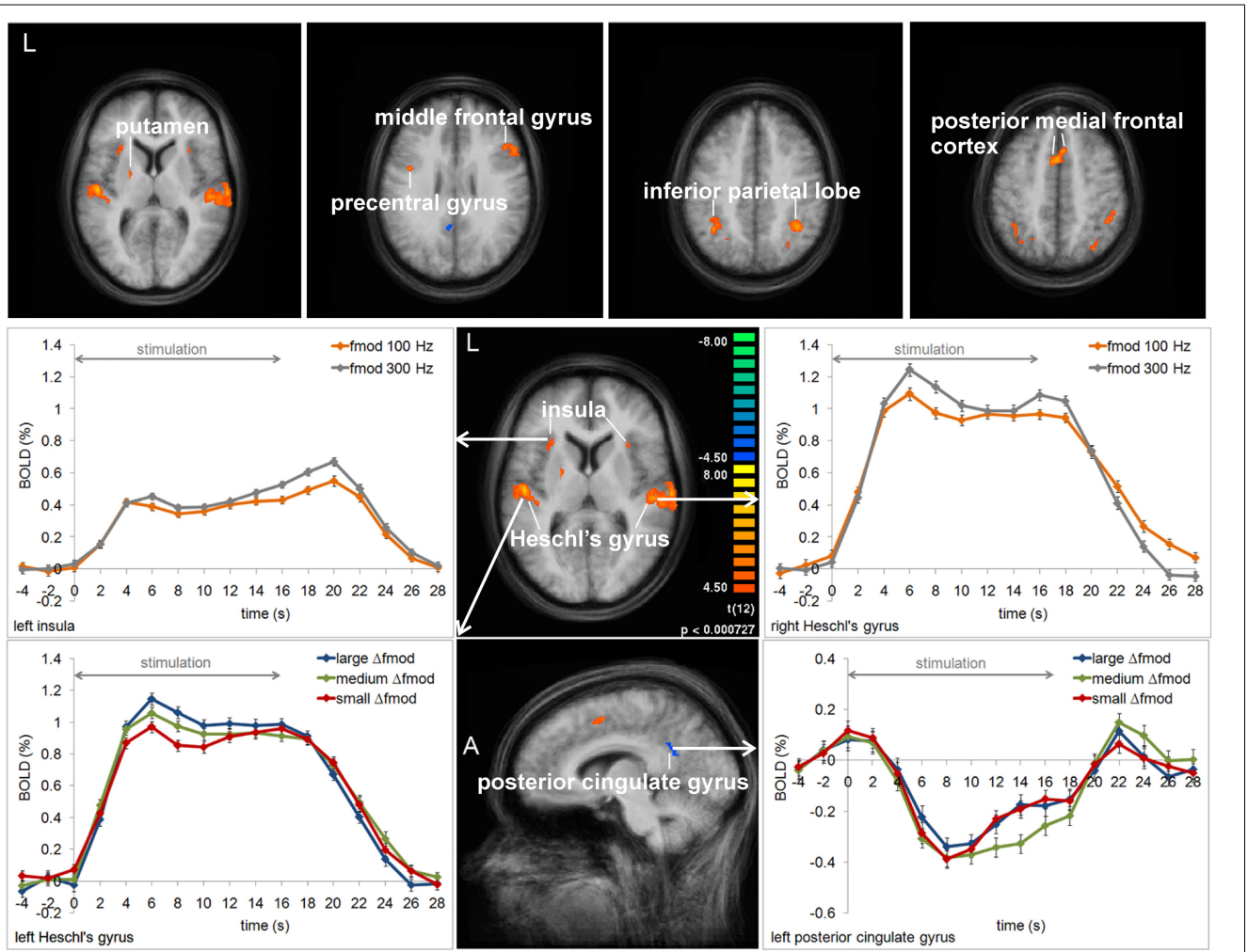


FIGURE 5 | Group average activation maps (13 subjects) and BOLD signal time courses within regions of interest in Experiment 1. The maps depict all brain regions showing positive or negative deflections of the BOLD signal in at least one of the three Δf_{mod} conditions compared to the baseline ($t = 4.5$, $p < 0.002$) for each of the two $f_{mod\ A}$ (100, 300 Hz). Several regions that showed significant differences between conditions and $f_{mod\ A}$ variants are labeled and the respective averaged BOLD signal time courses are assigned. Error bars represent SEM.

studies could be attributed to the differences in the range of $f_{mod\ A}$ and Δf_{mod} that was larger in the previous study by Dolležal et al. (2012a). When comparing the proportion of a 2-stream percept of the medium Δf_{mod} condition, that was explicitly chosen to compare stream segregation of temporal ($f_{mod\ A} = 100$ Hz) vs. spectral ($f_{mod\ A} = 300$ Hz) cues (**Table 1**), spectral cues appear not to further stream segregation more than temporal cues. Next to the evaluation of the subjective streaming percept the present study also applied an objective task of stream segregation to the same stimulus conditions to be able to directly compare the streaming percept across both psychoacoustical tasks. The shift detection thresholds obtained with the objective task increased with increasing Δf_{mod} between A and B SAM tones. Such an increase in the shift detection threshold with increasing feature differences between A and B signals has been observed in other studies that also presented time shifted signals in an objective task using a range of different features (frequency differences: Van Noorden, 1975; Neff et al., 1982; Cusack and Roberts, 2000; Micheyl and Oxenham, 2010; Thompson et al., 2011; differences in the starting phases of frequency components: Roberts et al., 2002; differences in fundamental frequencies: Vliegen et al., 1999). Furthermore, Divenyi and Danner (1977) also observed a sizable deterioration of the discrimination performance if the signals were made very dissimilar from each other (e.g., in frequency or intensity) even though they did not employ a paradigm that led to a streaming percept.

We also applied the shift detection task in an ABA- sequence with omitted A signals (-B---B---B--...) to determine the shift detection threshold in a condition providing no temporal reference to A signals. We observed higher shift detection thresholds in the B-only condition than for the large Δf_{mod} condition of ABA- sequences with A and B signals. That difference in threshold may indicate that even in sequences with well segregated A and B signals the A signal can provide support to the detection of the time shift of the B signal. If subjects would have solely relied on the B SAM tones for their performance in both the B only

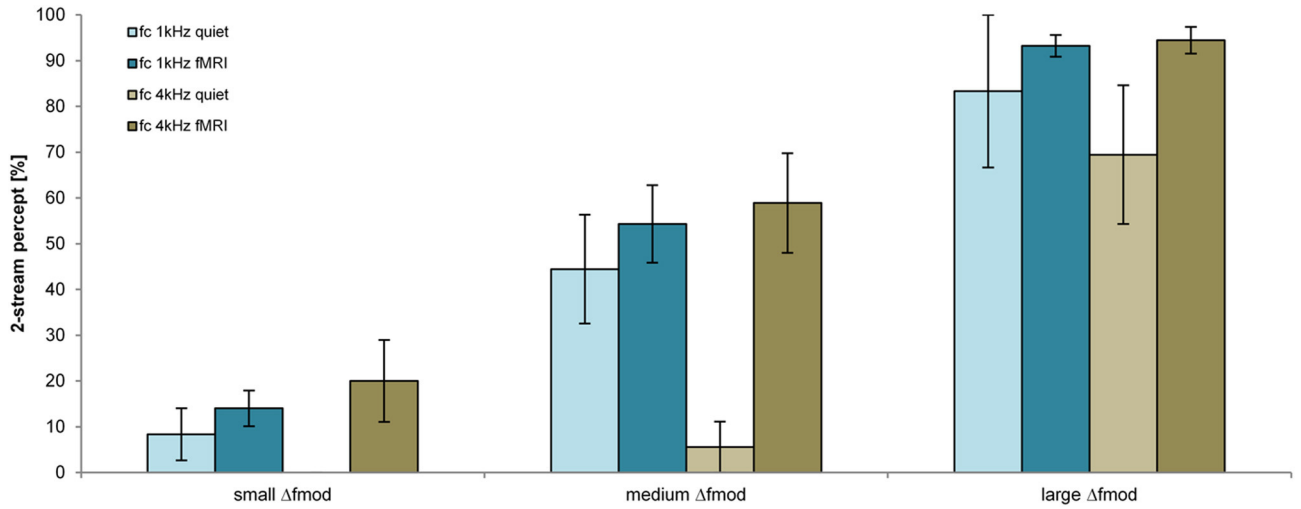


FIGURE 6 | Proportions of a 2-stream percept (mean and s.e.m.) are shown for the f_c of 1 kHz (blue) and 4 kHz (brown) for the measurements in quiet (lighter coloring: $n = 6$) and during fMRI (darker coloring: $n = 9$) for all Δf_{mod} conditions.

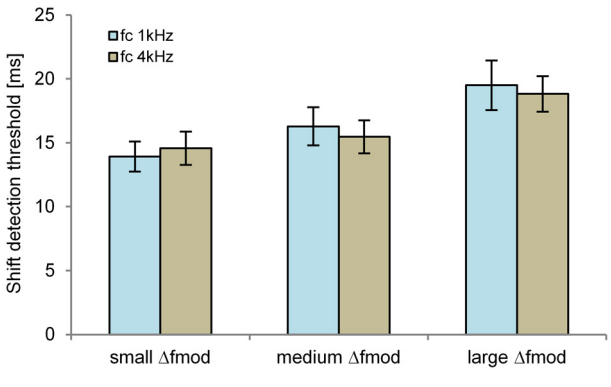


FIGURE 7 | Shift detection thresholds of the B SAM tone ($n = 6$; mean and SEM) are shown for the f_c of 1 kHz (blue) and 4 kHz (brown) for all Δf_{mod} conditions.

condition and in the ABA- condition the thresholds should be alike.

In Experiment 2 the subjective perception of stream segregation was affected by the Δf_{mod} condition and by f_c (1 and 4 kHz, respectively) of the SAM tones when analyzing the subjective data from fMRI and those obtained in quiet together. The effects of the f_c and Δf_{mod} and their interaction was also observed by Dolležal et al. (2012a) who reasoned that the difference in the proportion of a 2-stream percept may be due to the excitation pattern differences between A and B signals being assessed by the auditory system. In the present analysis, a higher proportion of a 2-stream percept was observed for the lower f_c of 1 kHz than for the higher f_c of 4 kHz. At a f_c of 1 kHz at least in the large Δf_{mod} condition spectral cues provided for stream segregation in addition to the temporal cues that were also the prominent cue for ABA- SAM tones presented at a f_c of 4 kHz. Thus, the spectral excitation pattern difference available for the lower f_c of 1 kHz providing additional cues to stream segregation may be the cause for the higher amount of a 2-stream percept in that condition. The significant interaction between the condition of presentation (quiet, during fMRI) and f_c, however, indicates that responses differed between both presentation conditions. The effect of f_c was only prominent in quiet conditions and not in the noisy fMRI condition that may have precluded the use of excitation pattern differences. The subjective segregation percept in the noisy fMRI condition match the pattern of BOLD responses (see below). If we focus on the large Δf_{mod} condition that allows comparing the amount of stream segregation elicited by spectral vs. temporal cues, we find no significant difference indicating that both type of cues have the potential to elicit the percept of well segregated streams.

In general, the proportion of a 2-stream percept was smaller for subjects that have been tested in quiet, than for subjects that have been tested in scanner noise during fMRI measurements. Especially in the medium Δf_{mod} condition we observed a small amount of stream segregation that was less than expected on the basis of the previous measurements (Dolležal et al., 2012a). A similar difference in the proportion of a 2-stream percept has been observed in Experiment 1, but it did not reach significance. Wilson et al. (2007) also compared the streaming perception of subjects in quiet and during fMRI. Their results show a non-significant but higher proportion of a 2-stream percept for subjects tested during fMRI than in the quiet booth revealing a tendency that is comparable to the results of the present study. Dolležal et al. (2012a) also observed a higher proportion of a 2-stream percept in pink noise than in quiet. A general explanation for the observed effect, however, cannot be provided.

When presenting the stimulus conditions of Experiment 2 in the objective task no effect of f_c on the shift detection threshold can be observed whereas an effect of Δf_{mod} remained. In the subjective task the effect size of Δf_{mod} was considerably larger than the effect size for f_c. Since the objective task will lead to better

Table 3 | Brain regions (BA-Brodmann area; *x,y,z*-Talairach coordinates) showing positive deflections of the BOLD signal in at least one of the three Δf_{mod} conditions compared to the baseline ($t = 4.5$, $p < 0.002$) for each of the two f_c (1, 4 kHz) tested in Experiment 2 and the results of ANOVAs within the resulting VOIs.

Brain region					VOI analysis (ANOVA)		
		Talairach coordinates			Main effect		
	BA	*x*	*y*	*Z*	Δf_{mod} condition	$f_{mod\ A}$	Volume (mm^3)
EXPERIMENT 2							
REGIONS OF ACTIVATION							
Left Hemisphere							
Heschl's gyrus	41	−44	−24	12	$F = 8.667$; $p = 0.002$	$F = 0.016$; $p = 0.900$	540
Inferior parietal lobe	40	−33	−52	40	$F = 1.014$; $p = 0.383$	$F = 0.016$; $p = 0.902$	126
Posterior medial frontal cortex	6	−3	14	46	$F = 3.598$; $p = 0.048$	$F = 1.000$; $p = 0.343$	257
Right Hemisphere							
Superior temporal gyrus	22	51	−13	10	$F = 19.634$; $p < 0.001$	$F = 3.177$; $p = 0.108$	592
Insula	13	36	19	4	$F = 0.859$; $p = 0.440$	$F = 0.483$; $p = 0.504$	113
Inferior parietal lobe	40	38	−45	46	$F = 3.501$; $p = 0.052$	$F = 1.650$; $p = 0.231$	541

Significant effects are highlighted in gray.

thresholds if the subjects integrate A and B signals into a single stream (e.g., Van Noorden, 1975; Neff et al., 1982; Vliegen et al., 1999; Cusack and Roberts, 2000; Roberts et al., 2002; Micheyl and Oxenham, 2010; Thompson et al., 2011), they may be inclined to integrate more than in a subjective evaluation of the stimuli. This may reduce smaller effects of the subjective task to non-significance in the objective task.

BOLD ACTIVITY DURING STREAM SEGREGATION BY SAM TONES

Corresponding to the psychoacoustical results, BOLD activity in auditory cortex regions depended on the Δf_{mod} between A and B SAM tones. With increasing Δf_{mod} the dominant percept changed from a 1-stream to a 2-stream and the BOLD response amplitudes gradually increased. The results of Experiment 1 and 2 differ, however, in that the Δf dependent effect was observed only in left auditory cortex in Experiment 1 and in both auditory cortices in Experiment 2. Previous human imaging studies on stream segregation found either an involvement of both auditory cortices (e.g., Gutschalk et al., 2007; Wilson et al., 2007) or a specific involvement of the left auditory cortex (Deike et al., 2004, 2010). Deike et al. suggested that the involvement of the left hemisphere was caused by the specific demands on sequential analysis in the active stream segregation task. Even though the present experiments require the sequential analysis of the sound sequences, the subjects were not forced to actively group the sounds into one or the other perceptual organization but had to monitor their spontaneous perception. Therefore, one might rather suggest a stimulus driven representation of Δf_{mod} in both auditory cortices and the failure to observe this in right auditory cortex in Experiment 1 might simply be explained by statistical thresholding.

The Δf_{mod} dependent effect in auditory cortex regions was observed for all stimulus parameters and thus, irrespective as to whether SAM tones provide spectral, temporal or both types of cues. Several human imaging studies have described increasing neural activity throughout the auditory cortex for both differences in spectral (Deike et al., 2004, 2010; Gutschalk et al., 2005; Snyder et al., 2006; Wilson et al., 2007) and in temporal (Gutschalk et al., 2007) properties between A and B signals in streaming sequences. Hence, our finding of increasing BOLD response amplitudes in auditory cortex regions with increasing Δf_{mod} between SAM tones is consistent with previous studies. Electrophysiological recording studies in animals using pure-tone paradigms suggested that frequency selectivity of tonotopically organized neurons in primary auditory cortical fields in combination with forward suppression leads to separate representations of A and B tones that contribute to the percept of two separate streams (Fishman et al., 2001, 2004; Kanwal et al., 2003; Bee and Klump, 2004, 2005; Micheyl et al., 2005). With increasing frequency separation between tones the populations of active neurons become more disjoined, leading to decreasing suppression between successive tones. It was supposed that this decrease in suppression causes the larger summed activity in auditory cortex measured using fMRI, EEG, or MEG (Gutschalk et al., 2005; Snyder et al., 2006; Wilson et al., 2007). Using harmonic tone complexes with only unresolved harmonics Gutschalk et al. (2007) suggested that suppression also accounts for the interaction of sounds with differences in temporal properties. For SAM tones Bartlett and Wang (2005) found that neurons in marmoset monkey auditory cortex show significant forward suppression of the preceding to the following SAM tone. Similarly, Itatani and Klump (2009), who used the same ABA- paradigm as in the present study and tested a large parameter space of SAM tones, observed forward suppression in multiunit responses of the auditory forebrain of awake European starlings. Related to this potential common cortical mechanism underlying stream segregation on temporal and spectral properties of sounds one may further ask the question of pitch representation at the cortex. In the present study, two stimulus conditions (Experiment 1: medium Δf_{mod}, Experiment 2: large Δf_{mod}) provided a direct comparison between spectral and temporal pitch cues on which stream segregation was based and we did not find any cortical region which showed a significant difference in BOLD responses

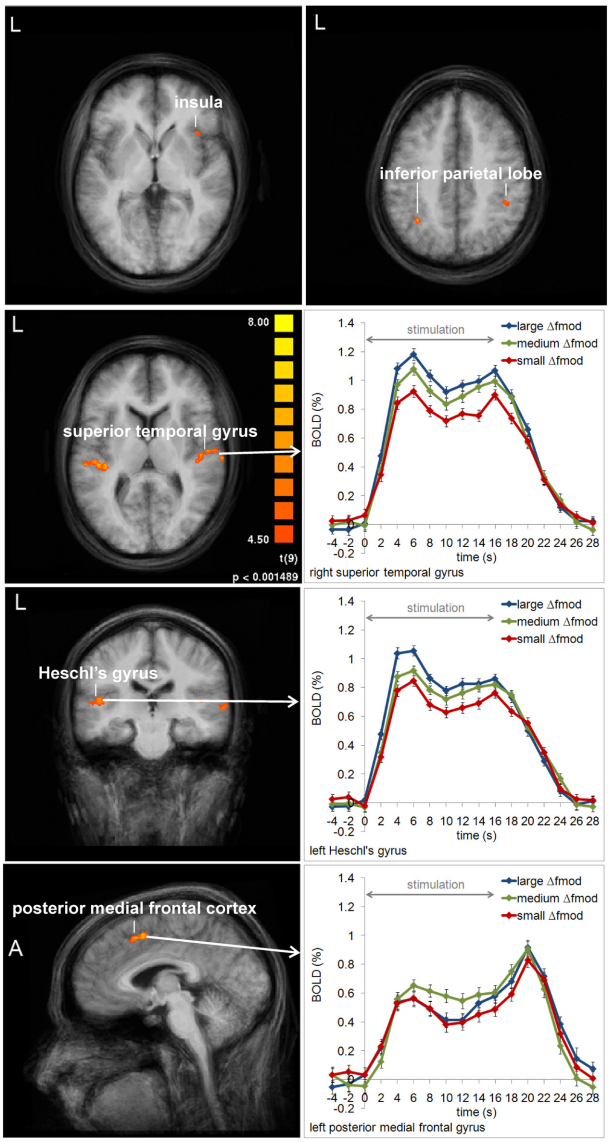


FIGURE 8 | Group average activation maps (10 subjects) and BOLD signal time courses within regions of interest in Experiment 2. The maps depict all brain regions with positive deflections of the BOLD response in at least one of the three Δf_{mod} conditions compared to the baseline ($t = 4.5$, $p < 0.002$) for each f_c (1, 4 kHz). Several regions that showed significant differences between conditions are labeled and the respective averaged BOLD signal time courses are assigned. Error bars represent SEM.

between both cues in this comparison. This finding is consistent with the results by Hall and Plack (2009) who tested a range of pitch-evoking stimuli with different spectral, temporal, and binaural characteristics and did not find any differentiated activation within auditory cortex regions. Although differing in anatomical location, there is supporting evidence for a cue independent common pitch region in auditory cortex coming from the neurophysiological study by Bendor and Wang (2005) who found pitch-selective neurons near the anterolateral low-frequency border of the primary auditory cortex field A1 in marmoset monkeys. At the same time, the underlying mechanism for pitch coding in this region was found to depend both on the temporal and spectral characteristics of the sounds (Bendor et al., 2012).

In Experiment 1, BOLD responses in left and right Heschl's gyrus depended on the $f_{mod\ A}$ of SAM tones with higher BOLD response amplitudes for the higher $f_{mod\ A}$ of 300 Hz compared to the smaller one of 100 Hz. As SAM tones are characterized by three spectral peaks, i.e., the central peak representing the f_c and the two sidebands (upper: $f_c + f_{mod\ A}$, lower: $f_c - f_{mod\ A}$), the stronger responses for the higher $f_{mod\ A}$ of 300 Hz might be explained by broader spectral excitation. In addition, higher BOLD response amplitudes for the higher $f_{mod\ A}$ were also observed in left and right insula and in the left medial part of Brodmann area 6 comprising the supplementary motor area (SMA). The involvement of these areas might be thought in the context of specific task demands other than motor processing in which both have a primary function. Specifically, the insula cortex has a role in different auditory processes, such as allocating auditory attention, temporal processing, phonological processing, and visual-auditory integration (for review, see Bamiou et al., 2003). The SMA is described as a part of the larger functional unit of posterior medial frontal cortex (pMFC) which has a function in cognitive control and particularly in performance monitoring including monitoring of response conflicts and decision uncertainty (for review, see Ridderinkhof et al., 2004). As the subjects' task was to assign their perception to one of the two perceptual alternatives, the stronger activity for the $f_{mod\ A}$ of 300 Hz in the pMFC might reflect the monitoring of a response conflict or uncertainty in perceptual decision. In the same way, Tregellas et al. (2006) observed in the pMFC and the insular/opercular cortex an increase in BOLD activity in a "difficult" compared to an "easy" auditory temporal processing task. In their study, the subjects had to discriminate the duration of the second tone within pairs of tones and the task difficulty was adapted by varying duration differences between tones. Increasing BOLD activity in the anterior insula bilaterally with increasing task demands were also observed in a pitch discrimination and a n-back pitch memory task (Rinne et al., 2009). In the present study the task required sequence processing and rhythmic pattern perception, namely comparing the galloping ABA- rhythm (1-stream percept) to the two different isochronous rhythms (A-A-A-…and -B---B---…; 2-stream percept). Although the proportion of 2-stream perception is very similar across conditions between both $f_{mod\ A}$ variants one might suppose that the perceptual decision might be more difficult for the higher $f_{mod\ A}$ of 300 Hz because of specific sound qualities (e.g., timbre) other than pitch. Corresponding to this, one might suggest that the stronger BOLD response in auditory cortex for the higher $f_{mod\ A}$ of 300 Hz also reflects the task difficulty. This notion finds support in human imaging studies providing evidences that even in sensory areas the activation can be modulated by task difficulty (Gerlach et al., 1999; Brechmann and Scheich, 2005; Reiterer et al., 2005; Harinen and Rinne, 2013).

In both experiments, cortical regions outside the auditory cortex were found which showed specific activity for medium Δf_{mod} between SAM tones compared to the small and the large Δf_{mod}'s. In particular, the left pMFC (Experiment 2) and the left posterior

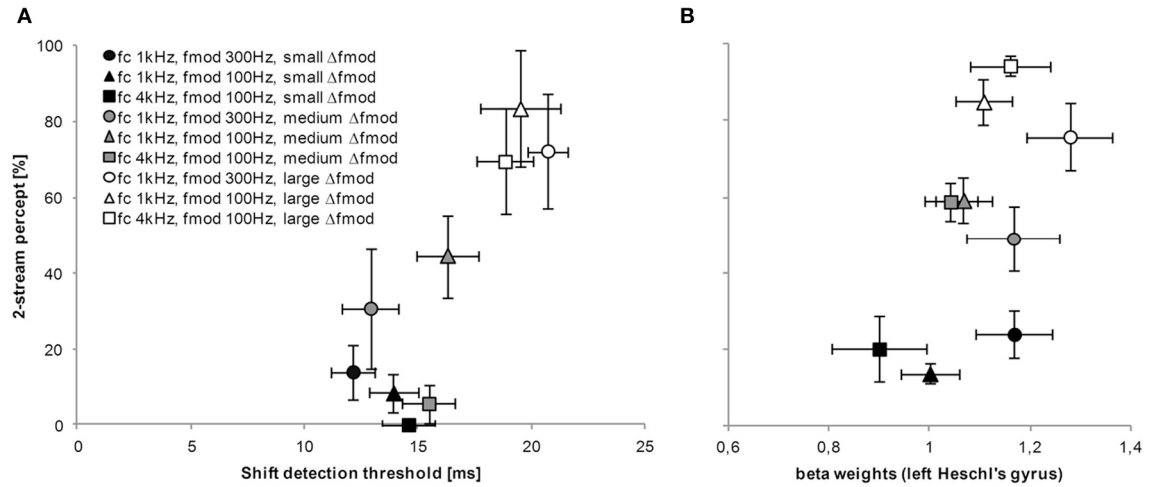


FIGURE 9 | Relationship between tasks and measures of stream segregation. The graph in **(A)** represents the proportion of a 2-stream percept for all tested Δf_{mod} stimulus conditions obtained in the subjective psychophysical task (y-axis) and the matching shift detection threshold obtained in the objective psychophysical task (x-axis). **(B)** For the purpose of comparison and as an example, the proportion of a 2-stream percept for all tested Δf_{mod} stimulus conditions obtained during fMRI (y-axis) is related to the strength of BOLD responses (beta weights) in left Heschl's gyrus (x-axis). The symbols represent the three possible combinations of f_c, $f_{mod\ A}$, and Δf_{mod} (see legend for values). Generally, the shading of the symbol represent the Δf_{mod} stimulus condition; The darkest shading represent values for the small Δf_{mod} stimulus condition, whereas the lightest shading represent values for the large Δf_{mod} stimulus condition. Mean and error bars (SEM) are presented.

cingulate gyrus (PCG) (Experiment 1) showed stronger positive and negative deflections of the BOLD signal, respectively, for the medium Δf_{mod} and very similar smaller BOLD responses for the two other conditions. This activation pattern is very different from the gradual increase in activation with increasing Δf_{mod} that was observed in auditory cortex regions. Whereas the activation gradient in auditory cortex rather reflects the physical differences between conditions, the BOLD responses in the pMFC and the PCG might rather be related to perceptional decision. As already mentioned above, the pMFC has a cognitive function in response conflicts and decision uncertainty. This is particularly the case in the ambiguous perceptual region where both perceptual alternatives are possible and compete with each other. Thus, the stronger BOLD response for ambiguous sequences in the pMFC might be explained by response conflicts and/or decision uncertainty. Similarly, response conflicts or decision uncertainty are equivalent to imposing higher task demands that might explain the stronger deactivation for ambiguous sequences in the PCG which is a part of the "task negative" default mode network showing decreasing activity with increasing task demands (Raichle et al., 2001; Corbetta and Shulman, 2002; Fox et al., 2005; Dosenbach et al., 2007).

Our fMRI results can be summarized as follows. In auditory cortex stream segregation on SAM tones showed the same Δf dependent BOLD responses as other streaming stimuli. In contrast, BOLD activity in regions outside the auditory cortex rather appear to reflect the perceptual decision and specifically the higher task demands caused by specific stimulus characteristics or by perceptual ambiguity leading to response conflicts and decision uncertainty, respectively. The involved regions differ from those observed in other studies and we did not find significant activation in any of the regions reported in Cusack (2005) (intraparietal sulcus), Kondo and Kashino (2009) (Thalamus), and Dykstra et al. (2011) (e.g., middle temporal and frontal gyri). This might be explained by general differences in the approaches: Cusack (2005) and Kondo and Kashino (2009) examined ambiguous streaming sequences to find correlates of different perceptual organizations and perceptual switches, respectively, whereas the present study examined stream segregation across the domains of perceptual dominance and ambiguity by varying the stimulus parameters. The study by Dykstra et al. (2011) also compared different Δf conditions and found that the middle temporal and frontal gyri showed the same increase in neural activity with increasing Δf as the auditory cortex. They, however, did not observe a specific response for ambiguous stimuli. This discrepancy must be resolved in future studies.

COMPARISON ACROSS TASKS AND MEASURES

A direct comparison across psychoacoustical tasks of stream segregation showed a correlation across all subjects and experiments (**Figure 9A**). The results of the objective task mirror the results obtained by the subjective task, thus the shift detection threshold as well as the proportion of a 2-stream percept increased with increasing Δf_{mod} between A and B SAM tones. Such a correlation reveals that both the proportion of a 2-stream percept (subjective task) as well as the shift detection threshold (objective task) can represent the amount of stream segregation. These results are in agreement with a study by Micheyl and Oxenham (2010) who presented pure tones in ABA- sequences with frequency differences between A and B tones and also correlated the proportion of a 2-stream percept with the shift detection experiment. A comparison across measures (subjective streaming percept and BOLD responses (beta weights) in left Heschl's gyrus) did not show a significant correlation even though a

relatively high Spearman's rho was observed for the mean values of both measures (**Figure 9B**). In the exemplary figure of the correlation of humans perception and BOLD responses in left Heschl's gyrus a trend similar to the one observed in the figure of the correlation across psychoacoustical tasks (see **Figure 9A**) can be observed, showing an increasing proportion of a 2-stream percept with increasing beta weights measured in BOLD responses.

ACKNOWLEDGMENTS

This study was supported by the DFG (SFB TRR 31, GRK 591). We thank Rainer Beutelmann, Holger Dierker, Monika Dobrowolny, and Antje Schasse for assistance.

REFERENCES

Bamiou, D.-E., Musiek, F. E., and Luxon, L. M. (2003). The insula (Island of Reil) and its role in auditory processing: Literature review. *Brain Res. Rev.* 42, 143–154. doi: 10.1016/S0165-0173(03)00172-3

Bartlett, E. L., and Wang, X. (2005). Long-lasting modulation by stimulus context in primate auditory cortex. *J. Neurophysiol.* 94, 83–104. doi: 10.1152/jn.01124.2004

Baumgart, F., Kaulisch, T., Tempelmann, C., Gaschler-Markefski, B., Tegeler, C., Schindler, F., et al. (1998). Electrodynamic headphones and woofers for application in magnetic resonance imaging scanners. *Med. Phys.* 25, 2068–2070. doi: 10.1118/1.598368

Bee, M. A., and Klump, G. M. (2004). Primitive auditory stream segregation: a neurophysiological study in the songbird forebrain. *J. Neurophysiol.* 92, 1088–1104. doi: 10.1152/jn.00884.2003

Bee, M. A., and Klump, G. M. (2005). Auditory stream segregation in the songbird forebrain: effects of time intervals on responses to interleaved tone sequences. *Brain Behav. Evol.* 66, 197–214. doi: 10.1159/000087854

Bendor, D., Osmanski, M. S., and Wang, X. (2012). Dual-pitch processing mechanisms in primate auditory cortex. *J. Neurosci.* 32, 16149–16161. doi: 10.1523/JNEUROSCI.2563-12.2012

Bendor, D., and Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex. *Nature* 436, 1161–1165. doi: 10.1038/nature03867

Bethmann, A., Tempelmann, C., De Bleser, R., Scheich, H., and Brechmann, A. (2007). Determining language laterality by fMRI and dichotic listening. *Brain Res.* 1133, 145–157. doi: 10.1016/j.brainres.2006.11.057

Brechmann, A., and Scheich, H. (2005). Hemispheric shifts of sound representation in auditory cortex with conceptual listening. *Cereb. Cortex* 15, 578–587. doi: 10.1093/cercor/bhh159

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Corbetta, M., and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3, 201–215. doi: 10.1038/nrn755

Cusack, R. (2005). The intraparietal sulcus and perceptual organization. *J. Cogn. Neurosci.* 17, 641–651. doi: 10.1162/0898929053467541

Cusack, R., and Roberts, B. (2000). Effects of differences in timbre on sequential grouping. *Percept. Psychophys.* 62, 1112–1120. doi: 10.3758/BF03212092

Deike, S., Gaschler-Markefski, B., Brechmann, A., and Scheich, H. (2004). Auditory stream segregation relying on timbre involves left auditory cortex. *Neuroreport* 15, 1511–1514. doi: 10.1097/01.wnr.0000132919.12990.34

Deike, S., Scheich, H., and Brechmann, A. (2010). Active stream segregation specifically involves the left human auditory cortex. *Hear. Res.* 265, 30–37. doi: 10.1016/j.heares.2010.03.005

Divenyi, P. L., and Danner, W. F. (1977). Discrimination of time intervals marked by brief acoustic pulses of various intensities and spectra. *Percept. Psychophys.* 21, 125–142. doi: 10.3758/BF03198716

Dolležal, L.-V., Beutelmann, R., and Klump, G. M. (2012a). Stream segregation in the perception of sinusoidally amplitude-modulated tones. *PLoS ONE* 7:e43615. doi: 10.1371/journal.pone.0043615

Dolležal, L.-V., Itatani, N., Günther, S., and Klump, G. M. (2012b). Auditory streaming by phase relations between components of harmonic complexes: a comparative study of human subjects and bird forebrain neurons. *Behav. Neurosci.* 126, 797–808. doi: 10.1037/a0030249

Dosenbach, N. U. F., Fair, D. A., Miezin, F. M., Cohen, A. L., Wenger, K. K., Dosenbach, R. A. T., et al. (2007). Distinct brain networks for adaptive and stable task control in humans. *Proc. Natl. Acad. Sci. U.S.A.* 104, 11073–11078. doi: 10.1073/pnas.0704320104

Dykstra, A. R., Halgren, E., Thesen, T., Carlson, C. E., Doyle, W., Madsen, J. R., et al. (2011). Widespread brain areas engaged during a classical auditory streaming task revealed by intracranial EEG. *Front. Hum. Neurosci.* 5:74. doi: 10.3389/fnhum.2011.00074

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., and Shamma, S. A. (2009). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron* 61, 317–329. doi: 10.1016/j.neuron.2008.12.005

Fishman, Y. I., Arezzo, J. C., and Steinschneider, M. (2004). Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration. *J. Acoust. Soc. Am.* 116, 1656–1670. doi: 10.1121/1.1778903

Fishman, Y. I., Reser, D. H., Arezzo, J. C., and Steinschneider, M. (2001). Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. *Hear. Res.* 151, 167–187. doi: 10.1016/S0378-5955(00)00224-0

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Essen, D. C. V., and Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc. Natl. Acad. Sci. U.S.A.* 102, 9673–9678. doi: 10.1073/pnas.0504136102

Gerlach, C., Law, I., Gade, A., and Paulson, O. B. (1999). Perceptual differentiation and category effects in normal object recognition A PET study. *Brain* 122, 2159–2170. doi: 10.1093/brain/122.11.2159

Green, D. M., and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics. 1966th Edn.* New York: Wiley.

Grimault, N., Bacon, S. P., and Micheyl, C. (2002). Auditory stream segregation on the basis of amplitude-modulation rate. *J. Acoust. Soc. Am.* 111, 1340–1348. doi: 10.1121/1.1452740

Gutschalk, A., Micheyl, C., Melcher, J. R., Rupp, A., Scherg, M., and Oxenham, A. J. (2005). Neuromagnetic correlates of streaming in human auditory cortex. *J. Neurosci.* 25, 5382–5388. doi: 10.1523/JNEUROSCI.0347-05.2005

Gutschalk, A., Oxenham, A. J., Micheyl, C., Wilson, E. C., and Melcher, J. R. (2007). Human cortical activity during streaming without spectral cues suggests a general neural substrate for auditory stream segregation. *J. Neurosci.* 27, 13074–13081. doi: 10.1523/JNEUROSCI.2299-07.2007

Hall, D. A., and Plack, C. J. (2009). Pitch processing sites in the human auditory brain. *Cereb. Cortex* 19, 576–585. doi: 10.1093/cercor/bhn108

Harinen, K., and Rinne, T. (2013). Activations of human auditory cortex to phonemic and nonphonemic vowels during discrimination and memory tasks. *Neuroimage* 77, 279–287. doi: 10.1016/j.neuroimage.2013.03.064

Itatani, N., and Klump, G. M. (2009). Auditory streaming of amplitude-modulated sounds in the songbird forebrain. *J. Neurophysiol.* 101, 3212–3225. doi: 10.1152/jn.91333.2008

Itatani, N., and Klump, G. M. (2011). Neural correlates of auditory streaming of harmonic complex sounds with different phase relations in the songbird forebrain. *J. Neurophysiol.* 105, 188–199. doi: 10.1152/jn.00496.2010

Kanwal, J. S., Medvedev, A. V., and Micheyl, C. (2003). Neurodynamics for auditory stream segregation: tracking sounds in the mustached bat's natural environment. *Network* 14, 413–435. doi: 10.1088/0954-898X/14/3/303

Kohlrausch, A., Fassel, R., and Dau, T. (2000). The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers. *J. Acoust. Soc. Am.* 08, 723–734. doi: 10.1121/1.429605

Kondo, H. M., and Kashino, M. (2009). Involvement of the thalamocortical loop in the spontaneous switching of percepts in auditory streaming. *J. Neurosci.* 29, 12695–12701. doi: 10.1523/JNEUROSCI.1549-09.2009

Micheyl, C., and Oxenham, A. J. (2010). Objective and subjective psychophysical measures of auditory stream integration and segregation. *J. Assoc. Res. Otolaryngol.* 11, 709–724. doi: 10.1007/s10162-010-0227-2

Micheyl, C., Tian, B., Carlyon, R. P., and Rauschecker, J. P. (2005). Perceptual organization of tone sequences in the auditory cortex of awake macaques. *Neuron* 48, 139–148. doi: 10.1016/j.neuron.2005.08.039

Moore, B. C. J., and Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acust. United Acust.* 88, 320–333.

Moore, B. C. J., and Gockel, H. E. (2012). Properties of auditory stream formation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 919–931. doi: 10.1098/rstb.2011.0355

Neff, D. L., Jesteadt, W., and Brown, E. L. (1982). The relation between gap discrimination and auditory stream segregation. *Percept. Psychophys.* 31, 493–501. doi: 10.3758/BF03204859

Pressnitzer, D., Sayles, M., Micheyl, C., and Winter, I. M. (2008). Perceptual organization of sound begins in the auditory periphery. *Curr. Biol.* 18, 1124–1128. doi: 10.1016/j.cub.2008.06.053

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proc. Natl. Acad. Sci.U.S.A.* 98, 676–682. doi: 10.1073/pnas.98.2.676

Reiterer, S. M., Erb, M., Droll, C. D., Anders, S., Ethofer, T., Grodd, W., et al. (2005). Impact of task difficulty on lateralization of pitch and duration discrimination. *Neuroreport* 16, 239–242. doi: 10.1097/00001756-200502280-00007

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., and Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science* 306, 443–447. doi: 10.1126/science.1100301

Rinne, T., Koistinen, S., Salonen, O., and Alho, K. (2009). Task-dependent activations of human auditory cortex during pitch discrimination and pitch memory tasks. *J. Neurosci.* 29, 13338–13343. doi: 10.1523/JNEUROSCI.3012-09.2009

Roberts, B., Glasberg, B. R., and Moore, B. C. J. (2002). Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. *J. Acoust. Soc. Am.* 112, 2074–2085. doi: 10.1121/1.1508784

Snyder, J. S., Alain, C., and Picton, T. W. (2006). Effects of attention on neuroelectric correlates of auditory stream segregation. *J. Cogn. Neurosci.* 18, 1–13. doi: 10.1162/089892906775250021

Szalárdy, O., Bendixen, A., Tóth, D., Denham, S. L., and Winkler, I. (2013). Modulation-frequency acts as a primary cue for auditory stream segregation. *Learn. Percept.* 5, 149–161. doi: 10.1556/LP.5.2013.Suppl2.9

Thompson, S. K., Carlyon, R. P., and Cusack, R. (2011). An objective measurement of the build-up of auditory streaming and of its modulation by attention. *J. Exp. Psychol. Hum.* 37, 1253–1262. doi: 10.1037/a0021925

Tregellas, J. R., Davalos, D. B., and Rojas, D. C. (2006). Effect of task difficulty on the functional anatomy of temporal processing. *Neuroimage* 32, 307–315. doi: 10.1016/j.neuroimage.2006.02.036

Van Noorden, L. P. A. S. (1975). Temporal coherence in the perception of tone sequences. Ph.D. thesis, Eindhoven University of Technology.

Vliegen, J., Moore, B. C. J., and Oxenham, A. J. (1999). The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. *J. Acoust. Soc. Am.* 106, 938–945. doi: 10.1121/1.427140

Vliegen, J., and Oxenham, A. J. (1999). Sequential stream segregation in the absence of spectral cues. *J. Acoust. Soc. Am.* 105, 339–346. doi: 10.1121/1.424503

Wilson, E. C., Melcher, J. R., Micheyl, C., Gutschalk, A., and Oxenham, A. J. (2007). Cortical FMRI activation to sequences of tones alternating in frequency: relationship to perceived rate and streaming. *J. Neurophysiol.* 97, 2230–2238. doi: 10.1152/jn.00788.2006

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

Received: 19 December 2013; accepted: 03 May 2014; published online: 06 June 2014.
Citation: Dolležal L-V, Brechmann A, Klump GM and Deike S (2014) Evaluating auditory stream segregation of SAM tone sequences by subjective and objective psychoacoustical tasks, and brain activity. Front. Neurosci. ***8****:119. doi: 10.3389/fnins.2014.00119*
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in
NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 12 June 2014
doi: 10.3389/fnins.2014.00144

Decreased ability in the segregation of dynamically changing vowel-analog streams: a factor in the age-related cocktail-party deficit?

Pierre Divenyi [1,2]*

[1] *Department of Music, Center for Computer Research in Music and Acoustics, Stanford University, Stanford, CA, USA*
[2] *Speech and Hearing Research, Veterans Affairs Northern California Health Care System, Martinez, CA, USA*

Edited by:
Elyse S. Sussman, Albert Einstein College of Medicine, USA

Reviewed by:
Mounya Elhilali, Johns Hopkins University, USA
Torsten Rahne, Universitätsklinikum Halle (Saale), Germany
Jennifer Lentz, Indiana University, USA

***Correspondence:**
Pierre Divenyi, Center for Computer Research in Music and Acoustics, Stanford University, 360 Lomita Court, Stanford, CA 94305, USA
e-mail: pdivenyi@ccrma.stanford.edu

Pairs of harmonic complexes with different fundamental frequencies f0 (105 and 189 Hz or 105 and 136 Hz) but identical bandwidth (0.25–3 kHz) were band-pass filtered using a filter having an identical center frequency of 1 kHz. The filter's center frequency was modulated using a triangular wave having a 5-Hz modulation frequency f_{mod} to obtain a pair of vowel-analog waveforms with dynamically varying single-formant transitions. The target signal S contained a single modulation cycle starting either at a phase of $-\pi/2$ (up-down) or $\pi/2$ (down-up), whereas the longer distracter N contained several cycles of the modulating triangular wave starting at a random phase. The level at which the target formant's modulating phase could be correctly identified was adaptively determined for several distracter levels and several extents of frequency swing (10–55%) in a group of experienced normal-hearing young and a group of experienced elderly individuals with hearing loss not exceeding one considered moderate. The most important result was that, for the two f0 differences, all distracter levels, and all frequency swing extents tested, elderly listeners needed about 20 dB larger S/N ratios than the young. Results also indicate that identification thresholds of both the elderly and the young listeners are between 4 and 12 dB higher than similarly determined detection thresholds and that, contrary to detection, identification is not a linear function of distracter level. Since formant transitions represent potent cues for speech intelligibility, the large S/N ratios required by the elderly for correct discrimination of single-formant transition dynamics may at least partially explain the well-documented intelligibility loss of speech in babble noise by the elderly.

Keywords: speech perception, aging, auditory scene analysis, formants, frequency modulation

INTRODUCTION

In aging various auditory functions of the individual are often impaired. Perhaps the most disturbing aspect of this impairment is the significantly reduced ability to understand speech in social noise or a reverberant environment, commonly referred to as the loss of the "cocktail-party effect" (CPE). Although, expectedly, this deficit is exacerbated by presbycusic—the typical age-related sensorineural high-frequency elevation of auditory thresholds (Carhart and Tillman, 1970)—it is often also experienced by elderly individuals with normal audiograms or having, at worst, a mild-to-moderate hearing loss (Dubno et al., 1984; Divenyi and Haupt, 1997; Snell et al., 2002). Causes of the CPE deficit in the elderly are complex. On the peripheral end, hearing loss has been for long known to affect speech understanding in babble noise, regardless of age (Humes et al., 1994). On the other end of the spectrum, age-related cognitive decline has also been implicated, be it decreased selective attention to concurrent speech (Sommers, 1997), impaired short-term recall of words (Murphy et al., 2000), or reduced working memory capacity (Ng et al., 2013). These factors are among those recognized to increase in the mental effort required when the elderly listens to speech in a CPE setting (Zekveld et al., 2011). But, between peripheral and cognitive extremes there is a host of sensory/nervous system processes indispensable for understanding speech in interference that are also deficient. One group of these are deficits of temporal processing in diverse time ranges, such as gap detection and discrimination necessary for the perception of stop consonants and affricates (Snell and Frisina, 2000), duration discrimination (Fitzgibbons and Gordon-Salant, 1994) affecting accurate perception of subsyllabic and syllabic segments, temporal modulation transfer functions (He et al., 2008) and resistance to modulation interference (Bacon and Takahashi, 1992; Humes et al., 2013), formant transition discrimination (Elliott et al., 1989), and temporal-order discrimination (Fitzgibbons and Gordon-Salant, 1998). A second group is related to localization, which is also known to be impaired in aging (Herman et al., 1977; Abel et al., 2000). This impairment makes CPE performance poorer by reducing or altogether canceling the 2.5-to-4 dB release from masking provided by spatial separation of the target and the interference (Ihlefeld and Shinn-Cunningham, 2008).

From a strictly auditory standpoint, CPE can be regarded as an instance of masking with target speech as the signal and interference as the noise. However, since in speech the respective frequency ranges of the target and the interference seldom perfectly overlap, the rules derived from decades' worth of tone-in-noise *energetic* masking research will be applicable only to specific speech segments. As proposed by authors investigating CPE in laboratories across different continents (Brungart et al., 2001; Cooke et al., 2008) the overwhelming portion of speech in babble noise the masking is *informational*, due to the similarity of the target and the interference (Lee and Richards, 2011). Research over the last few decades uncovered many aspects of informational masking (Watson, 1987; Lutfi, 1990; Kidd et al., 1994; Oh and Lutfi, 1998; Freyman et al., 2004) and was able to quantitatively specify the differences between the two modes of masking (Arbogast et al., 2002; Brungart and Simpson, 2002; Durlach et al., 2003b). As to informational masking in aging, some studies have shown elderly listeners to be more affected than the young (Freyman et al., 2008; Rajan and Cainer, 2008), while others found no age differences (Agus et al., 2009; Ezzatian et al., 2011). The disagreement between these results may stem from the lack of uniformly accepted definition of informational masking other than being different from energetic masking—a tautology pointed out by Durlach (Durlach et al., 2003a)—but also from the inherent difficulty of controlling for the ensemble of physical parameters of speech. However, the lack of age effect could also be due to the elderly listener using his/her experience to compensate for a low speech-to-noise ratio by relying on predictions derived from overlearned patterns (Divenyi, 2005).

CPE can be also viewed as the instance of auditory scene analysis (ASA, Bregman, 1990) most important for verbal communication. In fact, in a CPE situation the listener must continuously segregate a speech target stream from the babble stream or streams. While young normal-hearing individuals are able to understand speech in CPE settings even under quite unfavorable signal-to-noise ratios (SNR's), the SNR elderly individuals require is significantly higher (Gelfand et al., 1988; Snell et al., 2002), even when these individuals suffer from no or only mild presbycusic hearing loss (Divenyi and Haupt, 1997). The way ASA understands speech segregation of target from non-target speech is that harmonics of the fundamental frequency (f0) each of the simultaneous voices are grouped, thereby allowing the listener to focus on the harmonics of the target voice alone—as demonstrated by experiments on the segregation of non-speech harmonic complexes (Micheyl and Oxenham, 2010) and synthesized as well as natural speech sounds (Darwin et al., 2003; Roman and Wang, 2006; Lavandier and Culling, 2008). Segregation of concurrent vowels (Assmann and Summerfield, 1990), or speech of concurrent talkers (Darwin et al., 2003), is easier when their f0's are widely separated (e.g., as in the voices of different gender talkers) and becomes increasingly difficult as the difference between f0's decreases. Temporal asynchrony of vowels (Darwin and Hukin, 1998) or words (Lee and Humes, 2012) also facilitates their segregation. The ability to segregate one vowel in an ensemble of concurrent vowels increases when the f0 of one or several in the ensemble is modulated by a low-frequency sinusoid (i.e., when it undergoes a *vibrato*) (McAdams, 1990). In rooms, the target and non-target talkers are in spatially separated locations allowing the auditory system to segregate them, as shown in binaural experiments (Brungart and Simpson, 2002, 2007; Hawley et al., 2004).

But, looking from a broad perspective, speech is a dynamic signal characterized by constant *changes*. The changes can be defined in various ways, such as on the level of acoustics (fluctuating envelope, fundamental frequency variations, formant transitions, etc.), articulatory phonetics (gestural movements), descriptive phonetics and phonology (sequences and clusters of phonemic and sub-phonemic units), or higher-order linguistics (sequences of morphemes, words, word strings, sentences, sentence strings). Although computational characterization, and modeling, of these changes is nearly impossible at the higher levels of analysis, a mathematical formulation of the transform of acoustic signals to activity patterns observed at the cortical level, the complex modulation spectrum based on Gabor's wavelet transform (Gabor, 1946) has been gaining acceptance. Although Gabor conceived it for the reduction of information "atoms" in audio (i.e., telephone) communication, the transform and its inverse have been widely used for the analysis and synthesis of images (Levi and Stark, 1983) before being adopted for the analysis of audio signals (Pitton et al., 1996) and to models of the auditory system beyond peripheral analysis (Kowalski et al., 1996). Recognizing that both the temporal and spectral envelopes of natural (i.e., complex) sounds contain peaks that change over time, the transform represents the *spectrum* of these peaks as *modulations* in the temporal (rate, in Hz) and spectral (scale in cycles per octave) domains. By choosing appropriate parameters for this model, called the "spectro-temporal receptive field" (STRF) model, it has been demonstrated that auditory cortex activity in the ferret (Chi et al., 2005) or in the song bird (Singh and Theunissen, 2003), as well as temporal-parietal cortical responses recorded with an electrode grid placed on the surface of patients awaiting epilepsy surgery (Mesgarani et al., 2008), can be fairly accurately modeled using this transform. In agreement with Plomp (1983)—"…speech is a signal slowly varying in amplitude and frequency"—and with the 4-Hz major mode of the temporal modulation spectrum of speech (Greenberg et al., 2003), shown to be language-independent (see e.g., Arai and Greenberg, 1997) speech input to this model shows that the predominant temporal modulation rate is slow and so is the scale of frequency peak shifts during relatively stable segments (e.g., vowels, fricatives, nasals, Elliott and Theunissen, 2009). Because the transform effectively uncovers patterns and features of complex signals—speech, music, animal sounds, and environmental sounds—it has been used as a tool for the separation of concurrent auditory streams in a cocktail-party situation (Elhilali and Shamma, 2008; Mesgarani et al., 2011). Continuing this line of thought, if the auditory system uses temporal and spectral modulations to picture our acoustic world and to separate auditory objects, then studying the perception of signals modulated in amplitude and/or frequency, as well as its impairments, should bring us closer to the understanding of the success and failures of listening in a CPE setting. Thus, data on modulation detection/discrimination interference (MDI) in the amplitude (Moore et al., 1995; Moore and Sek, 1996), or frequency (Lyzenga and Carlyon, 1999) domains

not only reveal parametric limitations of the Gabor transform applied to audition but also quantitatively describe the dynamic temporal and spectral map inside the existence limit of the CPE.

The present study continues the above line of reasoning in a set of experiments aimed at better understanding components of the deficiency elderly individuals display when listening to speech in the presence of speech interference. Since an earlier study showed the effect of duration and velocity on the perception of vowel transitions (Divenyi, 2009), and since the perception of frequency transitions in aging has been shown to correlate with intelligibility (Gordon-Salant et al., 2007), the experiments were focused on the way, and the extent to which, identification of a transition of interest is affected by the presence of an similar transition. The experiments used a single-formant simplified analogs of a target vowel and of an interfering vowel, each having a fixed f0 and a formant peak modulated in frequency. A similar stimulus configuration was used in studies by Lyzenga and Carlyon (1999, 2005) focused on the effects of the difference between either modulating or fundamental frequency, and of spectral content. In contrast, the question the present experiments addressed was the target-to-distracter ratio (TDR) necessary for a formant peak modulation pattern to be identified by normal-hearing young, and by elderly listeners without appreciable hearing loss.

MATERIALS AND METHODS

Stimuli in the experiments consisted of pairs of harmonic complexes: a target stream presented simultaneously with an interfering distracter stream. The 800-ms distracter stream started 200 ms before the 400-ms target stream. The two streams had different fundamental frequencies (f0), one always 107 Hz for one of the streams and either 136 or 189 Hz for the other, thereby producing two different fundamental frequency separations (Δf0), one wide (Δf0/f0 = 0.77, approximately corresponding to the minor seventh musical interval) and one narrow (Δf0/f0 = 0.27, approximately corresponding to the major third). At each Δf0 separation, the higher f0 was assigned to the target in half of the conditions while it was assigned to the distracter in the other half. The spectrum of both streams contained only harmonics inside the 250- to 3000-Hz band. Because this constraint resulted in a certain degree of difference between the perceived salience of the two streams, Terhardt's algorithm (Terhardt et al., 1982) was used to generate streams the salience of the dominant temporal ("virtual") pitch of which was comparable. Both streams were spectrally shaped to produce single-formant pseudo-vowels by passing them through second-order band-pass filters with a 6 dB/octave falloff, i.e., filters with formant characteristics not unlike that of natural vowels. The target stream's formant frequency F_T was held constant at 1 kHz during the first and last 100 ms of its duration, while during the central 200 ms the F_T was modulated with a 5-Hz triangular wave that went for 100 ms in one direction and for 100 ms in the other, thereby creating formant trajectory patterns in which F_T was changing either up-down or down-up starting from, and returning to, an F_T of 1 kHz, with a maximum formant swing of ΔF_T Hz. In the distracter stream, the formant frequency F_D was also modulated with a 5-Hz triangular wave, except that the modulator waveform was during the whole duration of the distracter, creating a continuous up-down-up-down pattern. The extent of the distracter's formant trajectory was also larger than that of the target: the top and the bottom formant frequency extremes were 1800 and 525 Hz, i.e. two frequencies equally distant from 1-kHz, a frequency at the

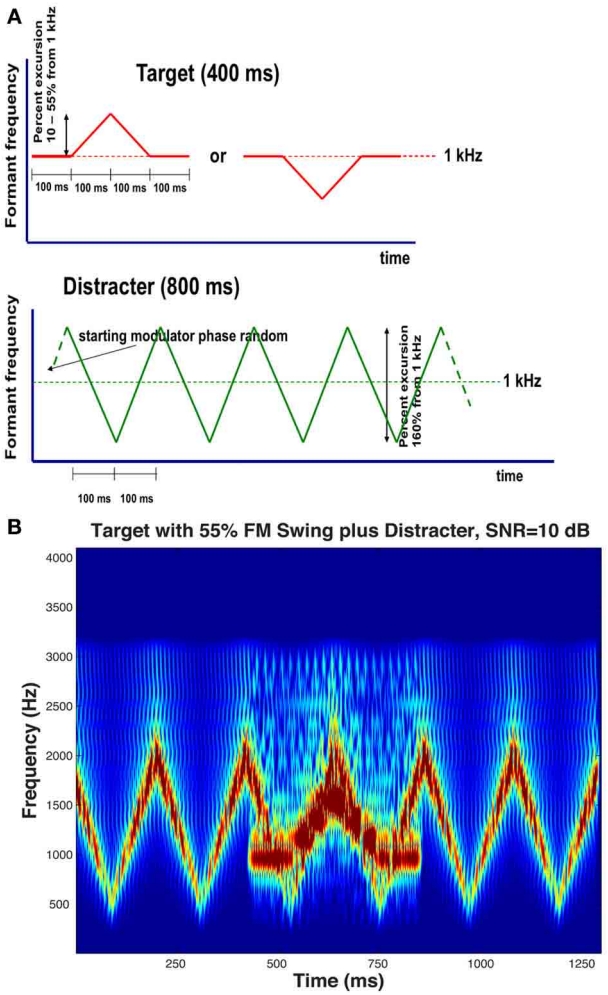


FIGURE 1 | (A) Schematic stimulus diagram, illustrating trajectory of the single formant's peak in the target and in the distracter, frequency-modulated using a 5-Hz triangular wave. The upper traces illustrate the two target formant peak trajectory patterns that the listeners had to identify: a 100-ms ascending followed by a 100-ms descending trajectory or a 100-ms descending followed by a 100-ms ascending trajectory, both trajectories flanked by two 100-ms steady-state portions with a constant formant peak at 1 kHz. The lower trace illustrates the trajectory of the distracter that consisted of an 800-ms pattern of repeating 100-ms ascending and descending trajectories between the 525 and 1800-Hz formant peak minimum and maximum that are equidistant from the 1-kHz center on a basilar membrane distance scale. The starting phase of the distracter's modulating waveform varied randomly from trial to trial. All formant trajectories were linear on this distance scale. Both the target and the distracter were gated using a 25-ms cosinusoidal window to ensure that the onset and the offset of both were smooth. **(B)** Spectrogram of a trial. The target is an up-down transition with an FM excursion reaching 1550 Hz, i.e., 55% higher than the resting 1-kHz formant frequency. The SNR of the example shown, 10 dB, is larger than most conditions used in the experiments.

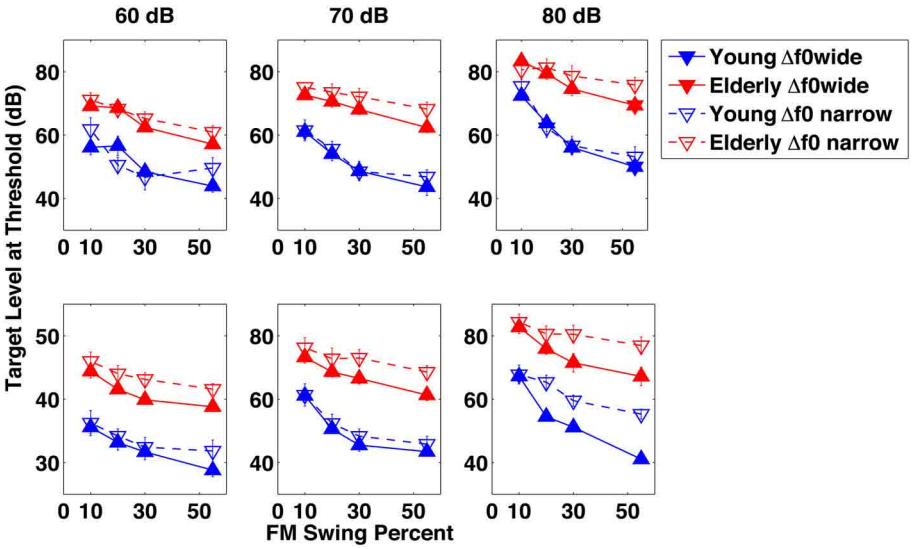


FIGURE 2 | Results of Experiment 1. Target level at identification threshold in dB as a function of the target formant's frequency modulation swing expressed as percent maximum displacement from 1 kHz, across three distracter levels in dB, in separate columns of graphs. The elderly group's data are in red and those of the young group in blue. Filled symbols represent results for conditions in which the fundamental frequency separation Δf0/f0 between the target and the distracter was wide (0.77), whereas data marked by the empty symbols are for conditions with a narrow (0.27) Δf0/f0. In the top row of graphs the data shown represent for conditions in which the fundamental frequency f0 of the target was always *lower* than that of the distracter, whereas in the bottom row of graphs they represent conditions in which the fundamental frequency f0 of the target was always *higher* than that of the distracter.

center of the low and high extremes on Greenwood's (1962) scale of basilar membrane distances. Schematic formant frequency-vs.-time diagrams of the target and the distracter are shown in **Figure 1A**, with an audio example presenting the target, the distracter, and their combination. Throughout the experiments, the starting modulation phase of the distracter randomly varied from trial to trial. **Figure 1B** displays a spectrogram representing a trial having an up-down target a large, 55%, formant excursion embedded in the distracter; the TDR is +10 dB—a ratio larger than most used in the experiments proper and is shown here mainly for illustrative purposes.

The study included two experiments. The objective of the first was to examine how the ensemble of stimulus parameters affected the threshold of *discriminability* of formant transition patterns. The objective of the second experiment was to examine the effect of the stimulus parameters on the threshold of *detectability* (i.e., audibility) of the target. In Experiment 1, the subject performed a single-interval two-alternative forced choice task that consisted of *identifying* whether the formant trajectory pattern in the target stream was up-down or down-up, while ignoring the distracter stream. In each block of trials $\Delta F_T/F_T$, the frequency excursion (i.e., the swing) of the target, remained fixed at 10, 20, 30, or 55 percent and the overall level of the distracter was held constant at 60, 70, or 80 dB SPL. The difference between the fundamental frequencies of the target and the distracter, Δf0/f0, was narrow or wide and varied from condition to condition, and so did the assignment of the higher or the lower of the fundamental frequencies to the target stream (and the other fundamental frequency to the distracter). In Experiment 2 the subject performed a two-interval two-alternative task in which he/she had to *detect* whether the target formant pattern was present in the first or the second interval, with the distracter being presented in both intervals. Two formant swing extents, 10 and 55 percent, were investigated with the distracter level constant at 60, 70, or 80 dB SPL. The higher fundamental frequency was always assigned to the target and the fundamental frequency separation was always the wide one ($\Delta f0/f0 = 0.77$).

Stimuli were digitally stored and delivered by a PC computer using an Echo Gina analog converter system connected to Tucker and Davis Technology filters and digital attenuators, and delivered diotically to headphones (Sennheiser SH 250). In each run of trials in both experiments the level of the target stream was varied adaptively from a starting point of 90 dB SPL to track the 79.4% correct performance threshold. The initial step size was 5 dB and was reduced with whenever the subject gave three consecutive correct responses first to 2, and then to 1 dB. The run was terminated at the tenth reversal and threshold in each run was calculated as the average of the target's dB level at the last eight reversals. The threshold estimate for each subject and each condition was the arithmetic mean of thresholds obtained in six to eight runs.

Listener performance was assessed for subjects in two groups. The young group included 17 normal-hearing individuals between 19 and 29 years of age (average 22.0 ± 3.4 years). The elderly group included 12 elderly individuals between 61 and 82 years of age (average 69.0 ± 6.7 years) in Experiment 1, 10 of whom also participated in Experiment 2. Their hearing impairment, when present, was a mild-to-moderate presbycusic sensorineural loss; the mean of the group's pure-tone average thresholds between 0.5 and 4 kHz was 19.3 ± 14.2 dB SPL. Several

Table 1 | ANOVA of target identification data.

Analysis of variance					
Source	**Sum sq.**	***df***	**Mean sq.**	***F***	**Prob > *F***
SUB(Sgroup)	58931.9	27	2182.7	77.92	0
Dlevel	28,678	2	14,339	511.91	0
Trghi/Lo	1941.7	1	1941.7	69.32	0
Df0	622.8	1	622.8	22.23	0
Swing	33,475	3	11158.3	398.36	0
Sgroup	48032.7	1	48032.7	1714.79	0
SUB(Sgroup) ∗ Dlevel	6475.1	54	119.9	4.28	0
SUB(Sgroup) ∗ Trghi/Lo	2826.3	27	104.7	3.74	0
SUB(Sgroup) ∗ Df0	737.6	27	27.3	0.98	0.5011
SUB(Sgroup) ∗ Swing	6641.1	81	82	2.93	0
Dlevel ∗ Trghi/Lo	485.4	2	242.7	8.66	0.0002
Dlevel ∗ Df0	13.4	2	6.7	0.24	0.7869
Dlevel ∗ Swing	685.6	6	114.3	4.08	0.0005
Dlevel ∗ Sgroup	1161.3	2	580.7	20.73	0
Trghi/Lo ∗ Df0	194.4	1	194.4	6.94	0.0085
Trghi/Lo ∗ Swing	1982.1	3	660.7	23.59	0
Trghi/Lo ∗ Sgroup	1292.2	1	1292.2	46.13	0
Df0 ∗ Swing	386.9	3	129	4.6	0.0033
Df0 ∗ Sgroup	89.6	1	89.6	3.2	0.0739
Swing ∗ Sgroup	956.2	3	318.7	11.38	0
Error	32016.3	1143	28		
Total	229346.1	1391			

Factors: Sgroup—elderly/young groups; Dlevel—Distracter level; Targhi/Lo—f0 of target higher or lower than f0 of distracter; Df0—f0 separation wide/narrow; Swing—Maximum of target formant's excursion.

elderly subjects had normal hearing and none had hearing loss exceeding 25 dB under 3 kHz, that is, in the frequency region of the stimuli. Subjects were tested individually in sessions that lasted 1 h a day. All subjects received training with stimuli used in both experiments. Data collection for each subject was started after he/she obtained a score of at least 95 percent correct in two contiguous 60-trial runs using the largest formant excursion (55%) without the distracter present, and a score of at least 80 percent correct in two 60-trial runs with the distracter present at 60 dB SPL, using the 55% formant excursion, and a constant target level of 80 dB SPL. Typically, young subjects needed one-and-half session to reach these criteria, whereas elderly subjects needed, on the average, two-and-half sessions. Subject testing procedures were fully consistent with experimental protocols approved by the V.A. Northern California Health Care System's Institutional Review Board.

RESULTS

EXPERIMENT 1—IDENTIFICATION OF FORMANT TRAJECTORY PATTERNS

Figure 2 illustrates the results of the experiments on the identification of a dynamically changing single- and formant target pattern in the background of a dynamically changing single-formant distracter. The figure represents threshold level of the target pattern for the average of the elderly (red lines and symbols) and the young (blue lines and symbols) subject groups, as a function of the extent of the target formant's excursion across the three distracter levels. All panels compare results of the wide (Δf0/f0 = 0.77, approximately minor seventh, solid lines) and narrow (Δf0/f0 = 0.27, approximately major third, dashed lines) f0 separations. In the top panels the target's f0 is lower and in the bottom panels it is higher than that of the distracter.

Looking at the young group's data within each and across all of the four figure panels, several general observations can be made (lower target thresholds indicating better performance):

(1) Increasing distracter level from 60 to 70 and 80 dB resulted in increased target levels. For a 20 dB increase in the distracter level a 10.6 dB target level increase was required, suggesting that the target level was a compressed nonlinear function of the distracter level.
(2) Decreasing the excursion resulted in an increase of the target threshold, although substantial increase was seen mainly at the smallest (10%) excursion extent. The difference between target levels at the easiest (55%) and hardest (10%) swing extents averaged across all conditions was large (16.2 dB). Because most of the target level increase occurred between the two smallest swings (10% and 20%), the target level was an expansive nonlinear function of the formant excursion.
(3) At both f0 separations, when the target's f0 was lower than distracter's the task was easier, resulting in target thresholds 2.75 dB lower on the average across all conditions.
(4) The large f0 separation was easier than the narrow one, resulting in target thresholds 4.6 dB lower on the average across all conditions.
(5) At the threshold of identifiability, the TDR was −22.40 dB at the easiest and −4.61 dB at the hardest condition, that is, the level of the target was below that of the distracter even when identifiability of the target was most difficult.

The trend of the elderly subjects' data mirrors that of the young subjects. In general, the differences within the elderly group's data with regard to swing, f0 separation, and target f0 are comparable to, or somewhat smaller than, those exhibited by the young subject group. The target level-distracter level nonlinearity for the elderly is a little smaller than for the young: a 12.6 dB target level increase for a 20 dB distracter level increase. Averaged across conditions, the target level difference between the easiest and hardest swing conditions was 9.6 dB, between the assignments of the higher f0 to the target or to the distracter was 4.02 dB, and between the two f0 separation was a mere 0.2 dB. However, when comparing the elderly and the young groups' results, one striking feature appears: the elderly subjects' target levels are about 20 dB higher than those of the young subjects, in all experimental conditions. In other words, in order to identify the target formant pattern as up-down or down-up, elderly listeners needed a target intensity about 20 dB higher than the young, regardless of the condition. Expressing the results as TDR, average TDR for the young was −22.4 dB in the easiest and −4.60 in the most difficult condition, whereas TDR for the elderly was −2.85 and 4.34 dB, respectively, for the two difficulty degrees. [The easiest

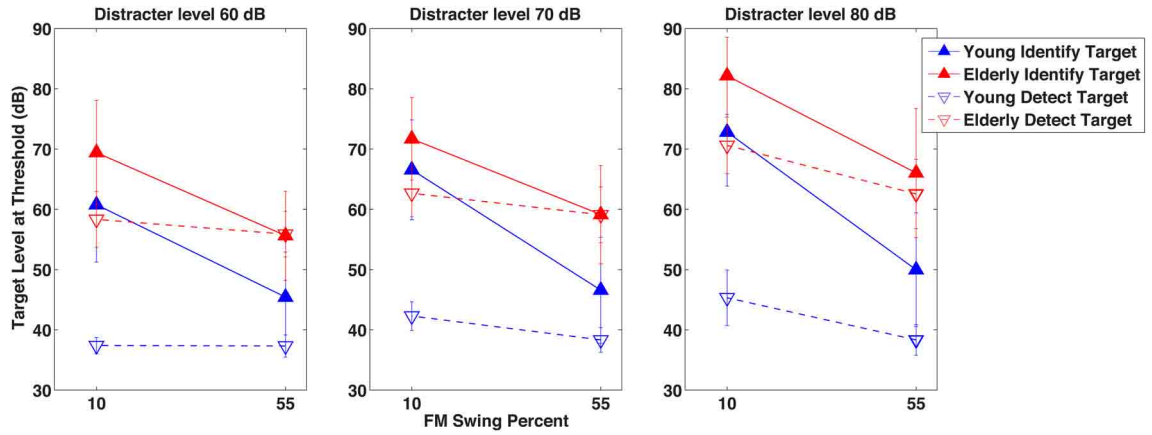


FIGURE 3 | Results of Experiment 2. A comparison of detection and identification of the target in the presence of the distracter. As in **Figure 2**, target level at threshold is shown in dB as a function of the target formant's FM excursion expressed as percent maximum displacement from 1 kHz. The three separate graphs indicate data for three distracter levels in dB. Unfilled symbols represent threshold levels for the *detection* of the target, whereas filled symbols represent thresholds for the *identification* of the target. The fundamental frequency separation Δf0/f0 of the target and the distracter was *wide* (0.77) and the fundamental frequency f0 of the target was always *higher* than that of the distracter.

Table 2 | ANOVA of data from detection vs. identification tasks.

Analysis of variance					
Source	**Sum sq.**	***df***	**Mean sq.**	***F***	**Prob > *F***
SUB(Sgroup)	8922.3	25	356.9	11.86	0
Sgroup	8352.9	1	8352.9	277.67	0
Dlevel	285.8	2	142.9	4.75	0.0097
Det/Ident	14427.4	1	14427.4	479.61	0
Swing	2724.9	1	2724.9	90.58	0
SUB(Sgroup) ∗ Dlevel	1855.4	50	37.1	1.23	0.1612
SUB(Sgroup) ∗ Det/Ident	3781.2	25	151.2	5.03	0
SUB(Sgroup) ∗ Swing	605.9	25	24.2	0.81	0.7318
Sgroup ∗ Dlevel	223.2	2	111.6	3.71	0.0263
Sgroup ∗ Det/Ident	19.3	1	19.3	0.64	0.4242
Sgroup ∗ Swing	303	1	303	10.07	0.0018
Dlevel ∗ Det/Ident	3686.5	2	1843.3	61.28	0
Dlevel ∗ Swing	247.3	2	123.7	4.11	0.0179
Det/Ident ∗ Swing	9287.4	1	9287.4	308.74	0
Error	5535.1	184	30.1		
Total	60526.2	323			

Factors: Sgroup—elderly/young groups; Dlevel—Distracter level; Det/Ident—Task (detection vs. identification); Swing—Maximum of target formant's excursion.

condition was that of the 60 dB SPL distracter, the widest (55%) swing, the wide f0 separation, and for the target having the higher f0 than the distracter. In the same vein, the hardest condition was that of the 80 dB SPL distracter, the narrowest (10%) swing, the narrow f0 separation, and for the target having the lower f0 than the distracter]. Thus, the elderly-young discrepancy when the task is easy is the same 20 dB as for the overall data shown in the figures but it diminishes to only 8.3 dB when the tasks are difficult.

To uncover details and to analyze the statistics of the observations, an analysis of variance was conducted. Results of the ANOVA are shown in **Table 1**. As the probability (p-) column indicates, all main effects—distracter level, formant swing extent, the size of f0 separation, and assignment of the higher f0 to target or distracter—were highly significant with $p = 0.0001$, both within and across subjects. The subject group effect was also highly significant and so were the within subject and main effect interactions, except the within subject-f0 separation effect, indicating that some subjects found the task for both Δf0/f0's equally difficult or easy. The significant interaction between both individual subjects and subject group vs. the assignment of the higher f0 to target or distracter indicates, as suggested by **Figure 2**, that the elderly, as well as some individual subjects, found it more consistently easier to identify the target pattern when the target f0 was the higher one. The significant interaction between formant swing extent and f0 assignment indicates, as both rows of graphs in **Figure 2** illustrate, that the target's f0 assignment to the higher f0 made the task easier only when the swing was relatively large, i.e., when the task itself was less difficult. The lack of significance of the Δf0/f0-distracter level and the Δf0/f0-subject group interactions indicate that frequency separation was a stable effect unaffected by the loudness of the distracter or the age of the listener.

EXPERIMENT 2—COMPARISON TARGET PATTERN DETECTION AND IDENTIFICATION

After seeing the subjects' performance in identifying the correct formant pattern, one is compelled to ask the question just how detectible the patterns were. This question was put to test in Experiment 2 in which detectability of a small number of selected target patterns was measured when its audibility was masked by the distracter used in Experiment 1. **Figure 3** illustrates with the closed symbols data of the detection experiment. The same subjects' results in the identification experiment at the same conditions are also shown for comparison with the open symbols.

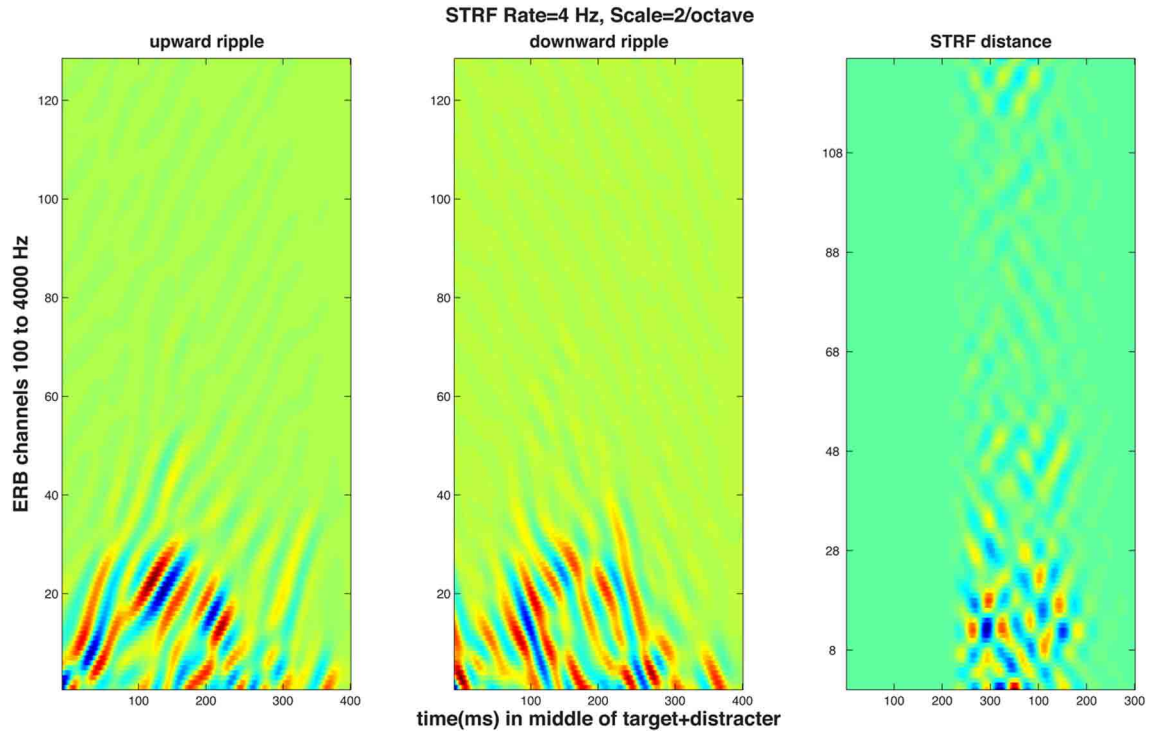


FIGURE 4 | STRF magnitude responses to a target-plus-distracter stimulus. The graphs represent simulated cortical response patterns of 128 channels tuned to frequencies aligned on an equivalent rectangular bandwidth (ERB) scale, as a function of time over the 600-ms middle portion of the stimulus containing 600-ms of the distracter and 400 ms of the target (see **Figure 1**) starting 100 ms after the portion of the distracter shown. All panels refer to STRF activity of regions tuned to 4-Hz temporal modulation rate and 2 cycles/octave spectral modulation. Left and middle panels: upward (left) and downward (middle) time-frequency modulation ripple STRF responses to an up-down target pattern embedded in the distracter, i.e., activity elicited by +4 Hz and −4 Hz temporal modulation, respectively. The rightmost panel displays pixel-by-pixel Euclidean distances of STRF magnitude evoked by the up-down and the down-up target in an identical distracter, i.e., two stimuli differing only in the very center of their duration.

Elderly subjects needed the target level to be 18.6 dB higher than the young at the 60 dB SPL distracter level and 24.22 dB higher at the 80 dB SPL distracter level. At the easy (55% swing) conditions detection of the target for the young subjects required a substantially lower (between 8.1 and 11.6 dB) target level than did its identification, whereas for the elderly subjects the two tasks required the same level, except at the most intense distracter level, where identification level was 3.5 dB higher than the detection level. Contrary to identifiability of the target (as seen in Experiment 1), the extent of formant swing did not change its detectability: the swing had no influence on whether the target could be heard. Analysis of variance of these results, shown in **Table 2**, uncovered highly significant main effects and, except for distracter level, also within subject-main effect interactions. The highly significant subject group- task (detection vs. identification) and individual subject-task interactions indicate that a definite age effect for the way detection and identification are performed (and perhaps also understood). The significant group-formant swing and group-distracter level interactions show that elderly and young listeners are differentially affected by the difficulty of the task, be it detection or identification. The highly significant task-by-swing and task-by-distracter level interactions mean that factors making identification easier or more difficult had no bearing on detection, i.e., once a target was audible, many of its properties were irrelevant. This conclusion seems to have been shared by the two groups, as indicated by the non-significant subject group-by-task interaction.

DISCUSSION

Formant transitions in vowels convey important information. Although a large excursion of transition indicates phonemic change and mostly signals the presence of a diphthong, even a relatively minor dynamic change in the frequency of a formant peak contributes to intelligibility because it is one of the markers of the consonant preceding and following the vowel (Hillenbrand et al., 2001; Fogerty and Kewley-Port, 2009). The present experiments investigated the identifiability and detectability of a simplified form of these transitions in the presence of intense distracter transitions, both in the range of those indicating phonemic change—like the 55% excursions—and those signaling the identity of preceding/following consonants—like the 10–20% excursions. The FM rate chosen, 5 Hz, was also similar to syllabic rate and thus makes the results comparable to speech and to the CPE. From a psychoacoustic standpoint, the results complement the vast and detailed body of information on modulation interference and modulation masking that has established parametric limits for detection and discrimination thresholds of a target in the presence of similar interference, with respect to differences

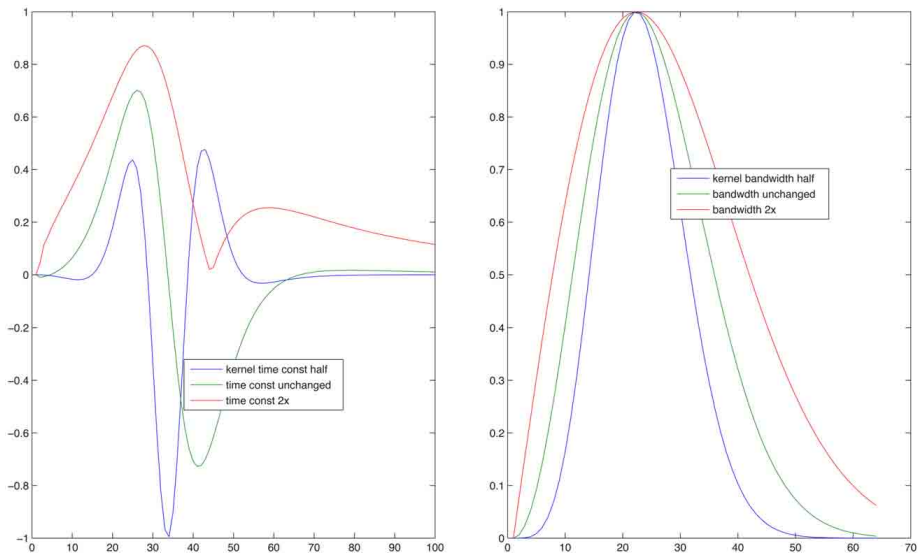


FIGURE 5 | Kernel functions used by the model of Chi et al. (2005) to generate temporal (left) and spectral (right) modulation planes at any given temporal rate and spectral scale specified. The line in the center of each of the three shown in the two panels is the unmodified kernel function and was used to simulate results of the young subjects. In both panels the lines showing functions narrower than the central one were modified to produce temporal or spectral resolutions higher than those that supposedly underlay the good performance by the young listeners, whereas the lines outlining a broader function were modified to produce lower temporal or spectral resolutions that were expected to simulate results of the elderly subjects.

in frequency, modulation rate, level, and some other properties. Although adding to this body was not the primary objective of the present study, it showed that increasing the level of interference resulted in a compressed growth in the level of the target, not only in the identification but also in the detection task. The present experiments used vowel-like harmonic target and distracter—a situation treated by only a relatively small number of studies (e.g., Shackleton and Carlyon, 1994; Lyzenga and Carlyon, 1999, 2005) that examined frequency modulation masking (FMD) for signals with f0's typically closer to each other than even those of our narrow f0 separation. Present in Lyzenga's and Carlyon's data, although not specifically pointed out in their papers, was the finding that FMD was larger when the target f0 was below that of the interferer. The present data shown in **Figure 2** clearly show a worse overall performance, especially in the difficult 10% swing conditions, when the target f0 was lower than the distracter's. One explanation could be that when the distracter f0 is higher, it will have more intense harmonics in the frequency range where the target's second-to fifth harmonics (those that are most important for carrying pitch information) are located. Western composers from the Renaissance period on (i.e., from the beginnings of accompanied melody and polyphony) have been well aware of this relationship and have customarily placed the melody intended to be heard in the treble.

Clearly, the important finding of the experiments is the impaired ability by the elderly listeners to detect and identify formant excursions in the target embedded in a distracter. Since deficits have been documented for a variety temporal processing tasks in elderly individuals with little or no presbycusic impairment (Gordon-Salant and Fitzgibbons, 1999; Humes et al., 2012), it is unlikely that our elderly listeners' deficiency may be due to the presence of their not more severe than mild-to-moderate hearing loss. The surprising finding is the large difference, 20 dB on the whole, between target identification thresholds of young and elderly subjects. Thus, one could hypothetically assume that, in addition to a small part attributable to high-frequency threshold elevation that would have diminished the contribution of higher harmonics to the strength of the definition of formants, decline of a more central, possibly cortical, site may account for the observed perceptual loss.

RELEVANCE OF THE RESULTS

Formant transitions in vowels convey important information. Although a large excursion of transition indicates phonemic change and mostly signals the presence of a diphthong, even a relatively minor dynamic change in the frequency of a formant peak contributes to intelligibility because it is one of the markers of the consonant preceding and following the vowel (Hillenbrand et al., 2001; Fogerty and Kewley-Port, 2009). The present experiments investigated the identifiability and detectability of a simplified form of these transitions in the presence of intense distracter transitions, both in the range of those indicating phonemic change—like the 55% excursions—and those signaling the identity of preceding/following consonants—like the 10–20% excursions. The FM rate chosen, 5 Hz, was also similar to syllabic rate and thus makes the results comparable to speech and to the CPE. From a psychoacoustic standpoint, the results complement the vast and detailed body of information on modulation interference and modulation masking that has established parametric limits for detection and discrimination thresholds of a target in the presence of similar interference, with respect to differences in frequency, modulation rate, level, and some other

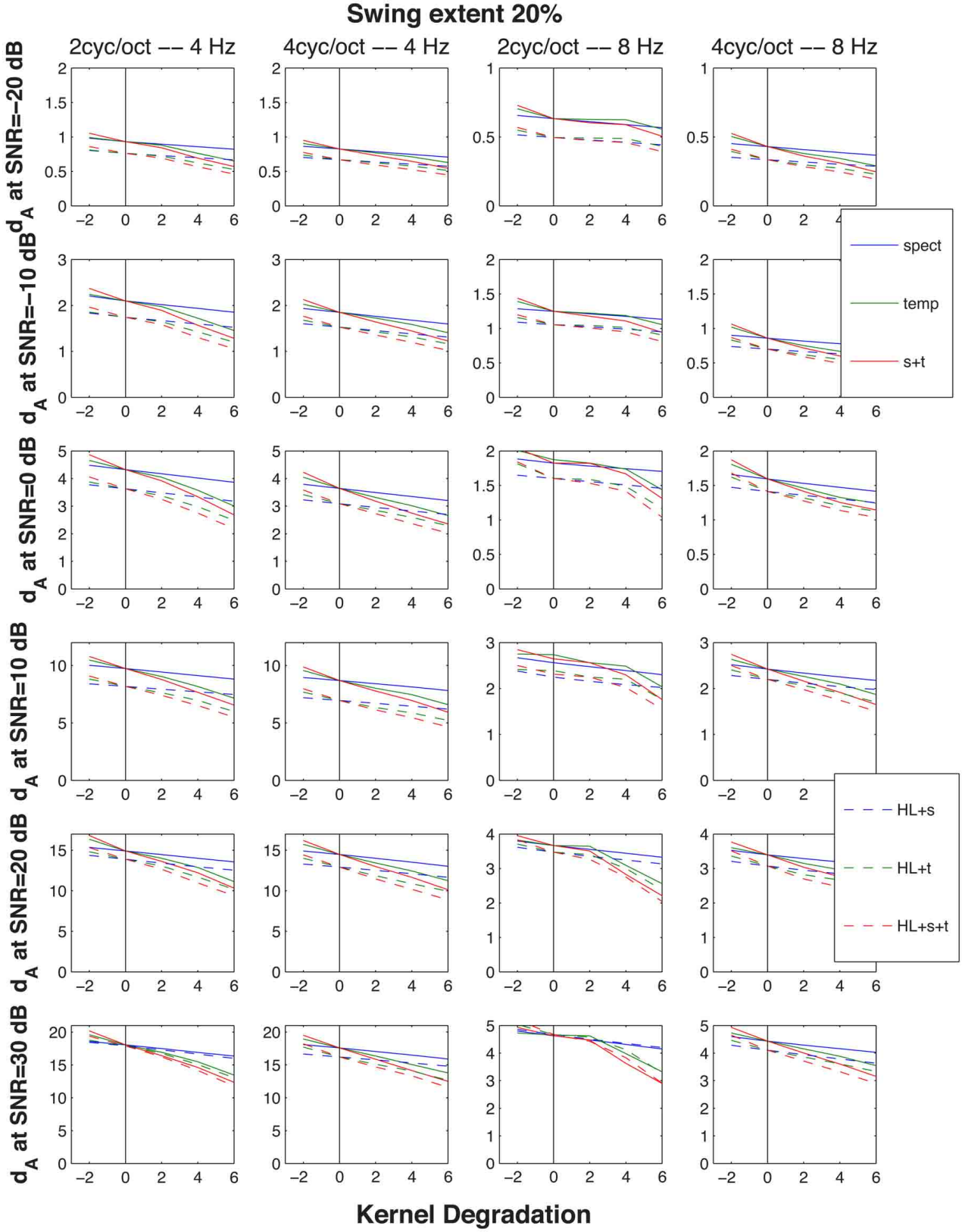


FIGURE 6 | Cumulative normalized Euclidean distances d_A between STRF patterns generated by the up-down and the down-up targets presented in an identical (magnitude and phase) distracter and identical stimulus parameters (extent of FM swing and SNR). The distance metric is considered to be proportional to the magnitude of the perceptual difference *(Continued)*

FIGURE 6 | Continued
between the two targets, hence its increasing magnitude with SNR indicated in the six rows of graphs. The abscissa indicates the size of degradation (i.e., the size of change of the kernel function): 6 dB corresponds to a 2-fold increase in the function's broadness and the vertical line at 0 marks the point at which the original, un-modified kernel functions were used, resulting in identification performance as good as demonstrated by the young subjects. The lines with negative slope show simulation by modification of the functions affecting resolution of the spectral, the temporal, or both the spectral and temporal modulations. Solid lines were generated with only the STRF kernels modified, whereas the broken lines resulted from entering the STRF module with a stimulus low-pass filtered and with a slightly reduced frequency selectivity, both intended to reflect changes encountered in elderly individuals with a moderate presbycusic sensorineural hearing loss. The leftmost tail of the lines indicate a point generated by an improved, rather than degraded, kernel function—hence the higher distance (e.g., better identification performance) it is associated with. The four columns represent the four combinations of two temporal modulation rates (4 and 8 Hz) and two spectral modulation scales (2 and 4 cycles/octave). Across all conditions in the figure a single extent of FM excursion, 20% maximum counting from the 1 kHz resting formant frequency, was used.

properties. Although adding to this body was not the primary objective of the present study, it showed that increasing the level of interference resulted in a compressed growth in the level of the target, not only in the identification but also in the detection task. The present experiments used vowel-like harmonic target and distracter—a situation treated by only a relatively small number of studies (e.g., Shackleton and Carlyon, 1994; Lyzenga and Carlyon, 1999, 2005) that examined FMD for signals with f0's typically closer to each other than even those of our narrow f0 separation. Present in Lyzenga's and Carlyon's data, although not specifically pointed out in their papers, was the finding that FMD was larger when the target f0 was below that of the interferer. The present data shown in **Figure 2** clearly show a worse overall performance, especially in the difficult 10% swing conditions, when the target f0 was lower than the distracter's. One explanation could be that when the distracter f0 is higher, it will have more intense harmonics in the frequency range where the target's second-to fifth harmonics (those that are most important for carrying pitch information) are located. Western composers from the Renaissance period on (i.e., from the beginnings of accompanied melody and polyphony) have been well aware of this relationship and have customarily placed the melody intended to be heard in the treble.

A comparison of the detection and identification results seen in **Figure 3** would be interpreted by some authors (e.g., Brungart et al., 2001) as a contrast between energetic masking and informational masking—energetic masking being considered as the process underlying detection and informational masking as a process of interference not attributable to energetic masking (Durlach et al., 2003a). One particular result, however, is incompatible with the energetic-informational masking contrast: while young listeners needed a higher SNR for identification than for detection regardless of the difficulty of the task, for the easiest condition (when the target has the highest extent of FM swing) the older listeners needed the same SNR for detection and identification.

Clearly, the major finding of the experiments is that the ability by the elderly listeners to identify and also to detect formant excursions in the target embedded in a distracter is impaired. This finding adds to a long list of deficits for a variety temporal processing tasks in elderly individuals who, just as our elderly listeners, had little or no presbycusic impairment (e.g., Gordon-Salant and Fitzgibbons, 1999; Humes et al., 2012). While peripheral auditory impairment could have had some contribution to the 20 dB effect in our results even if the threshold shifts indicated by the elderly person's audiogram were relatively minor, it is likely that some dysfunction higher up on the auditory pathway was more accountable for the loss illustrated in **Figure 2**. Obviously, it would be of interest to answer the question regarding what proportion of the loss observed in the data is attributable to peripheral and what to central impairment. This question could be addressed empirically by conducting a series of tests on the same subjects to measure a wide range of spectral and temporal auditory capabilities that, according to the literature, could differentiate peripheral and central auditory processing—such as amplitude and frequency modulation transfer functions, auditory filter width, pitch discrimination and salience, temporal processing in the 100-ms and longer ranges, auditory attention, and short-term memory for auditory stimulus details, only to cite a few. Unfortunately, such multidimensional data are not available for the subjects tested in the present experiments and the question can't be answered by analyzing the present data. Borrowing from physics, questions such as ours may be addressed indirectly by computational experiments using simulation. In our case, we could simulate normal and impaired processing of the present stimuli by using a model of the auditory system that includes both peripheral and central stages. The following subsection describes such a simulation with the help of tmodel mentioned in the Introduction, the STRF model (Chi et al., 2005). This model was chosen because it includes a peripheral and a central auditory stage, and because it permits manipulation of the efficacy of both stages.

SIMULATION OF THE RESULTS USING THE STRF MODEL

The STRF model first performs a multichannel filtering and compression akin those that take place in the cochlea and the auditory nerve, resulting in an "auditory spectrogram" with a critical band-type ERB (equivalent rectangular band) frequency scale (Patterson and Moore, 1986). This time-frequency response matrix is led to a subsequent stage in which temporal and spectral modulations are analyzed and decomposed to obtain a four-dimensional representation (time, frequency, temporal modulation by the rate of change, and spectral modulation by the scale of adjacent peaks in the spectrum). Such decomposition is known to take place in the cortex of animals (Kowalski et al., 1996) and humans (Mesgarani et al., 2008). The model is well suited for simulation of normal and impaired auditory processing on both levels because changing the threshold and the filtering in the first stage can mimic, to some extent, high-frequency presbycusic loss by the elderly. In the second stage, resolution of temporal and/or spectral modulation (i.e., grating) can be reduced by changing

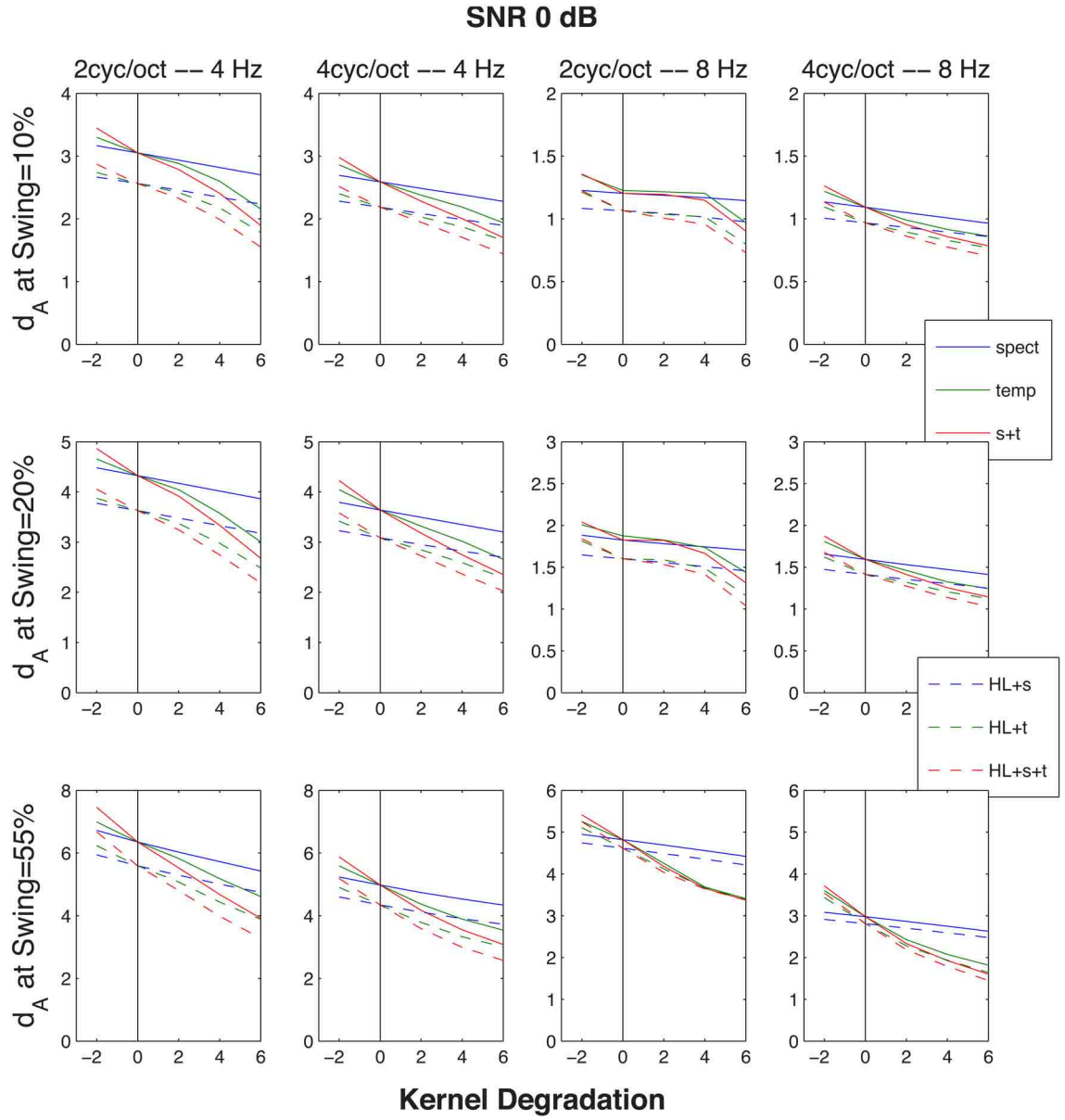


FIGURE 7 | Similar to Figure 6, except that the cumulative normalized Euclidean distances d_A between STRF patterns generated by the up-down and the down-up targets were evoked by the same target-distracter SNR of 0 dB, across three FM excursion extents (10, 20, and 55% maximum deflection re/the 1-kHz resting formant frequency), one for each row of graphs. The abscissa and the ordinate are the same as in **Figure 6** (distance vs. size of kernel degradation). The four columns of graphs represent the same four rate-cycle combinations (4, 8 Hz and 2, 4 cycles) as in **Figure 6**.

the appropriate parameters. Because of the magnitude of the effect obtained in the results, simulation of only the identification data was performed. It was assumed that identification of the up-down or the down-up pattern was based on the subject computing a distance between the four-dimensional STRF activity evoked by the two targets presented in the distracter. Since only one of the patterns was actually heard, it was further assumed that the STRF of the other target pattern was preserved and kept intact in memory, and it was available for the subject to perform the distance computation between the just-heard "pattern 1" and the previously hear "pattern 2." To perform a simulated psychophysical experiment, STRF distances could have been computed on a series of repeated trials using a distracter presented starting with a random modulating phase and a d' statistic could have been calculated from the distribution of the trial-by-trial distances. Such simulation, unfortunately, would have required computational resources that were not available. As a substitute, distances were computed between STRF patterns generated by the two targets embedded into a distracter having fixed magnitude and phase spectra. The targets were the single-formant FM patterns used in the experiments and illustrated in **Figure 1**; three of the five FM excursions used in the experiments were used in the

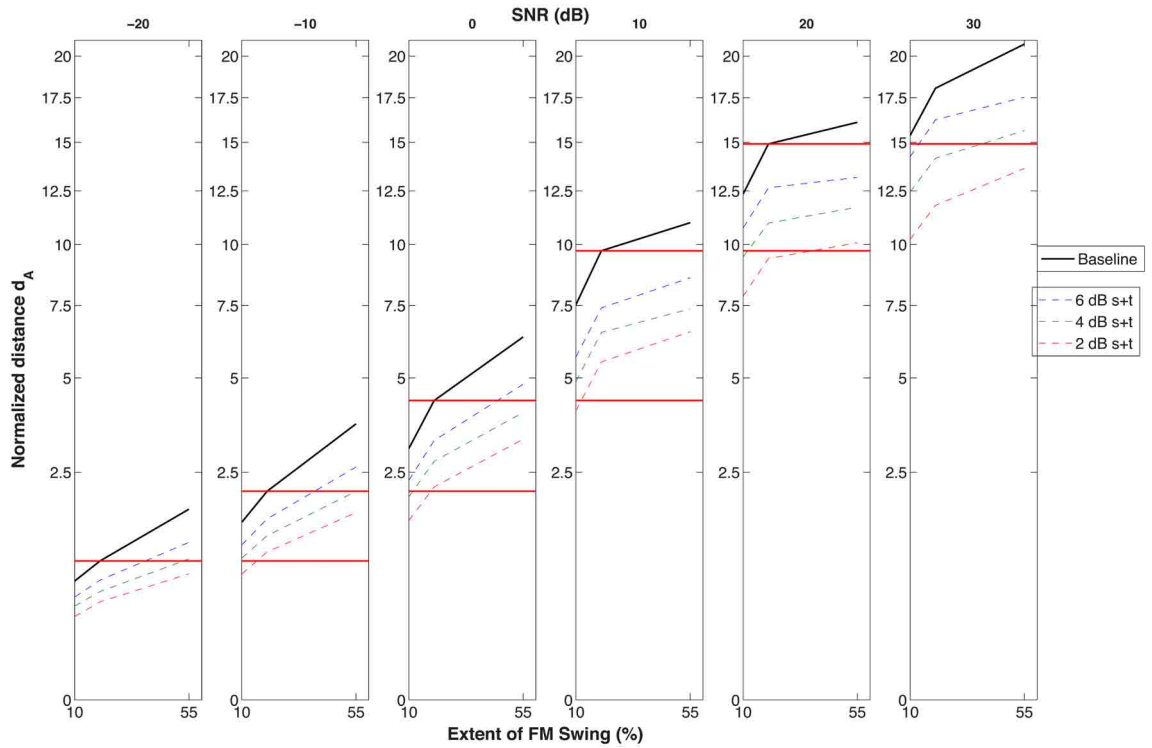


FIGURE 8 | Data simulating the almost-full set of conditions used in the identification experiments: identifiability of the targets (i.e., the cumulative normalized Euclidean distance between the STRF's evoked by the targets in distracter) as a function of three FM swing extents (10, 20, and 55%) across six SNR's. The figure wishes to illustrate that, according to this simulation at least, the distance obtained by the un-degraded STRF (simulating the young subject) at he 20% swing extent in the −20 to +20 dB SNR conditions is approximately equivalent to the distance obtained at a SNR 10 dB higher by the simulated elderly having a moderate presbycusic hearing loss and STRF's generated by spectral and temporal kernels twice their normal size, i.e., kernels that produced a markedly reduced resolution of both spectral and temporal modulations. This 10 dB loss, marked by the horizontal red lines, goes in the direction indicated by the experimental results shown in **Figure 2**, although it does not reach the 20 dB difference obtained in the psychophysical experiments.

simulation (10, 20, and 55% formant peak change with respect to the 1-kHz resting formant peak); the f0 of the target was always higher than that of the distracter and a single fundamental frequency difference Δf0 of 0.77 (the larger of the two tested in the experiments) was used. **Figure 4** illustrates STRF time-frequency response patterns to the 55% targets in the distracter presented at 0 dB SNR. The two STRF's at the left show the upward and downward grating responses to the up-down pattern, whereas the rightmost panel shows time-frequency distances between the up-down and the down-up targets. Because the two stimuli differed only in their middle 200 ms portions (see **Figure 1**), the time range in the pictures contains 400 ms starting 100 ms before and finishing 100 ms after the FM portion of the targets.

Since the objective of this simulation was to compare the model's predictions for normal and impaired listeners, auditory processing by elderly individuals was modeled in two ways. First, the typical high-frequency sloping hearing loss was emulated by passing the stimulus through a low-pass filter with a 1600-Hz corner frequency and a 6 dB/octave slope. In addition, the auditory spectrogram's filtering was made 10 percent less sharp, in order to mimic the broadening of cochlear frequency response often associated with age (Sommers and Gehr, 1998). Second, the spread of temporal (Takahashi and Bacon, 1992) and spectral (Sabin et al., 2013) modulation filters in aging was emulated by broadening the modulation filter kernels [the analogs to the Gabor (1946) transform's kernel] in the STRF model. This broadening of the two kernels is illustrated in **Figure 5**. For the data simulation, three different degrees of broadening (corresponding to factors of 1.26, 1.56, and 2, i.e., 2, 4, and 6 dB) and one degree of sharpening (a factor of 0.8, i.e., −2 dB) was used, either for the spectral, for the temporal, or both the spectral and temporal modulation filters. Results of these operations are illustrated in **Figure 6** showing d_A, the normalized (using standard deviations) cumulative Euclidean distance measure. These distances were computed between corresponding pixels of the time-frequency plane across SNR's ranging from −20 to 30 dB at one selected FM swing (20%) and are shown as a function of the degree of modulation filter degradation (with one negative degradation, i.e., improvement, as the leftmost point on each graph). The four columns of the figure display the distances for the four combinations of two the degrees of temporal modulation (2 and 4 Hz) and two degrees pf spectral modulation (2 and 4 cycles/octave). These modulation degrees were seen as being the most sensitive to the stimuli used in the study. Each graph illustrates the effect of two sets of

simulations, one (solid lines) for an intact first stage (=auditory spectrogram) input to the STRF stage, and one (broken lines) for the case in which the first stage contained a presbycusic analog low-pass filter. The difference between no-hearing-loss solid lines and the presbycusic-loss broken lines gauges the effect of the simulated peripheral hearing loss, best seen where they cross the vertical line indicating the condition in which the modulation filters were left undegraded. Such peripheral loss effect is present across all SNR's and all four rate/cycle combinations, although its size varies (between about 5% to more than 25%) and it is generally smaller for large SNR's. Comparing the three types of modulation filter degradation, combined widening of the spectral and temporal filters was the most destructive, widening of the spectral filter alone had the least effect, and widening of the temporal filter alone took place in the middle. The largest degradation effect, over 50%, is associated with a 6 dB (i.e., doubled) increase of the parameter controlling spectral and temporal modulation filter width was seen for high-SNR low-pass filtered (i.e., presbycusic) stimuli. Similar observations can be made when looking at **Figure 7** in which distance metrics across the three FM excursion extents are shown at the 0 dB SNR condition. Due to the relatively quiet signal level, absolute distance magnitudes and degradation effects are smaller than those seen in **Figure 6**.

Although the methods used and the d_A metric adopted were not optimally suited for simulating psychometric functions, a graphic projection of the effect of degradation on SNR was attempted in **Figure 8**. In this figure the d_A scale was used to allow comparison of data across SNR's and across the three FM excursion conditions (10, 20, and 55%). The performance of presbycusic-filtered inputs to the second stage that underwent the three types of degradation (broken lines) is compared with the no-filtering-no-degradation condition taken as the baseline (dark black line, emulating our young subjects). The 20% excursion taken as the criterion projected to the next, easier 10 dB higher SNR condition shows that although that particular performance level is exceeded by the baseline, it is within the range of degraded STRF processors. For instance, we see that a −10 dB SNR for the most degraded (thin red line) condition with an (interpolated) excursion between 10 and 20% will lead to a performance level identical to that of a 10 dB less loud stimulus at a 20% FM swing going through an intact (unfiltered/un-degraded) model. Similar 10 dB (or near 10 dB) effects comparing the baseline with the most degraded condition can be seen at all SNR's. While this difference between undegraded-normal and degraded-impaired simulated subjects is smaller than the 20 dB drop in performance by the elderly compared to the young in the experiments, it still suggests that a degradation *of the cortical processor* responsible for modulation filtering may at least partially account for the age effect seen.

Aside the age effect the objective was to simulate, there are some valuable hints offered by the STRF model data. As expected, one can see in both **Figures 6**, **7** that those stimuli that easier to discriminate (such as larger SNR's and larger FM swings) produce larger or much larger distances. There is, however, a potentially more important observation. Although the stimuli were not speech, they were tailored with parameters reflecting the spectral and temporal dynamics of speech. Thus, it may not be without relevance to speech processing in the cortex that the largest distances were seen for the 4-Hz temporal modulation filter at both the 2- and the 4-cycles/octave spectral modulation filter. This indicates that the most active temporal modulation filter coincides with the most prominent 4-Hz modulation rate observed for conversational speech across talkers and across languages (Greenberg et al., 2003) and that the 2-to-4-cycles/octave spectral grating coincides with distances between peaks in the spectrum that are optimal for resolving formants and formant changes in vowels (Kewley-Port and Zheng, 1999).

Despite the fact that the simulation discussed in this subsection provides only an imperfect analogy to a psychophysical experiment, it has been able to provide an answer to the question that led us to emulating the data presented in the previous sections with the help of a well-established model—the STRF—that includes peripheral and central stages of auditory processing. This simulation appears to suggest that the deficit shown by the elderly listeners in the experiments is due to two factors. The first is a peripheral loss affecting frequency and temporal resolution, but only to a lesser degree than the second. This second, more potent factor is a deficiency in the resolution of temporal and spectral modulations performed by the auditory system at a more central site, most likely in the cortex. To better understand the role of this brain mechanism in the perception of everyday speech, work should be directed toward extracting principal *features* of STRF activity and the *trajectory* of those features over time. Such future work would allow us to get a better grip on mechanisms likely to be responsible for the CPE and its decay with age.

ACKNOWLEDGMENTS

The author wishes to thank Alex Brandmeyer, Kara Haupt, and J. C. Sander for the help they provided during the experimental phase of the study, René Carré and Steven Greenberg for many helpful discussions on speech dynamics, Lakshmi Krishnan and Nima Mesgarani for their help in the STRF analysis of the signals, and the three reviewers for their very helpful comments and suggestions. The research was supported by grant AG-07998 from the National Institute on Aging and by the Veterans Affairs Medical Research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fnins.2014.00144/abstract

Audio examples—to be listened to while looking at Figure 1.

Audio example 1 | Ascending–descending target formant FM pattern (left pattern of top trace in **Figure 1**). High fundamental frequency (189 Hz) with an excursion wider than the widest used in the experiment (75%). The FM rate is 5 Hz.

Audio example 2 | Similar, but for the descending–ascending target formant FM pattern (right pattern of top trace in **Figure 1**).

Audio example 3 | 5-Hz FM pattern of the distracter. Low fundamental frequency (107 Hz). The formant swing is between the 525 and 100 Hz peak frequencies.

Audio example 4 | A waveform similar to whose presented to the listener in the identification experiment (Experiment 1). It is, actually, the pattern

of Audio example 2 presented in the middle of the distracter. The SPLs of the target and of the distracter in this example are identical. This large target-to-distracter ratio, together with the target's formant swing larger than those used in the experiment, produce a stimulus in which the target pattern would have been easily identified by all listeners in the study.

REFERENCES

Abel, S. M., Giguere, C., Consoli, A., and Papsin, B. C. (2000). The effect of aging on horizontal plane sound localization. *J. Acoust. Soc. Am.* 108, 743–752. doi: 10.1121/1.429607

Agus, T. R., Akeroyd, M. A., Gatehouse, S., and Warden, D. (2009). Informational masking in young and elderly listeners for speech masked by simultaneous speech and noise. *J. Acoust. Soc. Am.* 126, 1926–1940. doi: 10.1121/1.3205403

Arai, T., and Greenberg, S. (1997). "The temporal properties of spoken japanese are similar to those of English," in *Paper presented at the Eurospeech* (Rhodes).

Arbogast, T. L., Mason, C. R., and Kidd, J. G. (2002). The effect of spatial separation on informational and energetic masking of speech. *J. Acoust. Soc. Am.* 112, 2086–2098. doi: 10.1121/1.1510141

Assmann, P. F., and Summerfield, A. Q. (1990). Modeling the perception of concurrent vowels: vowels with different fundamental frequencies. *J. Acoust. Soc. Am.* 88, 680–697. doi: 10.1121/1.399772

Bacon, S. P., and Takahashi, G. A. (1992). Overshoot in normal-hearing and hearing-impaired subjects. *J. Acoust. Soc. Am.* 91, 2865–2871. doi: 10.1121/1.402967

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: Bradford Books (MIT Press).

Brungart, D. S., and Simpson, B. D. (2002). Within-ear and across-ear interference in a cocktail-party listening task. *J. Acoust. Soc. Am.* 112, 2985–2995. doi: 10.1121/1.1512703

Brungart, D. S., and Simpson, B. D. (2007). Effect of target-masker similarity on across-ear interference in a dichotic cocktail-party listening task. *J. Acoust. Soc. Am.* 122, 1724. doi: 10.1121/1.2756797

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *J. Acoust. Soc. Am.* 110, 2527–2538. doi: 10.1121/1.1408946

Carhart, R., and Tillman, T. W. (1970). Interaction of competing speech signals with hearing losses. *Arch. Otolaryngol.* 91, 273–279. doi: 10.1001/archotol.1970.00770040379010

Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* 118, 887–906. doi: 10.1121/1.1945807

Cooke, M., Garcia Lecumberri, M. L., and Barker, J. (2008). The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception. *J. Acoust. Soc. Am.* 123, 414–427. doi: 10.1121/1.2804952

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *J. Acoust. Soc. Am.* 114, 2913–2922. doi: 10.1121/1.1616924

Darwin, C. J., and Hukin, R. W. (1998). Perceptual segregation of a harmonic from a vowel by interaural time difference in conjunction with mistuning and onset asynchrony. *J. Acoust. Soc. Am.* 103, 1080–1084. doi: 10.1121/1.421221

Divenyi, P. (2005). "Masking the feature-information in multi-stream speech-analogue displays," in *Speech Separation by Humans and Machines*, ed P. Divenyi (New York, NY: Kluwer Academic Publishers), 269–281.

Divenyi, P. (2009). Perception of complete and incomplete formant transitions in vowels. *J. Acoust. Soc. Am.* 126, 1427–1439. doi: 10.1121/1.3167482

Divenyi, P. L., and Haupt, K. M. (1997). Audiological correlates of speech understanding deficits in elderly listeners with mild-to-moderate hearing loss. I. Age and laterality effects. *Ear Hear.* 18, 42–61. doi: 10.1097/00003446-199702000-00005

Dubno, J. R., Dirks, D. D., and Morgan, D. E. (1984). Effects of age and mild hearing loss on speech recognition in noise. *J. Acoust. Soc. Am.* 76, 87–96. doi: 10.1121/1.391011

Durlach, N. I., Mason, C. R., Kidd, G. Jr., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (2003a). Note on informational masking. *J. Acoust. Soc. Am.* 113, 2984–2987. doi: 10.1121/1.1570435

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., and Kidd, G. Jr. (2003b). Informational masking: counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. *J. Acoust. Soc. Am.* 114, 368–379. doi: 10.1121/1.1577562

Elhilali, M., and Shamma, S. A. (2008). A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. *J. Acoust. Soc. Am.* 124, 3751–3771. doi: 10.1121/1.3001672

Elliott, L. L., Hammer, M. A., Scholl, M. E., and Wasowicz, J. M. (1989). Age differences in discrimination of simulated single-formant frequency transitions. *Percept. Psychophys.* 46, 181–186. doi: 10.3758/BF03204981

Elliott, T. M., and Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Comput. Biol.* 5:e1000302. doi: 10.1371/journal.pcbi.1000302

Ezzatian, P., Li, L., Pichora-Fuller, K., and Schneider, B. (2011). The effect of priming on release from informational masking is equivalent for younger and older adults. *Ear Hear.* 32, 84–96. doi: 10.1097/AUD.0b013e3181ee6b8a

Fitzgibbons, P. J., and Gordon-Salant, S. (1994). Age effects on measures of auditory duration discrimination. *J. Speech Hear. Res.* 37, 662–670.

Fitzgibbons, P. J., and Gordon-Salant, S. (1998). Auditory temporal order perception in younger and older adults. *J. Speech Lang. Hear. Res.* 41, 1052–1060.

Fogerty, D., and Kewley-Port, D. (2009). Perceptual contributions of the consonant-vowel boundary to sentence intelligibilitya). *J. Acoust. Soc. Am.* 126, 847–857. doi: 10.1121/1.3159302

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *J. Acoust. Soc. Am.* 115, 2246–2256. doi: 10.1121/1.1689343

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (2008). Spatial release from masking with noise-vocoded speech. *J. Acoust. Soc. Am.* 124, 1627–1637. doi: 10.1121/1.2951964

Gabor, D. (1946). Theory of communication. *J. Inst. Elec. Eng.* 93, 429–457.

Gelfand, S. A., Ross, L., and Miller, S. (1988). Sentence reception in noise from one versus two sources: effect of aging and hearing loss. *J. Acoust. Soc. Am.* 83, 248–256. doi: 10.1121/1.396426

Gordon-Salant, S., and Fitzgibbons, P. J. (1999). Profile of auditory temporal processing in older listeners. *J. Speech Lang. Hear. Res.* 42, 300–311.

Gordon-Salant, S., Fitzgibbons, P. J., and Friedman, S. A. (2007). Recognition of time-compressed and natural speech with selective temporal enhancements by young and elderly listeners. *J. Speech Lang. Hear. Res.* 50, 1181–1193. doi: 10.1044/1092-4388(2007/082)

Greenberg, S., Carvey, H. M., Hitchcock, L., and Chang, S. (2003). Temporal properties of spontaneous speech – A syllable-centric perspective. *J. Phonetics* 31, 465–485. doi: 10.1016/j.wocn.2003.09.005

Greenwood, D. D. (1962). Approximate calculation of the dimensions of travelingwave envelopes in four species. *J. Acoust. Soc. Am.* 34, 1364–1369. doi: 10.1121/1.1918349

Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: effect of location and type of interferer. *J. Acoust. Soc. Am.* 115, 833–843. doi: 10.1121/1.1639908

He, N. J., Mills, J. H., Ahlstrom, J. B., and Dubno, J. R. (2008). Age-related differences in the temporal modulation transfer function with pure-tone carriers. *J. Acoust. Soc. Am.* 124, 3841–3849. doi: 10.1121/1.2998779

Herman, G. E., Warren, L. R., and Wagener, J. W. (1977). Auditory lateralization: age differences in sensitivity to dichotic time and amplitude cues. *J. Gerontol.* 32, 187–191. doi: 10.1093/geronj/32.2.187

Hillenbrand, J. M., Clark, M. J., and Nearey, T. M. (2001). Effects of consonant environment on vowel formant patterns. *J. Acoust. Soc. Am.* 109, 748–763. doi: 10.1121/1.1337959

Humes, L. E., Dubno, J. R., Gordon-Salant, S., Lister, J. J., Cacace, A. T., Cruickshanks, K. J., et al. (2012). Central presbycusis: a review and evaluation of the evidence. *J. Am. Acad. Audiol.* 23, 635–666. doi: 10.3766/jaaa.23.8.5

Humes, L. E., Kidd, G. R., and Lentz, J. J. (2013). Auditory and cognitive factors underlying individual differences in aided speech-understanding among older adults. *Front. Syst. Neurosci.* 7:55. doi: 10.3389/fnsys.2013.00055

Humes, L. E., Watson, B. U., Christensen, L. A., Cokely, C. G., Halling, D. C., and Lee, L. (1994). Factors associated with individual differences in clinical measures of speech recognition among the elderly. *J. Speech Hear. Res.* 37, 465–474.

Ihlefeld, A., and Shinn-Cunningham, B. (2008). Spatial release from energetic and informational masking in a divided speech identification task. *J. Acoust. Soc. Am.* 123, 4380–4392. doi: 10.1121/1.2904825

Kewley-Port, D., and Zheng, Y. (1999). Vowel formant discrimination: towards more ordinary listening conditions. *J. Acoust. Soc. Am.* 106, 2945–2958. doi: 10.1121/1.428134

Kidd, G. Jr., Mason, C. R., Deliwala, P. S., Woods, W. S., and Colburn, H. S. (1994). Reducing informational masking by sound segregation. *J. Acoust. Soc. Am.* 95, 3475–3480. doi: 10.1121/1.410023

Kowalski, N., Depireux, D. A., and Shamma, S. A. (1996). Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. *J. Neurophysiol.* 76, 3503–3523.

Lavandier, M., and Culling, J. F. (2008). Speech segregation in rooms: monaural, binaural, and interacting effects of reverberation on target and interferer. *J. Acoust. Soc. Am.* 123, 2237–2248. doi: 10.1121/1.2871943

Lee, J. H., and Humes, L. E. (2012). Effect of fundamental-frequency and sentence-onset differences on speech-identification performance of young and older adults in a competing-talker background. *J. Acoust. Soc. Am.* 132, 1700–1717. doi: 10.1121/1.4740482

Lee, T. Y., and Richards, V. M. (2011). Evaluation of similarity effects in informational masking. *J. Acoust. Soc. Am.* 129, EL280–EL285. doi: 10.1121/1.3590168

Levi, A., and Stark, H. (1983). Signal restoration from phase by projection onto a convex set. *J. Opt. Soc. Am.* 73, 810–822. doi: 10.1364/JOSA.73.000810

Lutfi, R. A. (1990). How much masking is informational masking? *J. Acoust. Soc. Am.* 88, 2607–2610. doi: 10.1121/1.399980

Lyzenga, J., and Carlyon, R. P. (1999). Center frequency modulation detection for harmonic complexes resembling vowel formants and its interference by off-frequency maskers. *J. Acoust. Soc. Am.* 105, 2792–2806. doi: 10.1121/1.426896

Lyzenga, J., and Carlyon, R. P. (2005). Detection, direction discrimination, and off-frequency interference of center-frequency modulations and glides for vowel formants. *J. Acoust. Soc. Am.* 117, 3042–3053. doi: 10.1121/1.1882943

McAdams, S. (1990). Interactions among cues contributing to concurrent sound segregation. *J. Acoust. Soc. Am.* 88, S146. doi: 10.1016/j.heares.2009.09.012

Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *J. Acoust. Soc. Am.* 123, 899–909. doi: 10.1121/1.2816572

Mesgarani, N., Thomas, S., and Hermansky, H. (2011). Toward optimizing stream fusion in multistream recognition of speech. *J. Acoust. Soc. Am.* 130, EL14–EL18. doi: 10.1121/1.3595744

Micheyl, C., and Oxenham, A. J. (2010). Pitch, harmonicity and concurrent sound segregation: psychoacoustical and neurophysiological findings. *Hear. Res.* 266, 36–51. doi: 10.1016/j.heares.2009.09.012

Moore, B. C., Sek, A., and Shailer, M. J. (1995). Modulation discrimination interference for narrow-band noise modulators. *J. Acoust. Soc. Am.* 97, 2493–2497. doi: 10.1121/1.411969

Moore, B. C. J., and Sek, A. (1996). Detection of frequency modulation at low modulation rates: evidence for a mechanism based on phase locking. *J. Acoust. Soc. Am.* 100, 2320–2331. doi: 10.1121/1.417941

Murphy, D. R., Craik, F. I., Li, K. Z., and Schneider, B. A. (2000). Comparing the effects of aging and background noise on short-term memory performance. *Psychol. Aging* 15, 323–334. doi: 10.1037/0882-7974.15.2.323

Ng, E. H., Rudner, M., Lunner, T., Pedersen, M. S., and Ronnberg, J. (2013). Effects of noise and working memory capacity on memory processing of speech for hearing-aid users. *Int. J. Audiol.* 52, 433–441. doi: 10.3109/14992027.2013.776181

Oh, E. L., and Lutfi, R. A. (1998). Nonmonotonicity of informational masking. *J. Acoust. Soc. Am.* 104, 3489–3499. doi: 10.1121/1.423932

Patterson, R., and Moore, B. (1986). "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selectivity in Hearing*, ed B. Moore (London: Academic Press), 123–178.

Pitton, J. W., Wang, K., and Juang, B.-H. (1996). Time-frequency analysis and auditory modeling for automatic recognition of speech. *Proc. IEEE* 84, 1199–1215. doi: 10.1109/5.535241

Plomp, R. (1983). "Perception of speech as a modulated signal," in *Proceedings of the 10th International Congress of Phonetic Science*, M. P. R. van der Broecke and A. Cohen (Dordrecht: Foris), 29–40.

Rajan, R., and Cainer, K. E. (2008). Ageing without hearing loss or cognitive impairment causes a decrease in speech intelligibility only in informational maskers. *Neuroscience.* 154, 784–795. doi: 10.1016/j.neuroscience.2008.03.067

Roman, N., and Wang, D. (2006). Pitch-based monaural segregation of reverberant speech. *J. Acoust. Soc. Am.* 120, 458–469. doi: 10.1121/1.2204590

Sabin, A. T., Clark, C. A., Eddins, D. A., and Wright, B. A. (2013). Different patterns of perceptual learning on spectral modulation detection between older hearing-impaired and younger normal-hearing adults. *J. Assoc. Res. Otolaryngol.* 14, 283–294. doi: 10.1007/s10162-012-0363-y

Shackleton, T. M., and Carlyon, R. P. (1994). The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination. *J. Acoust. Soc. Am.* 95, 3529–3540. doi: 10.1121/1.409970

Singh, N. C., and Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *J. Acoust. Soc. Am.* 114, 3394–3411. doi: 10.1121/1.1624067

Snell, K. B., and Frisina, D. R. (2000). Relationships among age-related differences in gap detection and word recognition. *J. Acoust. Soc. Am.* 107, 1615–1626. doi: 10.1121/1.428446

Snell, K. B., Mapes, F. M., Hickman, E. D., and Frisina, D. R. (2002). Word recognition in competing babble and the effects of age, temporal processing, and absolute sensitivity. *J. Acoust. Soc. Am.* 112, 720–727. doi: 10.1121/1.1487841

Sommers, M. S. (1997). Stimulus variability and spoken word recognition. II. The effects of age and hearing impairment. *J. Acoust. Soc. Am.* 101, 2278–2288. doi: 10.1121/1.418208

Sommers, M. S., and Gehr, S. E. (1998). Auditory suppression and frequency selectivity in older and younger adults. *J. Acoust. Soc. Am.* 103, 1067–1074. doi: 10.1121/1.421220

Takahashi, G. A., and Bacon, S. P. (1992). Modulation detection, modulation masking, and speech understanding in noise in the elderly. *J. Speech Hear. Res.* 35, 1413–1429.

Terhardt, E., Stoll, G., and Seewann, M. (1982). Algorithm for extraction of pitch and pitch salience from complex tonal signals. *J. Acoust. Soc. Am.* 71, 679–688. doi: 10.1121/1.387544

Watson, C. S. (1987). "Uncertainty, informational masking, and the capacity of immediate auditory memory," in *Auditory Processing of Complex Sounds*, ed W. A. Yost and C. S. Watson (Hillsdale, NJ: L. Erlbaum), 267–287.

Zekveld, A. A., Kramer, S. E., and Festen, J. M. (2011). Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response. *Ear Hear.* 32, 498–510. doi: 10.1097/AUD.0b013e31820512bb

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

Received: 13 February 2014; accepted: 22 May 2014; published online: 12 June 2014.
*Citation: Divenyi P (2014) Decreased ability in the segregation of dynamically changing vowel-analog streams: a factor in the age-related cocktail-party deficit? Front. Neurosci. **8**:144. doi: 10.3389/fnins.2014.00144*
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 24 June 2014
doi: 10.3389/fnins.2014.00170

Probing the time course of head-motion cues integration during auditory scene analysis

***Hirohito M. Kondo*[1,2]*, *Iwaki Toshima*[1], *Daniel Pressnitzer*[3,4] *and Makio Kashino*[1,5]**

[1] *NTT Communication Science Laboratories, NTT Corporation, Atsugi, Japan*
[2] *Department of Child Development, United Graduate School of Child Development, Osaka University, Kanazawa University, Hamamatsu University School of Medicine, Chiba University, and University of Fukui, Suita, Japan*
[3] *Laboratoire des Systèmes Perceptifs, CNRS UMR 8248, Paris, France*
[4] *Département d'études cognitives, École normale supérieure, Paris, France*
[5] *Department of Information Processing, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Japan*

Edited by:
Susann Deike, Leibniz Institute for Neurobiology, Germany

Reviewed by:
István Winkler, University of Szeged, Hungary
Sebastian Puschmann, Carl von Ossietzky University Oldenburg, Germany

****Correspondence:***
Hirohito M. Kondo, NTT Communication Science Laboratories, NTT Corporation, 3-1 Morinosato Wakamiya, Atsugi, Kanagawa 243-0198, Japan
e-mail: kondo.hirohito@lab.ntt.co.jp

The perceptual organization of auditory scenes is a hard but important problem to solve for human listeners. It is thus likely that cues from several modalities are pooled for auditory scene analysis, including sensory-motor cues related to the active exploration of the scene. We previously reported a strong effect of head motion on auditory streaming. Streaming refers to an experimental paradigm where listeners hear sequences of pure tones, and rate their perception of one or more subjective sources called streams. To disentangle the effects of head motion (changes in acoustic cues at the ear, subjective location cues, and motor cues), we used a robotic telepresence system, Telehead. We found that head motion induced perceptual reorganization even when the acoustic scene had not changed. Here we reanalyzed the same data to probe the time course of sensory-motor integration. We show that motor cues had a different time course compared to acoustic or subjective location cues: motor cues impacted perceptual organization earlier and for a shorter time than other cues, with successive positive and negative contributions to streaming. An additional experiment controlled for the effects of volitional anticipatory components, and found that arm or leg movements did not have any impact on scene analysis. These data provide a first investigation of the time course of the complex integration of sensory-motor cues in an auditory scene analysis task, and they suggest a loose temporal coupling between the different mechanisms involved.

Keywords: auditory streaming, bistable perception, build-up, cocktail party problem, crossmodal, hearing, head movement, virtual reality

INTRODUCTION

The structuring of a sensory scene determines what we perceive: rather than an indiscriminate mixture of acoustic events, a lively conversation between friends can be parsed into meaningful components. The sequential integration and segregation of frequency components for the formation of percepts, which is called auditory streaming, is essential for auditory scene analysis, as sound sources produce information over time. Traditionally, streaming has been studied with a highly simplified experimental paradigm (Miller and Heise, 1950; van Noorden, 1975; see Moore and Gockel, 2012 for a recent review). In such a paradigm, a sequence of two tones, A and B, is presented, with A and B set at different frequencies. The frequency difference between A and B biases the most likely perceptual organization: a small difference favors the perception of one stream, whereas a large separation favors the perception of two streams. Streaming is actually a bistable phenomenon for a range of A and B frequencies (see Schwartz et al., 2012 for a review), as a physically unchanging streaming sequence most often induces successive percepts of one or two streams, in a seemingly random fashion.

In the present study, we focus on the so-called build-up of streaming. This refers to the observation that streaming sequences tends initially to be heard as a single stream (van Noorden, 1975) before bistable alternations begin. The build-up has been widely used to probe streaming in behavior and physiology (e.g., Snyder and Alain, 2007 for a review). Note that, recently, the notion of build-up has been questioned by Deike et al. (2012). They pointed out an important experimental caveat: for build-up to be accurately estimated, the period of time between the onset of the sound and the first subjective report should be treated as missing data, which had not always been the case in previous investigations. When Deike et al. (2012) used a missing-data analysis, they found that build-up was not observed for all frequency separations. However, a build-up was still observed for moderate frequency separations (Deike et al., 2012). Other reports using the missing-data approach and moderate frequency separations also reported a build-up (for instance Pressnitzer and Hupé, 2006; Hupé and Pressnitzer, 2012). We adopted this methodology in the present study.

A last consideration of interest is that streaming may be "reset" by a sudden change in the stimulus. For instance, a

change in the ear of entry or in the spatial location of the stimulus (Anstis and Saida, 1985; Rogers and Bregman, 1998) tends to increase the proportion of one-stream reports, as is observed at the onset of a stimulus before build-up occurs. Other manipulations can have the same effect, such as introducing short silent gaps in the streaming sequence (Cusack et al., 2004; Denham et al., 2010) or even engaging and disengaging attention (Best et al., 2008; Thompson et al., 2011). The term resetting suggests that perceptual organization starts anew, but it should be noted that in most experiments the reset was only partial. Furthermore, the actual mechanisms of resetting are unknown. With these qualifications in mind, we will still use the term "resetting" in the following, for consistency with previous reports.

In a previous study, we used the build-up of streaming and its resetting as a tool to investigate the effects of head motion on auditory streaming (Kondo et al., 2012). The rationale was as follows. A change in acoustic cues at the ears, associated with a change in the subjective location of the sound, can induce resetting (Rogers and Bregman, 1998). A voluntary head motion also induces changes in the acoustic cues at the ear, but in theory no sizeable change in subjective spatial location of the sound in allocentric coordinates. What happens to resetting in this case? Perhaps surprisingly, we showed that voluntary head motion did produce some resetting, even though the acoustic scene had not changed (Kondo et al., 2012). Furthermore, we disentangled the various effects of head motion by using a telepresence robot, the "Telehead" system (Toshima et al., 2008). The structure of the various trials types used is summarized in **Figure 1** and presented in more details in the Material and Methods section of the present study. In some trials, the Telehead followed head motion, but in others it did not. This allowed to have trials with all possible combinations of three types of cues: (i) changes in acoustic cues at the ears (ΔA), (ii) changes in source location in allocentric coordinates (ΔS), and (iii) changes in non-auditory processes related to head motion (ΔH). A linear model was used to evaluate the effect of each cue. The results showed that all cues impacted perceptual organization, with no counter-balancing between, for instance ΔA and ΔH to avoid a resetting during natural head motion.

In the present study, we reanalyzed the data of Kondo et al. (2012) to focus on the precise time-course of perceptual organization after head motion. It would be possible to hypothesize, for instance, that merging the head position signal with auditory computations is sluggish, leading to the observed lack of exact compensation.

MATERIAL AND METHODS

LISTENERS

Ten strongly right-handed listeners were recruited for Experiment 1 (5 males and 5 females; mean age 25.3 years, range 19–30 years). Three different listeners participated in Experiment 2 (2 males and 1 female; mean age 31.3 years, range 24–38 years). All had normal hearing as clinically defined by their audiograms. None had any history of neurological or psychiatric illness or hearing-related disorders. All gave written informed consent, which was approved by the Ethics Committee of NTT Communication Science Laboratories.

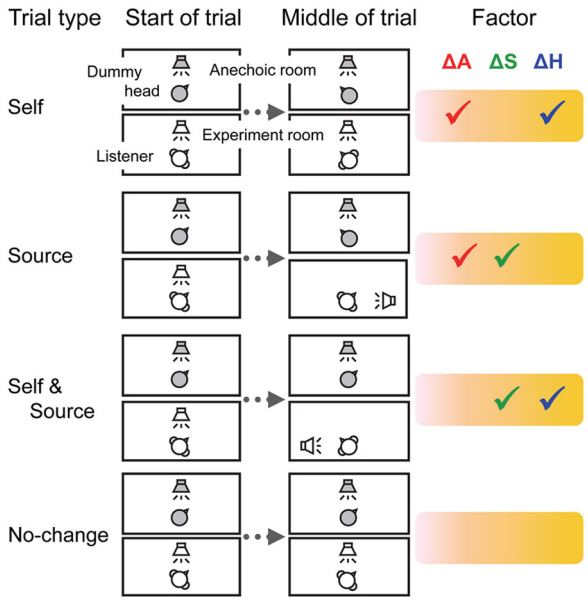


FIGURE 1 | The experimental setup and trial types. Auditory stimuli were presented to the Telehead system in an anechoic room. A loudspeaker was positioned in front of the robotic dummy head. Sounds were collected by microphones placed in the dummy head and transmitted to the listener via headphones. The head motion of the listener could be mimicked with minimal latency by the robotic head. Factors: ΔA, changes in acoustic cues at the ears; ΔS, changes in subjective sound localization; ΔH, changes in non-auditory factors related to head motion.

APPARATUS

Listeners were seated in the center of a double-walled soundproof room, wearing headphones (HDA 200, Sennheiser). Their head motion was tracked in real time and sent to the Telehead robot, which could mirror the 3D motion with minimal latency and distortion (Toshima et al., 2008). Auditory stimuli were delivered through a loudspeaker (MG10SD0908, Vifa) located 1 m in front of the Telehead dummy head, in an anechoic chamber. Sound was recorded by small microphones (ECM77B, Sony) placed 2 mm inside the entrance of the dummy head's outer ears and transmitted in real time to the headphones. Two light-emitting diodes (LEDs) were used as visual cues to direct head movements and positioned on the left and right sides of the listener at eye level and at a 2-m distance (visual angle re: midline = 60°). The room was darkened for the duration of the experiment.

The Telehead dummy head was made by molding a human head using impression material. The surface was covered with a 1-cm thick layer of soft polyurethane resin. The listeners' head positions were measured with a 3D head-tracker (FASTRAK, Polhemus) placed on the top of the headphones. The position data were obtained at a 120-Hz sampling rate and used to synchronize the yaw, pitch, and roll motions (maximum range 180, 80, and 60°, respectively) of the listener's head with those of the dummy head.

STIMULI AND TASK PROCEDURES

The auditory stimuli were composed of 50 repetitions of a triplet of narrow-band pink noises (roll-off = 3 dB/octave) arranged in

an ABA- pattern where A and B represent different noise bands and a hyphen represents a silent interval. The A and B bands were geometrically centered around 1 kHz with a 6-semitone frequency difference between them and a 4-semitone bandwidth. This yielded cut-off frequencies of (749–944) Hz for the A band and (1060–1335) Hz for the B band. The noise bands were generated in the frequency domain and equated in RMS amplitude. The duration of each noise was 62.5 ms, which included rising and falling cosine ramps of 10 ms. The onset asynchrony between successive bands was 100 ms. A background of pink noise was also included to mask any residual line noise of the Telehead system. The pink noise was generated in the frequency domain with cut-off frequencies of (0.1–5) kHz, with a level of −30 dB RMS relative to the A and B bands. The sound pressure level was measured by using ICE couplers with microphones and a measuring amplifier (Brüel and Kjær). The presentation level of the stimuli was set at 65 dB SPL.

Listeners were tested individually. We first explained the concept of auditory streaming by means of a visual illustration of the stimuli. They were instructed to report their percept by pressing one out of two buttons (one-stream when they heard a galloping rhythm ABA-ABA-, or two-stream when they heard A-A-…and –B—B—…each with an isosynchronous rhytm). Listener's responses were held between button presses. Before the first button press, responses were treated as missing data. Listeners were also instructed to move their head to track an LED and maintain it at the center of their gaze. Before the beginning of each trial, they oriented themselves toward the midline and their head positions were calibrated. Then, a blinking LED was randomly presented either on their left or right side and counterbalanced across the trials.

Experiment 1 consisted of four types of trials (**Figure 1**). In the Self trials, after 10 s of sound presentation, the LED was turned off and another LED was lit on the contralateral side. The listeners were instructed to track this change by moving their head as fast as possible so as to maintain their gaze on the light. The Telehead robot mimicked the head motion, so the Self trials simulated actual head motion. In the Source trials, the LED remained lit on the same side throughout the trial, so that there was no head motion required from the listener. However, the Telehead robot initiated a motion previously recorded from the same listener. This motion had the same acoustic cues at the ears as for the Self trials, but without their motor and volitional components. Such Source trials simulated the displacement of a sound source. In the Self and Source trials, listeners initiated a head motion to follow a change in the visual cue position, but the robot did not move. Such trials have all the motor and volitional components of the Self trials, but without any change in the acoustic cues at the ears. They resulted in an apparent motion of the source in allocentric coordinates, which appeared to follow exactly the orientation of the head (as when one listens to music over headphones). In the No-change trials, the visual cue position was maintained throughout the trial and neither the listener nor the robot moved. The No-change trials were used as a baseline.

Experiment 2 consisted of three types of trials. At the beginning of the Arm trials, an LED was lit on the right side. Listeners were asked to raise their left arm as quickly as possible if the LED was turned off 10 s after stimulus onset. The left arm was chosen as listeners used their right hand to report streaming. An LED on the left side was lit in the Leg trials, and then the listeners raised their left leg if the LED was turned off. The No-change trials were identical to those in Experiment 1 (the LED was maintained lit during the whole duration of the trial).

The trial types were randomly mixed within blocks, for each experiment. The order of the trial types was randomized. In addition to tracking the visual cue, the listeners were instructed to continuously report whether they heard one stream or two.

At least 12 practice trials were run before data collection began. The head movement of the listeners in the final practice trial was recorded to generate the Telehead motion in the Source trials. Experiments 1 and 2 consisted of 6 blocks of 24 trials and 6 blocks of 18 trials, respectively.

DATA ANALYSES

Thirty-six time-series data (temporal resolution, 1 ms) were collected for each trial type. We smoothed the probability of two-stream judgments with 10-ms, non-overlapping rectangular temporal windows (bins). For each bin, we computed a resetting index, R, for the Self, Source, and Self & Source trials. R was obtained by subtracting the baseline probability of two-stream judgments in the No-change trials to the actual probability of two-stream in the condition of interest. R was computed for each bin, trial type TT, and listener L, as:

$$R_{TT,L} = P_{TT,L}(2\text{ stream}) - P_{No-change,L}(2\text{ stream}) \quad (1)$$

We then built a linear model to estimate the contribution of ΔA, ΔS, and ΔH to resetting. R was modeled as:

$$R = K_A\Delta A + K_S\Delta S + K_H\Delta H \quad (2)$$

Three measures of R were available for each listener, one for each trial type. For each measure, the values of ΔA, ΔS, and ΔH were set at either 0 or 1 depending on whether the trial type included changes in the corresponding factor (see **Figure 1**, check marks indicate 1). The system of three equations and three unknowns was then solved for each listener.

We computed K_A, K_S, and K_H for each time bin from 10 to 20 s after stimulus onset, and performed a repeated-measures analysis of variance (ANOVA) on the values. Tukey honestly significant difference (HSD) tests were used for *post hoc* comparisons (α-level = 0.05).

RESULTS AND DISCUSSION

We found a build-up pattern for the first 10 s of sound presentation in Experiment 1. The initial report was always one stream and the probability of two streams increased gradually over time. The analysis of interest focused on the percepts reported for the second half of the stimuli, that is, between 10 s and 20 s relative to stimulus onset. The probability of two-stream at 10 s did not depend on trial type. The probabilities of two-stream reports were 61 ± 5, 54 ± 5, 58 ± 5, and 58 ± 5% (means ± *SE*) for the Self, Source, Self & Source, and No-change trials, respectively, $F_{(3,\ 27)} = 2.38$, $\eta^2 = 0.03$, $p = 0.09$. This indicates,

reassuringly, that listeners could not guess the trial types before the presentation of the visual cue.

As just described, at the 10-s point where stimulus manipulation occurred, listeners reported two-stream in roughly 60%, and one-stream in the remaining 40% of trials. According to the standard definition of resetting, a reset implies a switch from two-stream to one-stream. Therefore, according to this definition, resetting could only be measured for those trials where listeners reported two-stream at 10 s. We will term those trials two-stream trials and analyze them separately. However, it could also be that the manipulations had a different effect on perceptual organization, not accounted for by the standard view of resetting: for instance, a change could facilitate a perceptual switch, whatever the state of the listener (one- or two-stream). We thus also applied the same analyses to the one-stream trials, for which listeners reported one-stream at 10 s. In summary, all trials were classified as either one-stream trials or two-stream trials according to the perceptual state of the listener at 10 s. The probability of two-stream at 10 s was normalized to either 0 or 1, respectively, for each type of trials.

The two-stream trials (**Figure 2A**, top panel) were already analyzed in Kondo et al. (2012) in an a priori time window. Here, using a time-varying analysis and Tukey HSD tests (10-ms time bins, $p < 0.05$ as criterion), we found differences between conditions in a temporal window ranging from 11.7 s to 14.5 s after stimulus onset. For the one-stream trials (**Figure 2A**, bottom panel), the effect of stimulus manipulation was much less salient. No significant difference was found between any of trial types, at any time bin in the analysis. Thus, head motion and source location changes only affected those trials where perceptual organization was at two-stream at the time of the manipulation.

Because of the lack of effect for the one-stream trials, we computed the time-varying R index only for the two-stream trials. Results are shown in **Figure 2B**, which displays the time-series for the contributions of the ΔA, ΔS, and ΔH factors to R. The shaded area in the figure indicates the time interval encompassing the head motion produced by listeners in the *Self* trials: 0.80 s ± 0.07 s, with an onset of motion at 10.6 s. In Experiment 1, the duration of the sound motion was matched with that of head motion (see Material and Methods), so the shaded area also represents the time when stimulus changes were introduced. We compared the contributions for the three factors by a repeated-measures ANOVA. The contribution of ΔH was larger from 11.3 s to 11.7 s but was smaller from 13.2 s to 15.7 s than those of ΔA and ΔS. A further difference is that ΔH produced negative values, that is, a bias toward two-stream, for the later period. The peak amplitudes of the contributions did not differ for different factors: ΔA, 0.26 ± 0.03; ΔS, 0.21 ± 0.02, and ΔH, 0.22 ± 0.03, $F_{(2,\ 18)} = 1.71$, $\eta^2 = 0.16$, $p = 0.21$. However, the latency to the peak amplitude was earlier for ΔH (12.4 s ± 0.5 s) than for ΔA (13.1 s ± 0.3 s) and ΔS (13.4 s ± 0.4 s), $F_{(2,\ 18)} = 3.74$, $\eta^2 = 0.29$, $p < 0.05$. So the effects of ΔH on perceptual organization occurred earlier than those of ΔS and ΔA, and also lasted for a shorter time. The initial effect of ΔH was to contribute to resetting, but there was also a "negative" contribution to resetting for ΔH at later times. Such a negative contribution would be what is required for compensating the effects of ΔA and ΔS and prevent resetting when only the head moves and the scene does not change. However, this negative contribution was too slow and too small for canceling out the resetting effects of other factors, like e.g., ΔA in the case of natural head motion.

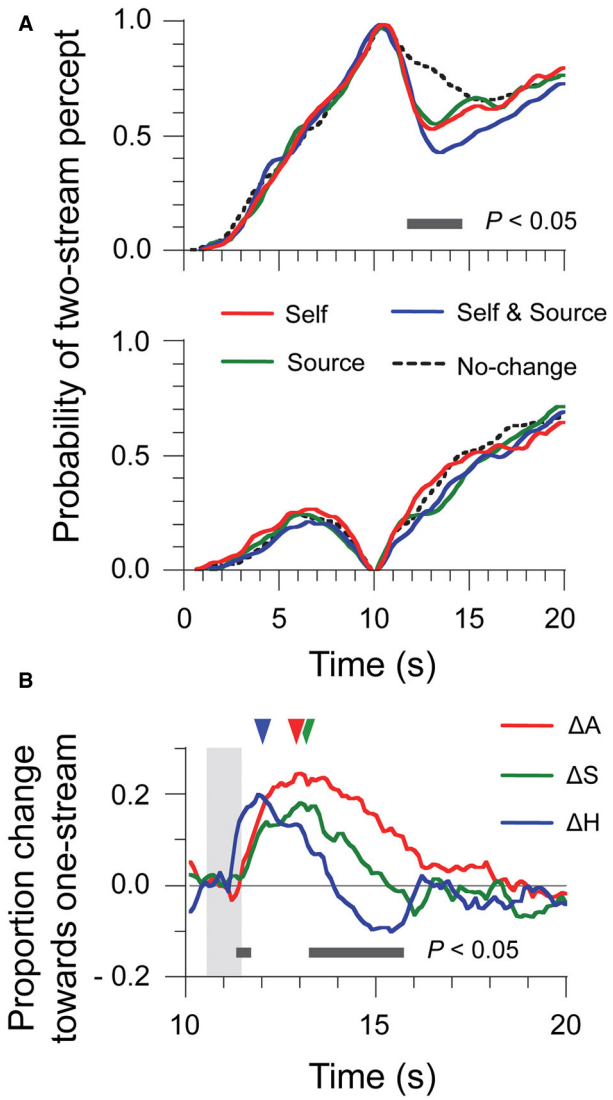


FIGURE 2 | Time-series data of percept probabilities and factor contributions for Experiment 1. (A) Normalized data were computed by selecting the trials where perception was either two-stream (top panel) or one-stream (bottom panel) at the 10-s point. **(B)** Contributions of the ΔA, ΔS, and ΔH factors to resetting were estimated for each time bin by means of a linear additive model considering all trial types. The shaded area represents the time window of sound motion and head motion. Triangles indicate the latency of the maximum amplitude for each factor.

These new observations on the time-course of ΔH suggest further possible interpretations for the lack of exact compensation between sensory cues and head motion. The ΔH factor reflected the overall effect of head motion, but it could possibly be further decomposed into a volitional component triggering the head movement, a component related to the

elaboration of motion commands, and somatosensory feedback information. Obviously, a volitional component would have to be present before any head motion could occur. In addition, it is known that volitional control affects spontaneous switching in auditory streaming (Pressnitzer and Hupé, 2006) as well as in visual bistable stimuli (Meng and Tong, 2004). Therefore, it could be that an early volitional signal could account for the early resetting effect of ΔH, which was only later followed by compensation mechanisms perhaps due to somatosensory feedback.

Experiment 2 aimed at controlling for the volitional component of ΔH. Listeners had to move their arm or leg, but not their head, during the streaming task (see Material and Methods). This task should have a comparable volitional component to the head-motion main task, without any relevant motor or sensory feedback cues for auditory scene analysis. The same analysis of streaming report was used as in the main experiment. Results are displayed in **Figure 3**. As before, build-up was observed and there was no difference in the probability of two streams at 10 s between the Arm, Leg, and No-change trials: 65 ± 2, 67 ± 2, and 67 ± 3%, $F_{(2,\ 4)} = 1.60$, $\eta^2 = 0.10$, $p = 0.31$. All the trials were classified under one- and two-stream trials, and the probability of two streams at 10 s was normalized into either 0 or 1. We did not find any difference in the probability of two streams between trial types at any time bin, for either two-stream or one-stream trials. This shows that a volitional component to body motion, as estimated with arm and leg movements, was not a major factor of the pattern of data. It also suggests that the early resetting effect observed for head motion in Experiment 1 was likely not due to volitional anticipation.

To summarize the new experimental findings, we found that acoustic cues (ΔA) and subjective location cues (ΔS) had a sustained and positive effect on resetting. In contrast, the head-motion cues (ΔH) had an early resetting effect followed by a later compensating effect. The early effect did not seem to be related to a volitional anticipation of the motion, as other types of body motion had no effect on auditory streaming.

It may be useful to consider those findings in the light of, on the one hand, the neural bases of sensory-motor integration during head motion, and, on the other hand, the neural bases of auditory streaming. The most obvious impact of head motion on auditory processes is for sound source localization. During head-motion, craniocentric binaural cues such as inter-aural time and inter-aural level differences must be transformed into allocentric coordinates: a stationary source will produce dynamic changes in binaural cues during head motion, changes that must be accounted for by comparing them to those expected because of head motion. This conversion from craniocentric to allocentric coordinates could use efferent copies of the head-motion command, afferent information from neck muscles, or afferent information from the vestibular system (Lewald and Ehrenstein, 1998; Lewald et al., 1999). The precise neural stage at which such signals contact the auditory pathways is not fully known. Human EEG data suggest that the allocentric map may only be fully completed after the primary auditory cortex (Altmann et al., 2009). However, there is also ample evidence from single-unit recordings that motor signals modulate early auditory spatial processing in

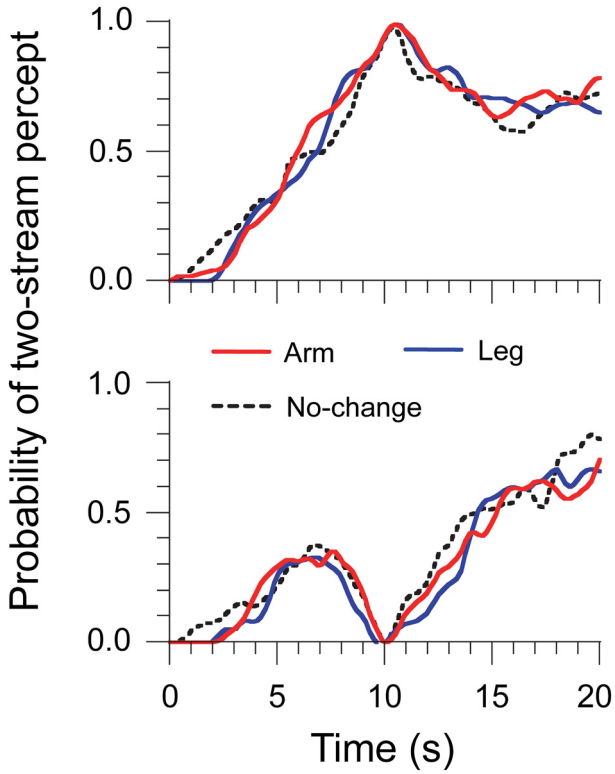


FIGURE 3 | Time-series data of percept probabilities for Experiment 2. Normalized data were computed by selecting the trials where perception was either two streams (top panel) or one stream (bottom panel) at the 10-s point.

the inferior (Groh et al., 2001) and superior (Jay and Sparks, 1984; Populin et al., 2004) colliculi, although those studies focused on eye position rather than head position. For auditory streaming, neural correlates are also a matter of debate. Correlates have been claimed at several stages in the auditory pathways: in the auditory cortex (Micheyl et al., 2007), in supra-modal areas such as the intraparietal sulcus (Cusack, 2005), but also subcortically in the auditory thalamus (Kondo and Kashino, 2009), the inferior colliculus (Schadwinkel and Gutschalk, 2011), and even before binaural convergence in the cochlear nucleus (Pressnitzer et al., 2008).

Therefore, a hypothesis to explain the surprising presence of partial perceptual resetting after head motion, even when the scene has not changed, could be that at least parts of network involved in auditory scene analysis are not fully modulated by head position signals during self-motion. This hypothesis would be consistent with findings related to sound source localization, independent of auditory scene analysis (Goossens and Van Opstal, 1999; Vliegen et al., 2004; Altmann et al., 2009). A parallel may exist with vision: eye movements are undoubtedly useful for apprehending a visual scene, but around the time of a saccade the compensation for self-induced motion is far from perfect (Ross et al., 2001). A compression of auditory space has also been reported just before the initiation of rapid head movements (Leung et al., 2008) and during passive

body movements (Teramoto et al., 2012). In other words, in this hypothesis, the resetting effect of head motion is not beneficial to auditory scene analysis, but it derives from other constraints on the neural architecture of the system (for instance the difficulty to have precise temporal alignment of all sources of information in two broadly distributed networks). Such an imperfect compensation may have been tolerated by the system as its computational efficiency outweighed any functional disadvantage.

There is another, more speculative interpretation of the observed time-course of head-motion signals on scene analysis. The early component of resetting due to head motion could be related specifically to head-motion volitional signals, anticipating motion and signaling the need to collect novel information (see for instance Kondo et al., 2012, for situations where the head motion disambiguate front/back location cues). In this perspective, at least a partial reconsideration of the current perceptual organization may be useful to integrate as rapidly as possible the new information revealed by head motion.

In any case, our data suggest a temporally-sluggish linkage between scene analysis and sensory-motor integration. Such a loose coupling may reduce the computational demands of combining the two complex functions, without any obvious functional disadvantage (or even a small benefit) in natural auditory scene analysis.

AUTHOR CONTRIBUTIONS

Hirohito M. Kondo, Iwaki Toshima, Daniel Pressnitzer, and Makio Kashino designed the research; Hirohito M. Kondo, Iwaki Toshima, and Daniel Pressnitzer performed the research; Hirohito M. Kondo, Iwaki Toshima, and Daniel Pressnitzer analyzed the data; and Hirohito M. Kondo and Daniel Pressnitzer wrote the paper.

REFERENCES

Altmann, C. F., Wilczek, E., and Kaiser, J. (2009). Processing of auditory location changes after horizontal head rotation. *J. Neurosci.* 29, 13074–13078. doi: 10.1523/JNEUROSCI.1708-09.2009

Anstis, S., and Saida, S. (1985). Adaptation to auditory streaming of frequency-modulated tones. *J. Exp. Psychol. Hum. Percept. Perform.* 11, 257–271. doi: 10.1037/0096-1523.11.3.257

Best, V., Ozmeral, E. J., Kopco, N., and Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proc. Natl. Acad. Sci. U.S.A.* 105, 13174–13178. doi: 10.1073/pnas.0803718105

Cusack, R. (2005). The intraparietal sulcus and perceptual organization. *J. Cogn. Neurosci.* 17, 641–651. doi: 10.1162/0898929053467541

Cusack, R., Deeks, J., Aikman, G., and Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 643–656. doi: 10.1037/0096-1523.30.4.643

Deike, S., Heil, P., Böckmann-Barthel, M., and Brechmann, A. (2012). The build-up of auditory stream segregation: a different perspective. *Front. Psychol.* 3:461. doi: 10.3389/fpsyg.2012.00461

Denham, S. L., Gyimesi, K., Stefanics, G., and Winkler, I. (2010). "Stability of perceptual organisation in auditory streaming," in *The Neurophysiological Bases of Auditory Perception*, eds E. A. Lopez-Poveda, A. R. Palmer, and R. Meddis (New York, NY: Springer), 477–488.

Goossens, H. H. L. M., and Van Opstal, A. J. (1999). Influence of head position on the spatial representation of acoustic targets. *J. Neurophysiol.* 81, 2720–2736.

Groh, J. M., Trause, A. S., Underhill, A. M., Clark, K. R., and Inati, S. (2001). Eye position influences auditory responses in primate inferior colliculus. *Neuron* 29, 509–518. doi: 10.1016/S0896-6273(01)00222-7

Hupé, J. M., and Pressnitzer, D. (2012). The initial phase of auditory and visual scene analysis. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 942–953. doi: 10.1098/rstb.2011.0368

Jay, M. F., and Sparks, D. L. (1984). Auditory receptive fields in primate superior colliculus shift with changes in eye position. *Nature* 309, 345–347. doi: 10.1038/309345a0

Kondo, H. M., and Kashino, M. (2009). Involvement of the thalamo-cortical loop in the spontaneous switching of percepts in auditory streaming. *J. Neurosci.* 29, 12695–12701. doi: 10.1523/JNEUROSCI.1549-09.2009

Kondo, H. M., Pressnitzer, D., Toshima, I., and Kashino, M. (2012). Effects of self-motion on auditory scene analysis. *Proc. Natl. Acad. Sci. U.S.A.* 109, 6775–6780. doi: 10.1073/pnas.1112852109

Leung, J., Alais, D., and Carlile, S. (2008). Compression of auditory space during rapid head turns. *Proc. Natl. Acad. Sci. U.S.A.* 105:6492–6497. doi: 10.1073/pnas.0710837105

Lewald, J., and Ehrenstein, W. H. (1998). Influence of head-to-trunk position on sound lateralization. *Exp. Brain Res.* 121, 230–238. doi: 10.1007/s002210050456

Lewald, J., Karnath, H. O., and Ehrenstein, W. H. (1999). Neck-proprioceptive influence on auditory lateralization. *Exp. Brain Res.* 125, 389–396. doi: 10.1007/s002210050695

Meng, M., and Tong, M. (2004). Can attention selectively bias bistable perception? Differences between binocular rivalry and ambiguous figures. *J. Vis.* 4, 539–551. doi: 10.1167/4.7.2

Micheyl, C., Carlyon, R. P., Gutschalk, A., Melcher, J. R., Oxenham, A. J., Rauschecker, J. P., et al. (2007). The role of auditory cortex in the formation of auditory streams. *Hear. Res.* 229, 116–131. doi: 10.1016/j.heares.2007.01.007

Miller, G. A., and Heise, G. A. (1950). The trill threshold. *J. Acoust. Soc. Am.* 22, 637–638. doi: 10.1121/1.1906663

Moore, B. C. J., and Gockel, H. E. (2012). Properties of auditory stream formation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 919–931. doi: 10.1098/rstb.2011.0355

Populin, L. C., Tollin, D. J., and Yin, T. C. (2004). Effect of eye position on saccades and neuronal responses to acoustic stimuli in the superior colliculus of the behaving cat. *J. Neurophysiol.* 92, 2151–2167. doi: 10.1152/jn.00453.2004

Pressnitzer, D., and Hupé, J. M. (2006). Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Curr. Biol.* 16, 1351–1357. doi: 10.1016/j.cub.2006.05.054

Pressnitzer, D., Sayles, M., Micheyl, C., and Winter, I. M. (2008). Perceptual organization of sound begins in the auditory periphery. *Curr. Biol.* 18, 1124–1128. doi: 10.1016/j.cub.2008.06.053

Rogers, W. L., and Bregman, A. S. (1998). Cumulation of the tendency to segregate auditory streams: resetting by changes in location and loudness. *Percept. Psychophys.* 60, 1216–1227. doi: 10.3758/BF03206171

Ross, J., Morrone, M. C., Goldberg, M. E., and Burr, D. C. (2001). Changes in visual perception at the time of saccades. *Trends Neurosci.* 24, 113–121. doi: 10.1016/S0166-2236(00)01685-4

Schadwinkel, S., and Gutschalk, A. (2011). Transient bold activity locked to perceptual reversals of auditory streaming in human auditory cortex and inferior colliculus. *J. Neurophysiol.* 105, 1977–1983. doi: 10.1152/jn.00461.2010

Schwartz, J. L., Grimault, N., Hupé, J. M., Moore, B. C. J., and Pressnitzer, D. (2012). Multistability in perception: binding sensory modalities, an overview. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 896–905. doi: 10.1098/rstb.2011.0254

Snyder, J. S., and Alain, C. (2007). Toward a neurophysiological theory of auditory stream segregation. *Psychol. Bull.* 133, 780–799. doi: 10.1037/0033-2909.133.5.780

Teramoto, W., Sakamoto, S., Furune, F., Gyoba, J., and Suzuki, Y. (2012). Compression of auditory space during forward self-motion. *PLoS ONE* 7:e39402. doi: 10.1371/journal.pone.0039402

Thompson, S. K., Carlyon, R. P., and Cusack, R. (2011). An objective measurement of the build-up of auditory streaming and of its modulation by attention. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 1253–1262. doi: 10.1037/a0021925

Toshima, I., Aoki, S., and Hirahara, T. (2008). Sound localization using an auditory telepresence robot: telehead II. *Presence: Teleoper. Virtual Environ.* 17, 392–404. doi: 10.1162/pres.17.4.392

van Noorden, L. P. A. S. (1975). *Temporal Coherence in the Perception of Tone Sequences.* Ph.D. thesis, Eindhoven University of Technology, Eindhoven.

Vliegen, J., Van Grootel, T. J., and Van Opstal, A. J. (2004). Dynamic sound localization during rapid eye-head gaze shifts. *J. Neurosci.* 24, 9291–9302. doi: 10.1523/JNEUROSCI.2671-04.2004

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

Received: 31 January 2014; accepted: 04 June 2014; published online: 24 June 2014.
*Citation: Kondo HM, Toshima I, Pressnitzer D and Kashino M (2014) Probing the time course of head-motion cues integration during auditory scene analysis. Front. Neurosci. **8**:170. doi: 10.3389/fnins.2014.00170*
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 21 July 2014
doi: 10.3389/fpsyg.2014.00775

Time course of auditory streaming: do CI users differ from normal-hearing listeners?

Martin Böckmann-Barthel[1]*, Susann Deike[2], André Brechmann[2], Michael Ziese[1] and Jesko L. Verhey[1]

[1] *Department of Experimental Audiology, Faculty of Medicine, Otto von Guericke University Magdeburg, Magdeburg, Germany*
[2] *Special Lab Non-invasive Brain Imaging, Leibniz Institute for Neurobiology, Magdeburg, Germany*

Edited by:
Susan Denham, Plymouth University, UK

Reviewed by:
Alexandra Bendixen, Carl von Ossietzky University of Oldenburg, Germany
Maria Chait, University College London, UK

****Correspondence:***
Martin Böckmann-Barthel, Department of Experimental Audiology, Faculty of Medicine, Otto von Guericke University Magdeburg, Leipziger Street 44, D-39120 Magdeburg, Germany
e-mail: martin.boeckmann@med.ovgu.de

In a complex acoustical environment, the auditory system decides which stimulus components originate from the same source by forming auditory streams, where temporally non-overlapping stimulus portions are considered to originate from one source if their stimulus characteristics are similar. The mechanisms underlying streaming are commonly studied by alternating sequences of A and B signals which are often tones with different frequencies. For similar frequencies, they are grouped into one stream. Otherwise, they are considered to belong to different streams. The present study investigates streaming in cochlear implant (CI) users, where hearing is restored by electrical stimulation of the auditory nerve. CI users listened to 30-s long sequences of alternating A and B harmonic complexes at four different fundamental frequency separations, ranging from 2 to 14 semitones. They had to indicate as promptly as possible after sequence onset, if they perceived one stream or two streams and, in addition, any changes of the percept throughout the rest of the sequence. The conventional view is that the initial percept is always that of a single stream which may after some time change to a percept of two streams. This general build-up hypothesis has recently been challenged on the basis of a new analysis of data of normal-hearing listeners. Using the same experimental paradigm and analysis, the present study found that the results of CI users agree with those of the normal-hearing listeners: (i) the probability of the first decision to be a one-stream percept decreased and that of a two-stream percept increased as Δf increased, and (ii) a build-up was only found for 6 semitones. Only the time elapsed before the listeners made their first decision of the percept was prolonged as compared to normal-hearing listeners. The similarity in the data of the CI user and the normal-hearing listeners indicates that the quality of stream formation is similar in these groups of listeners.

Keywords: cochlear implant, auditory streaming, build-up, perception, psychoacoustics

INTRODUCTION

Making sense of the complex auditory environment is of eminent importance in life. For hearing-impaired listeners and for those with their hearing restored by a cochlear implant (CI), a task such as the focusing on a conversational partner in a noisy environment often poses serious problems. Since the pioneering work of Miller and Heise (1950), Bregman and Campbell (1971), and van Noorden (1975), laboratory experiments have used the auditory streaming paradigm to understand the processes that help to disentangle several sound sources. In this paradigm, tones from two sets A and B differing in a specific sound feature are presented in rapid alternation, and listeners are asked if they hear one stream or two co-existing streams. Since fundamental frequency is the sound feature most often used to study the formation of streams, in the following, stream formation is discussed in relation to this feature only. At sufficiently high presentation rates, a two-stream percept is basically possible if the frequency separation Δf of the two sets exceeds a minimum value, the so-called fission boundary (cf. van Noorden, 1975). Whereas for large frequency separations, the segregated percept predominates, at intermediate frequency separations the perceptual organization is ambiguous, that is, both integration into one stream and segregation into two streams are possible (see Bregman, 1990, for a review).

The type of percept may change over the duration of a sequence. It has been generally assumed that "ABAB" sequences of alternating A and B sounds are initially heard as one stream, and that only after some time they are perceived as two separate streams which the listener can follow separately (see, e.g., Bregman, 1978; Anstis and Saida, 1985; Carlyon et al., 2001; Micheyl et al., 2005; Pressnitzer et al., 2008). Recently, Deike et al. (2012) used harmonic tone complexes in an ABAB sequence to study segregation in normal-hearing listeners. The average frequency separation between A and B tones ranged from 2 to 14 semitones. The rate was set to 6 tones per second. Listeners were asked to indicate as soon as possible their current percept and any change of it over the duration of the presented sequence. The probability of the one-stream and the two-stream percepts was tracked independently, thus providing the current probability of segregation unbiased toward any type of initial percept. They observed an increase in the two-stream probability only for the most ambiguous stimulus condition with a frequency separation

of 6 semitones between the A and B harmonic complexes. At larger frequency separations, the first percept was that of two-stream, in contrast to the general assumption of an initial one-stream percept. The probability of a two-stream percept was above 80% at the beginning of the sequence and hardly changed over the course of the sequence. This finding is in agreement with a recent study by Denham et al. (2013), who also found an initial two-stream percept at large frequency separations and high presentation rates. Based on their results in normal-hearing listeners, Deike et al. (2012) argued that a build-up of segregation is not generic and requires ambiguity of the sound sequences to occur.

Only few studies investigated stream segregation in CI users (Chatterjee et al., 2006; Hong and Turner, 2006; Cooper and Roberts, 2007, 2009; Marozeau et al., 2013). CIs can restore, to a certain extent, hearing in patients with a severe to complete cochlear hearing loss by stimulating the auditory nerve electrically. The signal is picked up by an external microphone and is then mapped to a linear array of 12–22 independent electrodes which was inserted into the cochlea. Each of these electrodes covers a certain frequency range.

All previous studies consistently found that effects related to stream segregation are observed in some of the CI users only. These studies either asked their listeners explicitly what they perceive at the end of the sequence (e.g., Chatterjee et al., 2006) or investigated stream segregation with a task where it is advantageous to perceive one stream (Hong and Turner, 2006; Cooper and Roberts, 2009) or two streams (Marozeau et al., 2013). Some of these studies used direct electrical stimulation of the electrodes, thus investigating if a stimulation of separate parts of the auditory nerve may generate stream segregation (Chatterjee et al., 2006; Cooper and Roberts, 2007, 2009). The others studies presented the stimuli through a loudspeaker and the microphone of the CI device (Hong and Turner, 2006; Marozeau et al., 2013).

A limitation of most studies is that they used short sequences with durations of several seconds where it is not clear if the percept is allowed to fully evolve. Only in one experiment, the sequence was 30 s long without changing the stimulus parameters (Cooper and Roberts, 2007). Unfortunately, in their study, listeners were instructed to press a single response button to indicate when the percept changed without defining the type of percept. Thus, their results do not allow to investigate the initial percept and how this was affected by the frequency separation.

Cooper and Roberts (2007, 2009) argued that their stream segregation results as well as those of Hong and Turner (2006) could be accounted for by assuming that listeners counted the number of different tones requiring a simple discrimination of A and B tones. Due to physiological and technical constraints of the CI, frequency discrimination of these patients is in order of several semitones (see for example, Pretorius and Hanekom, 2008). Thus, if stream segregation (with tone stimuli) is solely based on frequency discrimination, their ability to segregate streams should be worse than in normal-hearing listeners. One aim of this study is to investigate if this restriction results in a different stream formation in CI users compared to normal-hearing listeners.

Marozeau et al. (2013) also used long ABAB sequences of harmonic tone complexes but in contrast to Cooper and Roberts (2007) the frequency separation Δf of the A and B sounds gradually increased or decreased. Listeners were instructed to continuously rate the difficulty to follow the melody of the A tone set. When starting at a large average value of one octave, CI users could follow the melody rather easily, and increased the difficulty rating slowly with decreasing Δf. When increasing Δf from completely overlapping A and B tone sets, however, they found it much harder to hear out the melody, even at a Δf of one octave. Because each of the tone sets itself covered a range of 7 semitones, the results were not easily explained by simple discrimination of A and B tones but suggested an explicit segregation. Since the parameters continually changed in the experiment, the development and stability of a segregated percept could not be obtained from these data.

The main aim of the present study is to investigate the time course of the formation of streams in CI users and if this formation is similar in quality to that observed in normal-hearing listeners. The experiment of Deike et al. (2012) was replicated in experienced CI users with a reduced selection of the same stimuli and the identical task. The time course of perception is analyzed in the conventional way under the assumption of a default one-stream percept in comparison to the new analysis method used in Deike et al. (2012) which does not make any assumption of the percept before the first perceptual decision has been made. The proportions of time over which the sounds were perceived as either one-stream or two-streams are investigated as well as the time needed for the first perceptual decision.

MATERIALS AND METHODS

LISTENERS

Eight CI users aged 60–83 years (median age: 69 years), four female and four male, participated in the study. CI experience ranged from 8 months to 16 years (median: 2 years, 1 month). Demographic data of the participants are specified in **Table 1**. All participants used only one implant in the experiments. Bilaterally implanted listeners (CI01, CI06, CI08) usually listened through the earlier implanted side. The exception is listener CI08, who used the more recently implanted device due to significantly better performance with this implant than with the contralateral one. One participant with residual hearing (CI07) was equipped with an attenuating ear plug. All participants used their everyday device settings. The study was approved by the Ethics Committee of the Otto von Guericke University of Magdeburg to conform to the declaration of Helsinki. Each participant provided written informed consent to the study and was compensated for his expenses. Two additional CI users (not included in **Table 1**) had to be excluded from the experiment due to a consequent misunderstanding of the task or because recording of the responses failed due to a technical problem of the experimental setup.

APPARATUS, STIMULI, AND PROCEDURE

Whereas in the previous study on streaming in normal-hearing listeners (Deike et al., 2012) headphones were used, here sounds were presented to the CI users in a free-field condition in a sound-attenuated room through a single, frontally located active monitor loudspeaker (Reveal 6D, Tannoy Ltd., Coatbridge, UK).

Table 1 | Demographic data of the CI users participating in the experiment.

ID	Sex	Age (years)	CI experience (years; months)	Side	Contralateral hearing	Implant type	Sound processor	Processing strategy	Number of active channels	Lower cut-off frequency (Hz)
CI01	m	65	02;05	L	CI	Sonata	Opus2	FSP	12	100
CI02	f	83	01;08	R	–	Sonata	Opus2	FS4	12	100
CI03	m	65	16;01	R	–	CI40+	Opus2	FSP	11	100
CI04	m	73	02;01	L	–	Concerto	Opus2	FS4	12	100
CI05	f	69	03;02	L	–	CI512	CP810	ACE	19	188
CI06	f	73	01;08	R	CI	Sonata	Opus2	FS4	12	100
CI07	f	69	00;08	R	Residual	Concerto	Opus2	FS4	11	100
CI08	m	60	15;08	R	CI	CI40+	Opus2	FSP	9	100

All listeners were studied using only one CI device. The side is specified in the respective column labeled "side." One listener (CI07) was equipped with an earplug on the contralateral side due to residual hearing.

The level was individually adjusted to a comfortable level. Stimulus presentation and response collection was administered using Presentation (Neurobehavioral Systems Inc., San Francisco, CA, USA).

Otherwise, stimulus material and presentation were identical to that of Deike et al. (2012). Tones from a high-frequency set A and a low-frequency set B alternated in an ABAB order at a rate of 6 tones per second. The harmonic tone complexes consisted of the first five partials, including the fundamental frequency, at equal amplitudes. The duration of each tone complex was 25 ms including 3.8-ms cosine-squared onset and offset ramps. To construct ABAB sequences with different frequency separations (Δf) of A and B tone sets, the fundamental frequencies of both the A and B tones were arranged symmetrically on a semitone scale around 392 Hz. In contrast to Deike et al. (2012), only four conditions instead of seven with average values Δf of 2, 6, 10, and 14 semitones were used. The center values of the fundamental frequencies for the different conditions are given in **Table 2**. In addition, within each condition, individual exemplars of both A and B tones varied in the fundamental frequency in five discrete steps (± 2, ± 1, and 0 semitones). Forty sequences, each with a certain average Δf and duration of 30 s, were presented in a random order. The tone complexes were generated by MATLAB (The Mathworks Inc., Natick, MA, USA) prior to the experiment. Since the stimulus settings were identical to Deike et al. (2012), their data of normal-hearing listeners were used as a control.

The task of the listeners was to indicate their current percept continuously on a computer mouse. The listeners were asked to press the left button (marked as "1") as long as a single, see-saw stream of high and low tones (or of more or less the same pitch, termed as "one-stream") was perceived, and the right button (marked as "2") as long as a separation into two separate, parallel streams of high tones and low tones (termed as "two-stream") was perceived. In addition, sketches of an oscillating curve and two parallel curves were shown as graphical analogies of the two percept types prior to the experiment. The listeners were asked to enter a decision as soon as they had one of the two percepts, and to switch to the other button whenever the percept changed to the other type. Prior to the experiment, two sequences with the narrowest separation ($\Delta f = 2$ semitones) and two sequences with the widest separation ($\Delta f = 14$ semitones) were presented in an alternating manner to familiarize the listeners with the task.

Table 2 | Median fundamental frequencies fo of the A and B tone complexes in the different Δf conditions.

	fo/Hz	Condition: Δf/semitones
A tones	587	14
	523	10
	466	6
	415	2
B tones	370	2
	330	6
	294	10
	262	14

In addition, these values are randomly varied in semitone steps over a range of ±2 semitones within the A and B sounds (for details, see text).

DATA ANALYSIS

From the recorded instants of the responses, the proportions of time that the sound sequence was perceived as either one-stream or two-stream were calculated for each CI user and each condition. The resulting proportions of time were subjected to repeated-measures analyses of variance (SPSS Version 21, IBM, Armonk, NY, USA), testing for the within-subject factor Δf condition. Where necessary (due to violation of the sphericity assumption), p values were corrected according to Greenhouse–Geisser. Since it was expected that the proportion of time the sequences were perceived

as two streams increases with increasing Δf, planned contrasts were studied by calculating the polynomic trends.

In addition, for each CI user the median of the time to the first response was calculated across the 10 sequence presentations at each Δf condition. The median was used instead of mean values as the data were skewed toward smaller values, between as well as within listeners. Because the data were thus not normally distributed, a non-parametric Friedman test was conducted. *Post hoc* Wilcoxon tests are reported with Bonferroni correction of multiple testing.

The time course of the probability of each of the two percepts was calculated as described in Deike et al. (2012). Each sequence was divided into 1-s bins. For the probability of a one-stream percept, a value of one was assigned to each bin where the last answer had been a left-mouse button ("1") press, and a value of zero if it had been a right-mouse button ("2") press. This probability was averaged across the 10 presentations of a certain Δf value. In the same way, the probability of a two-stream percept was calculated by assigning a value of 1 to each bin where the last answer had been a right-mouse button ("2") press, and a value of 0 if it had been a left-mouse button ("1") press. The average across the 10 presentations for a Δf value yields to the time course of the two-stream percept for this frequency separation. In each average time course, the first bins were taken only as long as at least one response of at least two participants had occurred.

RESULTS

PROPORTIONS OF PERCEPT TYPES

To assess if listeners perceived alternating sequences as one (single) stream or two (segregated) streams at the four different Δf values, the individual proportions of both percepts were calculated over the whole duration of the sequence. In case the listeners showed stream segregation, the proportion of a two-stream percept should increase with Δf, whereas the proportion of a one-stream percept should decrease. The results are displayed in **Figure 1**. In five listeners (CI02, CI03, CI05, CI06, and CI08), an increase of the proportion of the two-stream percept was observed, in agreement with the streaming hypothesis. Three listeners (CI01, CI04, and CI07) had similar proportions of one-stream and two-stream percepts at all Δf, indicating that they did not experience an increase of the two-stream percept in any of the tested conditions.

The grand means of the data are shown in **Figure 2**. The proportion of the two-stream percept increases with Δf, whereas the proportion of the one-stream percept is reduced. The most ambiguous situation with equivocal proportions of both percepts is found at a Δf of 6 semitones. Note that the summed proportions do not add up to 1. This "missing" proportion represents the time from stimulus onset until the first perceptual decision was reported. The difference to 100% corresponds to the mean reaction time for the first response relative to the total duration.

For the analysis of the mean proportion of the two-stream percept, a repeated-measures analysis of variance was conducted on the factor Δf condition. For the proportion of the two-stream percept, Mauchly's test indicated the violation of the sphericity assumption. Using Greenhouse–Geisser corrected values ($\varepsilon = 0.61$), this proportion depended with high significance on Δf [$F(3,21) = 7.39$, $p < 0.01$]. Planned contrasts yielded a significantly increasing linear trend of the proportion of the two-stream percept with increasing Δf [$F(1,7) = 12.30$, $p < 0.01$]. The mean proportions of the one-stream percept were analyzed in the same way, using Greenhouse–Geisser correction ($\varepsilon = 0.58$). Here, Δf also had a significant effect [$F(3,21) = 9.21$, $p < 0.01$]. Planned contrasts yielded a significantly decreasing linear trend of the proportion of the one-stream percept with increasing Δf [$F(1,7) = 9.21$, $p < 0.01$].

FIRST RESPONSE

Each of the listeners provided at least one response within the duration of each single sequence. **Figures 3A,B** displays the time elapsed before the first response as individual data and box plot. Shown are individual values of the eight listeners evaluated before. Only three of the listeners consequently provided the first response within 3 s. Two listeners (CI02 and CI07) reacted very slowly with first button presses after more than 10 s. The effect of the factor Δf on the time of the first response was investigated using a non-parametric, Friedman test for related samples. This yielded no significant effect of Δf [$\chi^2(3) = 4.65$, $p > 0.1$].

If the listeners showed a build-up type of response, the first percept in each sequence should have been that of one-stream, independent of Δf. In normal-hearing listeners, however, the first decision was found to be two-stream at large Δf values (Deike et al., 2012). For the CI users, the probability that the first decision was a two-stream percept is shown in **Figure 3C**. As there are only two alternatives, these data also reflect the probability of the first decision being one stream, and both probabilities add up to 100%. The probability of a two-stream percept is low at small Δf but rises to a mean close to 100% at large Δf. As in the normal-hearing listeners, this suggests that there is no default one-stream percept in CI users. A repeated-measures analysis of variance on the factor Δf (Greenhouse–Geisser corrected, $\varepsilon = 0.57$), showed that this effect is significant [$F(3,21) = 5.50$, $p < 0.05$].

TIME COURSE OF PERCEPT PROBABILITIES

To study the time evolution of the percept, the current probability of the two-stream percept was calculated at every instant. This is displayed in **Figure 4A**. Since only very few responses occurred within the first second for almost all frequency differences, these time courses started at 2 s for all conditions except for the Δf of 6 semitones. As expected, the two-stream probability settles at higher values for larger Δf (darker lines). Furthermore, all two-stream courses take several seconds to grow to the final probability value. To assess if this is indeed a build-up behavior, the two-stream data were rescaled by the probability that any response has occurred yet (i.e., the sum of the probabilities of both percepts) according to Deike et al. (2012). The course of this normalized two-stream probability reaches a maximum already after the first two seconds for the highest Δf condition (**Figure 4B**). However, for lower Δf values, the probability still increases slowly.

DISCUSSION

PERCEPTUAL SEGREGATION IN CI USERS

The present study investigated whether, with increasing frequency separation Δf between the A and B harmonic complexes, CI users showed an increase of the two-stream percept that is similar

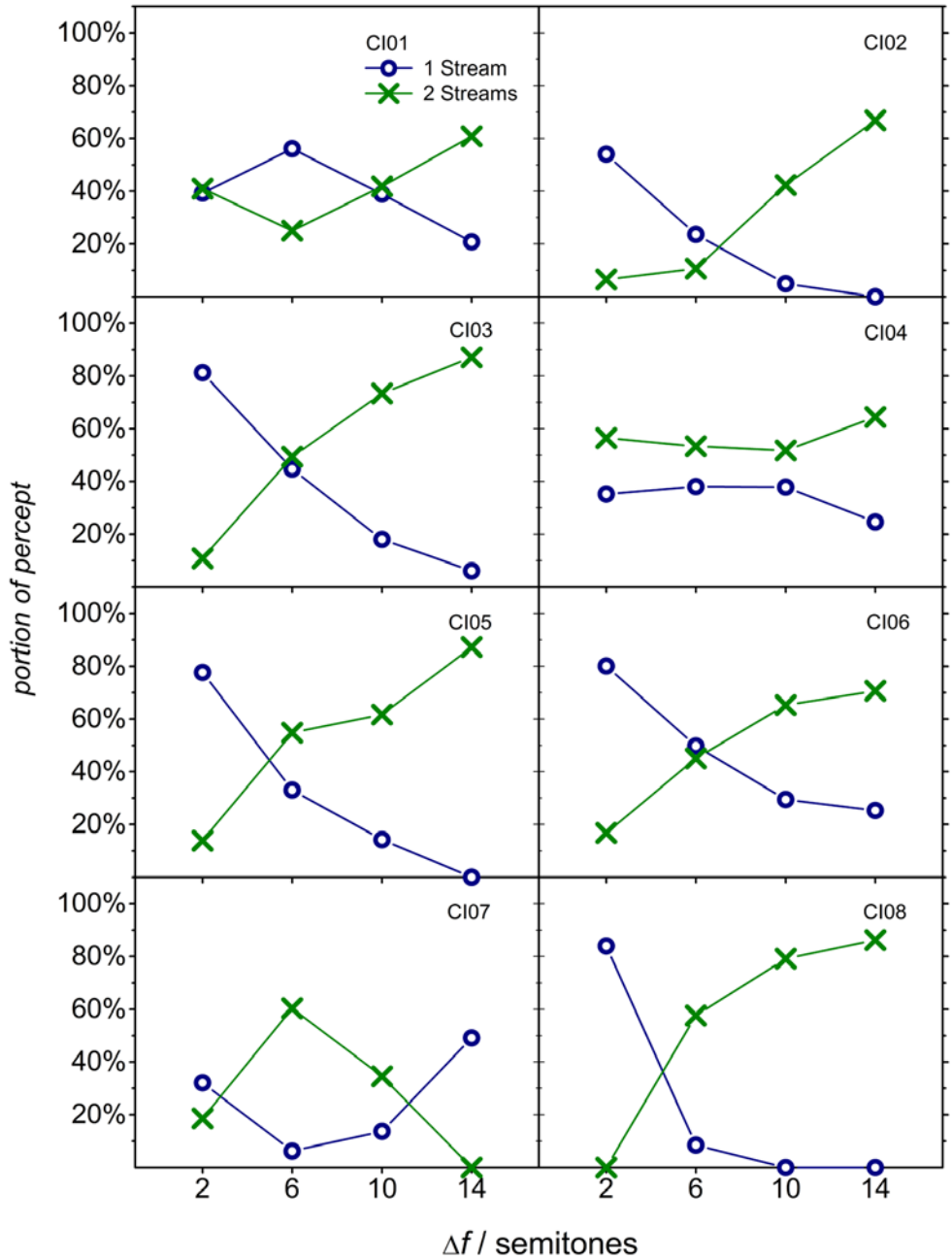


FIGURE 1 | Individual durations of answering options "two-streams" (green cross marks) and "one-stream" (blue circles), calculated as proportions of the total sequence duration (30 s), are plotted as a function of the (average) Δf for each listener.

to that found in normal-hearing listeners. In contrast to previous studies on stream segregation in CI users which stimulated the implant electrodes either directly (Chatterjee et al., 2006) or at the center frequencies of custom-fitted channels, the present study tested stream segregation with the same stimuli as used for normal-hearing listeners in an every-day listening situation (free-field presentation) where the listeners used their familiar CI processor settings. An additional roving by 5 semitones within the A and B sequences minimized effects due to the position of the signal frequencies with respect to the frequency setting of the CI device. The proportions of the one-stream percept and the two-stream percept resemble the results obtained with the same paradigm in normal-hearing listeners (cf. Figure 2B in Deike et al., 2012): With increasing Δf, the proportion assigned to a two-stream percept increased significantly and the proportion assigned to a one-stream percept decreased. In this respect, the average results are consistent with the presence of stream segregation in CI users.

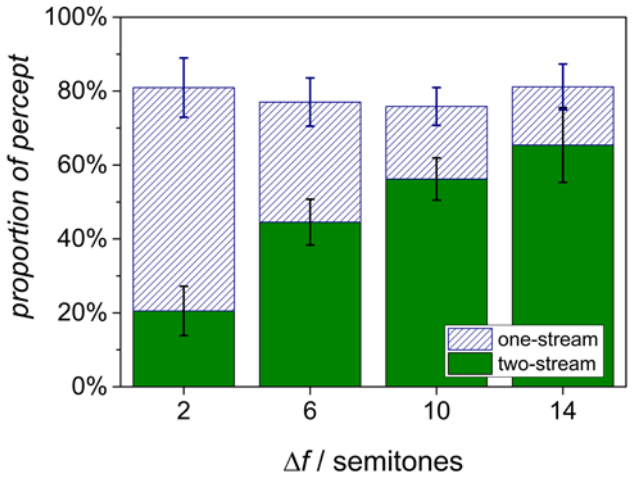


FIGURE 2 | Grand means of the proportions of the two percepts from eight listeners for different Δ*f* conditions. Error bars denote the standard error of mean.

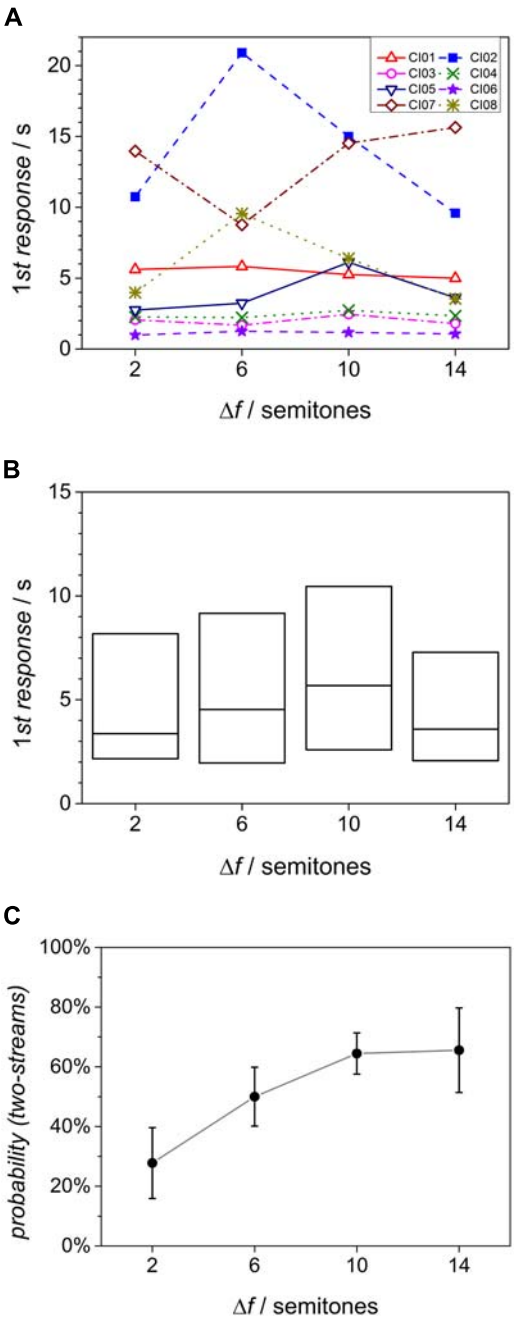


FIGURE 3 | (A) Time of the first response; individual values from eight listeners. **(B)** Box plot of the time of first response. Box boundaries mark the lowest and highest quartiles, center lines the median. **(C)** Type of first decision for different Δ*f* conditions.

The average proportion of a two-stream percept varied from about 20% at a Δ*f* of 2 semitones to 65% at a Δ*f* of 14 semitones. The corresponding average values of normal-hearing listeners were about 5 and 80%. Thus, even at a Δ*f* of 14 semitones there remains a considerable larger proportion attributed to a one-stream percept in CI users than in normal-hearing listeners. While the chosen Δ*f* range covers both extremes of the one-stream and two-stream percepts for the normal-hearing listeners, the CI users showed on average a smaller dynamic range in their response. This is mainly due to the large inter-individual variation found in the participating CI users: Only five of the eight studied listeners responded with a monotonic change in response from one stream to two streams as the frequency separation increased. The other three listeners (CI01, CI04, and CI07) provided responses rather independent of Δ*f*, indicating that these listeners had a similar perception for all frequency differences. These individual data are not consistent with typical results of normal-hearing and hearing-impaired listeners, where the amount of segregation increased as the difference between A and B sounds increased (Rose and Moore, 2005). In the absence of stream segregation one may expect only one-stream responses. For these CI users of the present study, however, the proportions of one-stream and two-stream percepts were comparable at all Δ*f* levels. It is conceivable that, for these particular CI users, the two percepts are possible and equally probable for all conditions used in the present study, even at an average frequency separation of only 2 semitones. Alternatively, this result may be interpreted as an indication of a large uncertainty of the task. According to this interpretation, these listeners experienced a percept that is similar in all conditions but is different from segregation. Taken together, three of the eight CI listeners of this study responded in a way that is different from normal-hearing listeners, i.e., without a monotonic increase of the two-stream percept with frequency difference. Previous studies on stream segregation in CI users reported a similar ratio with three out of eight (Cooper and Roberts, 2007), or one out of five (Hong and Turner, 2006) listeners failing to show a classical stream segregation. As stated for numerous other studies on psychophysical tasks in CI users, for example music perception, it is difficult to isolate factors that might be responsible for individual performance (cf. Looi, 2008; Limb and Roy, 2014). These may be found on all stages from the coupling to the auditory nerve to the listening experience of the listeners with non-speech material.

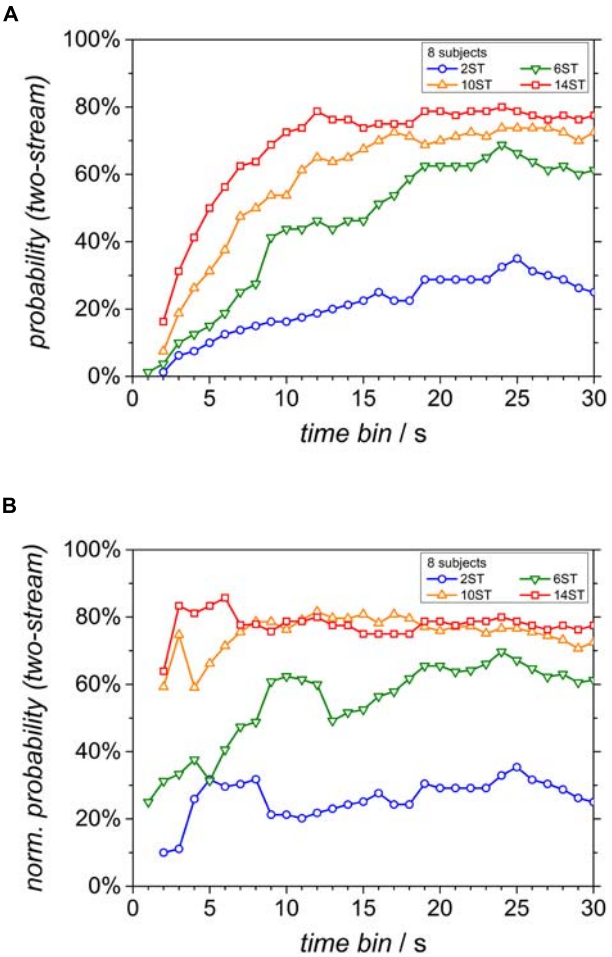


FIGURE 4 | Time courses of the current probability of the two-stream percept for the different Δf conditions. (A) Conventional display and **(B)** normalized probability, rescaled by the probability that any response has been given yet (see text).

It is noteworthy that the available CI users in the present study were all aged above 60, whereas the normal-hearing listeners of Deike et al. (2012) were young adults below 40. Thus, one may argue that the comparison was confounded by age. In agreement with this hypothesis, Rose and Moore (2005) found increased fission boundaries in some but not all bilaterally hearing-impaired listeners. Those listeners were also of higher age than the normal-hearing control listeners. However, whereas the increased fission boundaries could not be explained completely by hearing loss, there was also no evident effect of age from their data. When controlling for the hearing loss, Snyder and Alain (2007) found for the psychoacoustical performance in sequential stream segregation no effect that could be attributed to age itself. Furthermore, there is little evidence for a deterioration of a general CI performance in progressive age. Moreover, the stability of the performance over decades in several audiological tests has been shown in a large population of CI users (Lenarz et al., 2012). Therefore, the age effect has presumably only little effect in the comparison of the CI users tested here to the data of Deike et al. (2012).

The present results are in quantitative agreement with two preceding studies investigating stream segregation in CI users, showing that the occurrence of a two-stream percept requires a separation of two to three electrodes (Chatterjee et al., 2006; Cooper and Roberts, 2007). Whereas the first study did not provide information on the processor settings, the latter reported a frequency ratio of about 1.3 for a separation of two electrodes and about 1.5 for three electrodes. This corresponds to 5–7 semitones on the musical scale. The present study measured a separation of 6 semitones eliciting an ambiguous percept and 10 semitones a prevalence of segregation. Apart from the large inter-individual variability, the average data of the listeners showing stream segregation are very similar to those of the normal-hearing listeners. The fact that the most ambiguous condition with equivalent proportions of both percepts is found at 6 semitones, as in the data of normal-hearing listeners, may be surprising since the frequency difference limen in CI users is much larger than in normal-hearing listeners: In free-field presentation, on average about 3–4 semitones were found (Pretorius and Hanekom, 2008). In light of this strongly reduced frequency discrimination of CI users, one may have expected a significantly higher Δf necessary to elicit stream segregation. The similarity of the data of CI users and normal-hearing listeners argues against the hypothesis that stream segregation is determined by the ability to discriminate frequencies. This is consistent with the findings of a poor correlation of the fission boundary and the frequency difference limen of hearing-impaired listeners (Rose and Moore, 2005). As the fundamental frequency increased, the fission boundary remained constant when expressed in units of equivalent rectangular bandwidth, whereas the frequency difference limen increased. Thus the ratio of fission boundary and frequency discrimination is not constant, at least not in listeners with an impaired hearing. The finding of the present study is also consistent with another study suggesting that the necessary fundamental frequency difference for stream segregation in CI users was of the order of the values found in non-musical, normal-hearing listeners (Marozeau et al., 2013).

TIME ELAPSED BEFORE FIRST RESPONSE

The traditional hypothesis of build-up is that a two-stream percept emerges from an initial, default one-stream percept (Anstis and Saida, 1985; Carlyon et al., 2001; Boehnke and Phillips, 2005; Micheyl et al., 2005; Pressnitzer et al., 2008). Following this hypothesis, one may expect for the current experiment a shorter response time if the first percept is that of one-stream (the default) than if it is a two-stream percept. In the latter condition, the listener would have to change the impression from the default to the two-stream percept before responding. The current data do not support this hypothesis since the response time was not affected by Δf. On the contrary, it was on average shorter for the 14 semitones condition than for 2 semitones.

The median response times of the CI listeners range from about 3.5–6.5 s depending on the condition. It is evident from Figure 3B in Deike et al. (2012) that except for the 6 semitones condition the normal-hearing listeners provided 50% of the responses within the first two seconds of a sequence. This

corresponds roughly to the lowest quartile of the CI listeners. Therefore, the time needed for the first decision is considerably shorter for the normal-hearing listeners than for the CI users (Deike et al., 2012). It is unlikely that this is due to the sound processing in the CI device. It rather indicates perceptual differences between the two groups of listeners. The CI users seem to be generally less confident in their perceptual decision than normal-hearing listeners. This interpretation agrees well with a "context effect" reported by Marozeau et al. (2013): For the same large frequency separation between A and B tones, the difficulty rating to segregate A tones was rated as lower when the average frequency separation of the A and B tones continuously decreased from large to small than when it increased from small to large values. This effect was not observed in the normal-hearing listeners and indicates that it is more difficult for the CI users to find the appropriate cue for segregation. One should note that the time needed for the first decision was also large in normal-hearing listeners (Deike et al., 2012), considerably larger than what is commonly found in, e.g., simple pitch discrimination experiments (Ritter et al., 1972). This indicates that the task of the present study is of higher complexity than that for simple auditory discrimination.

DO CI USERS SHOW A BUILD-UP OF SEGREGATION?

The time courses of the two types of percept based on the traditional analysis with a one-stream percept as the initial default resemble that of previous studies with normal-hearing listeners, including the data shown in Deike et al. (2012): The probability of the two-stream percept increased toward the end of the sequence (see **Figure 4A**) and is very similar to the results of normal-hearing listeners shown in Figure 3A of Deike et al. (2012). Note, however, that this slow increase cannot immediately be attributed to a build-up since it ignores the fact that any response requires some time to be given. Taking into account the probability that any response had occurred, the probability of the two-stream percept started at a high level in conditions with a high Δf (**Figure 4B**), indicating that the two-stream percept prevailed immediately. This is again similar to the data of normal-hearing listeners when analyzed in this way (see in Figure 3C of Deike et al., 2012). In the normal-hearing listeners it started at about 100% two-stream percept and tended to decrease as the presentation time increased. In contrast, for the CI users, this value increased slightly within the first three seconds even for the largest Δf. However, it is clearly in favor of the two-stream percept already at the moment of the first decision. One should note that the data do not necessarily reveal the perception at the very beginning of the stimulus. It could either be that the initial perception is already that of the first decision or that they do not perceive any of the alternatives (neither one nor two streams). The data of normal-hearing listeners and CI users argue against the build-up hypothesis, i.e., the initial one-stream percept which may change to a two-stream percept (Micheyl et al., 2005; Pressnitzer et al., 2008).

Consistent with the results in the normal-hearing listeners, a build-up of the two-stream percept was only observed at the 6-semitones condition, settling at a probability close to 60%, i.e., about equal prevalence of both percepts. This is consistent with the data of normal-hearing listeners. They also showed at a Δf of 6 semitones a continuous increase asymptotically converging to 60%. It is also in line with the results of Denham et al. (2013). There, the initial percept was also two-stream at large frequency separations and high presentation rates. Furthermore, without normalization of the probability the time constant of build-up strongly decreased with increasing frequency separations. For CI users as well as for normal-hearing listeners, a build-up of segregation is thus restricted to a region of high ambiguity of the sequence of stimuli. The data argue against build-up as a general description for the time course of stream segregation.

In summary, the present experiment provides evidence that most of the CI users show stream segregation that is comparable with normal-hearing listeners. For large frequency differences of the A and B tones, a two-stream percept is predominant. The similarity between the results of the CI users and normal-hearing listeners indicates that the quality of stream formation is not altered when the auditory input is provided via a CI. This argues against a strong relation of stream segregation and frequency discrimination since the latter is affected by the limitations of the CI. For the experimental design of stream segregation in CI users, the present findings suggest that durations of more than 20 s are advisable to account for the prolonged time needed by the CI users to form a certain type of streaming percept.

ACKNOWLEDGMENTS

The authors are indebted to two reviewers for valuable comments and suggestions on a previous version of the manuscript. This work was supported by the "Deutsche Forschungsgemeinschaft" (SFB/TRR31) and MED-EL GmbH Germany.

REFERENCES

Anstis, S., and Saida, S. (1985). Adaptation to auditory streaming of frequency-modulated tones. *J. Exp. Psychol. Hum. Percept. Perform.* 11, 257–271. doi: 10.1037/0096-1523.11.3.257

Boehnke, S. E., and Phillips, D. P. (2005). The relation between auditory temporal interval processing and sequential stream segregation examined with stimulus laterality differences. *Percept. Psychophys.* 67, 1088–1101. doi: 10.3758/BF03193634

Bregman, A. S. (1978). Auditory streaming is cumulative. *J. Exp. Psychol. Hum. Percept. Perform.* 4, 380–387. doi: 10.1037/0096-1523.4.3.380

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge, MA: MIT Press, 1990.

Bregman, A. S., and Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *J. Exp. Psychol.* 89, 244–249. doi: 10.1037/h0031163

Carlyon, R. P., Cusack, R., Foxton, J. M., and Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *J. Exp. Psychol. Hum. Percept. Perform.* 27, 115–127. doi: 10.1037/0096-1523.27.1.115

Chatterjee, M., Sarampalis, A., and Oba, S. I. (2006). Auditory stream segregation with cochlear implants: a preliminary report. *Hear. Res.* 222, 100–107. doi: 10.1016/j.heares.2006.09.001

Cooper, H. R., and Roberts, B. (2007). Auditory stream segregation of tone sequences in cochlear implant listeners. *Hear. Res.* 225, 11–24. doi: 10.1016/j.heares.2006.11.010

Cooper, H. R., and Roberts, B. (2009). Auditory stream segregation in cochlear implant listeners: measures based on temporal discrimination and interleaved melody recognition. *J. Acoust. Soc. Am.* 126, 1975–1987. doi: 10.1121/1.3203210

Deike, S., Heil, P., Böckmann-Barthel, M., and Brechmann, A. (2012). The build-up of auditory stream segregation: a different perspective. *Front. Psychol.* 3:461. doi: 10.3389/fpsyg.2012.00461

Denham, S. L., Gyimesi, K., Stefanics, G., and Winkler, I. (2013). Perceptual bistability in auditory streaming: how much do stimulus features matter? *Learn. Percept.* 5(Suppl. 2.6), 73–100. doi: 10.1556/LP.5.2013

Hong, R. S., and Turner, C. W. (2006). Pure-tone auditory stream segregation and speech perception in noise in cochlear implant recipients. *J. Acoust. Soc. Am.* 120, 360–374. doi: 10.1121/1.2204450

Lenarz, M., Sönmez, H., Joseph, G., Büchner, A., and Lenarz, T. (2012). long-term performance of cochlear implants in postlingually deafened adults. *Otolaryngology* 147, 112–118. doi: 10.1177/0194599812438041

Limb, C. J., and Roy, A. T. (2014). Technological, biological, and acoustical constraints to music perception in cochlear implant users. *Hear. Res.* 308, 13–26. doi: 10.1016/j.heares.2013.04.009

Looi, V. (2008). The effect of cochlear implantation on music perception: a review. *Otorinolaringologia* 58, 169–190.

Marozeau, J., Innes-Brown, H., and Blamey, P. J. (2013). The acoustic and perceptual cues affecting melody segregation for listeners with a cochlear implant. *Front. Psychol.* 4:790. doi: 10.3389/fpsyg.2013.00790

Micheyl, C., Tian, B., Carlyon, R. P., and Rauschecker, J. P. (2005). Perceptual organization of tone sequences in the auditory cortex of awake macaques. *Neuron* 48, 139–148. doi: 10.1016/j.neuron.2005.08.039

Miller, G. A., and Heise, G. A. (1950). The trill threshold. *J. Acoust. Soc. Am.* 22, 637–638. doi: 10.1121/1.1906663

Pressnitzer, D., Sayles, M., Micheyl, C., and Winter, I. M. (2008). Perceptual organization of sound begins in the auditory periphery. *Curr. Biol.* 18, 1124–1128. doi: 10.1016/j.cub.2008.06.053

Pretorius, L. L., and Hanekom, J. J. (2008). Free field frequency discrimination abilities of cochlear implant users. *Hear. Res.* 244, 77–84. doi: 10.1016/j.heares.2008.07.005

Ritter, W., Simson, R., and Vaughan, H. G. (1972). Association cortex potentials and reaction time in auditory discrimination. *Electroencephaolgr. Clin. Neurophysiol.* 33, 547–555. doi: 10.1016/0013-4694(72)90245-3

Rose, M. M., and Moore, B. C. J. (2005). The relationship between stream segregation and frequency discrimination in normally hearing and hearing-impaired subjects. *Hear. Res.* 204, 16–28. doi: 10.1016/j.heares.2004.12.004

Snyder, J. S., and Alain, C. (2007). Sequential auditory scene analysis is preserved in normal aging adults. *Cereb. Cortex* 17, 501–512. doi: 10.1093/cercor/bhj175

van Noorden, L. P. A. S. (1975). *Temporal Coherence in the Perception of Tone Sequences.* Ph.D. dissertation, University of Technology, Eindhoven.

Conflict of Interest Statement: The study was partly supported by a research grant of MED-EL GmbH Germany, providing means for equipment and subject expenses.

Received: 18 March 2014; accepted: 01 July 2014; published online: 21 July 2014.
*Citation: Böckmann-Barthel M, Deike S, Brechmann A, Ziese M and Verhey JL (2014) Time course of auditory streaming: do CI users differ from normal-hearing listeners? Front. Psychol. **5**:775. doi: 10.3389/fpsyg.2014.00775*
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Psychology.

frontiers in
NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 21 July 2014
doi: 10.3389/fnins.2014.00203

Task-dependent neural representations of salient events in dynamic auditory scenes

Lan Shuai and Mounya Elhilali*

Laboratory of Computational Audio Perception, Department of Electrical and Computer Engineering, Center for Speech and Language Processing, Johns Hopkins University, Baltimore, MD, USA

Edited by:
Elyse S. Sussman, Albert Einstein College of Medicine, USA

Reviewed by:
Peter Lakatos, Hungarian Academy of Sciences, Hungary
Iria SanMiguel, Leipzig University, Germany

***Correspondence:**
Mounya Elhilali, Department of Electrical and Computer Engineering, Johns Hopkins University, 3400 N. Charles Street, Barton Hall Rm. 105, Baltimore, MD 21218, USA
e-mail: mounya@jhu.edu

Selecting pertinent events in the cacophony of sounds that impinge on our ears every day is regulated by the acoustic salience of sounds in the scene as well as their behavioral relevance as dictated by top-down task-dependent demands. The current study aims to explore the neural signature of both facets of attention, as well as their possible interactions in the context of auditory scenes. Using a paradigm with dynamic auditory streams with occasional salient events, we recorded neurophysiological responses of human listeners using EEG while manipulating the subjects' attentional state as well as the presence or absence of a competing auditory stream. Our results showed that salient events caused an increase in the auditory steady-state response (ASSR) irrespective of attentional state or complexity of the scene. Such increase supplemented ASSR increases due to task-driven attention. Salient events also evoked a strong N1 peak in the ERP response when listeners were attending to the target sound stream, accompanied by an MMN-like component in some cases and changes in the P1 and P300 components under all listening conditions. Overall, bottom-up attention induced by a salient change in the auditory stream appears to mostly modulate the amplitude of the steady-state response and certain event-related potentials to salient sound events; though this modulation is affected by top-down attentional processes and the prominence of these events in the auditory scene as well.

Keywords: attention, auditory scene, stead-state response, bottom-up, top-down

INTRODUCTION

Selecting relevant information from an auditory scene is guided either by the salience of the acoustic events (bottom-up driven) or by behavioral goals (top-down driven). While bottom-up and top-down attentional mechanisms engage different neural circuits of sensory processing and cognitive control (Buschman and Miller, 2007), their net effect on the neural representation of acoustic events appear to share many similarities (Katsuki and Constantinidis, 2014). Indeed, both forms of bottom-up and top-down attentional mechanisms have been reported to give rise to enhanced or better tuned responses to a relevant stimulus relative to the background, both at the single neuron level as well as the population level reflected in network analyses, event-related potentials and interactions across brain areas (Näätänen, 1992; Katsuki and Constantinidis, 2014). Specifically, saliency-driven attentional mechanisms are greatly reflected in the stimulus representation at the level of sensory cortex as well as the propagation of information to frontal areas, all while modulating key markers of the neural response as reflected in event-related components such as Mismatch Negativity (MMN) (Näätänen and Michie, 1979; Näätänen et al., 2007; Kim, 2014). Such changes are often accompanied by further modulation due to task-demands and guided by top-down attention which also affects sensory and frontal cortical networks (Picton and Hillyard, 1974b; Luck et al., 2000; Fritz et al., 2007a; Kiefer, 2007; Elhilali et al., 2009; Müller et al., 2009; Xiang et al., 2010). These effects are not unique to the auditory modality. They have been widely reported in vision as well (Buschman and Miller, 2007; Zhaoping, 2008; Andersen et al., 2012; Zhang et al., 2012), with stronger suggestions of independence between bottom-up and top-down attention despite some of their shared measureable effects (Pinto et al., 2013).

The role of attention in guiding processing of auditory scenes parsing is further regulated by the complexity of the scene itself, as well as the dynamic nature of the sound streams competing for a listener's attention. There is strong evidence suggesting that some of the neural mechanisms engaged in parsing complex acoustic scenes are in fact independent of top-down attention (Bregman, 1990; Sussman and Steinschneider, 2001). These attention-free processes are mostly a reflection of the acoustic properties and statistical nature of the stimulus which can bias its organization into mental representations perceived as segregated streams or salient events in the scene. In other words, parsing an acoustic scene is partly dictated by the physical nature of the acoustic cues present and the statistical evolution of these cues over time; which not only define the perceptual boundaries of the different auditory objects in the soundscape, but also direct the brain's computational resources to the relevant events in the scene based on their conspicuity and sensory relevance. These in turn can facilitate the process of top-down attention (rather than only vice versa).

In the current study, we try to tease apart the neural signature of bottom-up and top-down components of attention in a series of experiments, by focusing on the representation of salient events under or away from the spotlight of top-down attention. We embedded a salient change in a dynamic sound stream, and controlled listeners' attention to or away from this stream; both in presence or absence of a second competing sound stream; while measuring their neurophysiological responses using Electroencephalography (EEG). Key to our paradigm is that salient events are not defined by a simple form of deviance detection. Salience is defined here in a statistical sense that salient events diverge from the feature distribution of the dynamic auditory stream. This definition is chosen to clearly dissociate it from commonly used deviance-detection paradigms or odd-ball designs which rely on a standard sound event that is either physically repeated a number of times or with at least one of its acoustic properties of interest held fixed and repeated to establish a standard reference. In contrast, the paradigm presented here relies on a notion of standard that is defined only in a statistical sense, in that the feature of interest in the signal never repeats, but rather is drawn from a constrained statistical distribution. The implication of this design is that our brain is collecting or estimating these statistics as the signal evolves over time; and it infers that any changes from the underlying distribution as a violation of the statistical structure of the signal that would be deemed salient. Our working hypothesis is that the presence of a salient sound token as part of an auditory stream would trigger neural changes that reflect both its deviance from the existing object as well as reflect its salience as dictated by its prominence in presence or absence of competing streams. Such salient neural representation would be modulated by the attentional state of listeners, though parts of its neural signature would be independent of it. This hypothesis implies a degree of independence in the observed neural changes between bottom-up and top-down attention, whereby variations in the neural response due to salient changes in the acoustic structure of the scene could be observed and comparable in size independently of whether listeners were attending to the sound stream or not.

Our analysis focused on the auditory steady-state response (ASSR) as well as event-related potentials (ERPs) that were associated with the target sound stream which contained salient tokens. The ASSR analysis provides us a window into entrainment effects caused by the repeating nature of the stimulus and the degree to which they are modulated by both attentional state and complexity of the scene. These results complement the analysis based on event-related potentials; since the ASSR analysis is a frequency-based decomposition of a sequence of neural responses assessing their phase-locked nature, while the ERP analysis is a time-locked profile of the neural response that ignores the phase and pattern-locked changes over time though would indirectly influence the ASSR since a stronger stimulus-locked steady-state response would reflect a stronger match in the repeating pattern and hence a stronger alignment between responses to individual stimulus events. While there is an indirect association between these two components, they are not directly mapped from one to the other (Capilla et al., 2011; Zhang et al., 2013) and provide complementary views into the effect of stimulus and attentional manipulations on the neural response.

MATERIALS AND METHODS

STIMULI

A total of three experiments were performed where both stimuli and attentional state of listeners were manipulated. One experiment included single-stream stimuli, and the other two included dual-stream stimuli (**Figure 1A**). The single-stream stimuli consisted of a sequence of concatenated synthesized musical notes. The notes were synthesized from spectral templates of three instruments: trumpet, saxophone and clarinet from the RWC instrument database (Goto et al., 2003; Goto, 2004). Each note in the sequence was defined by four features: intensity, duration, pitch and timbre (spectral template). Each of the four features was drawn from a discrete distribution consisting of 12 levels. These 12 levels were separated into three groups. Level 1–4 belonged to the first group, 5–8 belonged to the second group, and 9–12 belonged to the third group. By separating notes this way, feature changes within each group were smaller while changes between

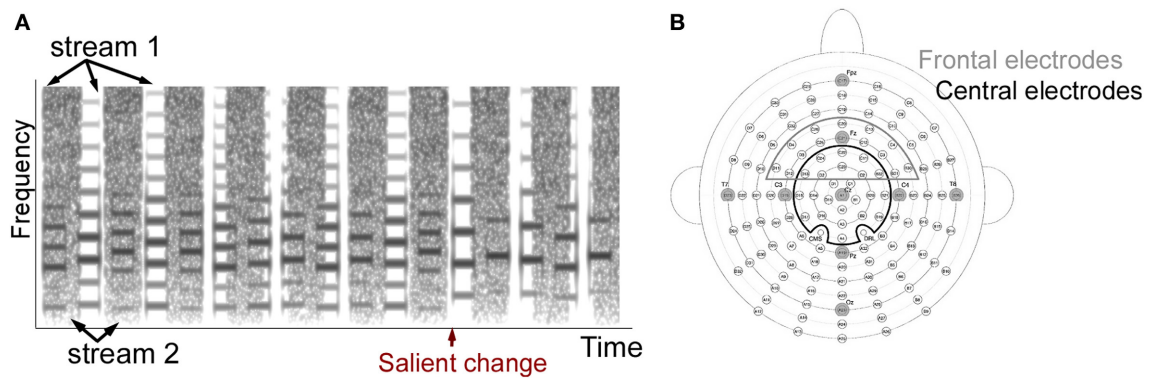


FIGURE 1 | (A) Example stimulus spectrogram used in the two-stream experiments. The spectrogram of one sample trial from the two-stream experiment is shown. The stimulus consists of a musical notes stream and a sequence of noise bursts. The one-stream experiment did not include the noise stream. **(B)** Electrode positions in the 128-channel Biosemi system. Cz and 23 electrodes around Cz were included as the Cz electrode group, and Fz and 21 electrodes were included as the Fz electrode group. Original layout is available at: http://www.biosemi.com/pics/cap_128_layout_medium.jpg.

groups were larger on average. Detailed parameters of each feature dimension and each group are listed in **Table 1**.

Each experimental trial started with an auditory stream consisting of a sequence of notes drawn from the same feature group. Salient notes would then appear in the stream at a pseudo-random moment in the sequence, in which the sound tokens were drawn from a different feature group. Control trials did not contain any salient notes, but consisted of only a stream whose notes varied within one feature group. There were a total of 400 experimental trials and 80 control trials. The order of trials was randomized across subjects. In each trial, only one dimension of the features was changing both during the auditory stream and salient events segment. For example, a given trial would be made of music notes with varying pitch values (e.g., group1 pitches: G3–A3#) but fixed timbre, intensity and duration features. The salient change would share the same timbre, intensity and duration but introduce a higher pitch drawn from group 2 or group 3 (see **Table 1**). Each note within one group had the same probability to appear, and no adjacent notes were the same. The SOA (stimulus onset asynchrony) was set to 213 ms, hence the tempo of the entire sequence was fixed at 4.7 Hz. Each note in the sequence included a 10 ms cosine ramp at onset and offset. The auditory stream contained 6, 12, 18, 24, or 30 notes; each repeated 80 experimental trials. Within each 80 trials, 20 trials had one of the four feature dimension changes. Ten of 20 trials had greater level jump (group 1 to 3 or vice versa) and 10 had smaller level jump (e.g., group 1 to 2). The salient change segment was composed of 12 notes. The control trials consisted of 16 trials of 18, 24, 30, 36, or 42 notes. The length of control trials matched the whole duration of experimental trials.

The dual-stream stimuli consisted of two streams playing simultaneously (**Figure 1A**). The first stream was similar to the note sequence described above. Simultaneously, a second stream of modulated noise tokens was played (**Figure 1A**). The second stream consisted of a sequence of flat spectrum white noise tokens that were sinusoidally-modulated at 3 cycles/octave (Chi et al., 1999) with either modulation depth 0 (flat) or 0.35 (modulated). A trial consisted of either flat noise tokens with a modulated deviant; or modulated noise tokens with a flat deviant. The deviant always appeared in one of the last 3 tokens in the sequence; with each of the 3 last positions being a deviant with equal probability. Control trials did not contain any deviant tokens. Each instance of noise was generated separately in order to avoid memorizing the temporal sequence of the noise. A modulation depth of 0.35 was chosen based on a pilot study indicating that this modulation depth was around the threshold of detection for average listeners, as modulation of 0.4 was easily detectable and 0.3 was hardly detectable. This value was slightly higher than previous reports of modulation detection thresholds in white noise (Chi et al., 1999) and could be due to the existence of the competing sound stream which increases the difficulty of the task.

The noise streams also comprised 480 trials; half consisting of flat noise standards and half of modulated noise standards. Fifty percent of all trials were controls with no deviant (half of them with flat noise standards, and half with modulated noise standards). Each noise token was 230 ms long with 10 ms onset and offset cosine ramps; and the SOA was set to 382 ms (corresponding to a tempo of about 2.6 Hz). The entire noise stream was composed of 11, 14, 17, 21, or 24 sound tokens to match the length of the music notes stream. The noise streams were randomly added to the music notes stream to form the dual-sound stream after setting the intensity of both sound streams to the same value over the full trial length.

The experiments were organized in three conditions: (1) attend to a salient change in the music sound stream (one-stream attend); (2) attend to change in the music sound stream while a noise sound stream is simultaneously playing in the background (two-stream attend); and (3) ignore music sound stream while attending to the noise sound stream (two-stream ignore).

PARTICIPANTS AND PROCEDURE

Thirty-six healthy volunteers between 18 and 35 years old (mean age: 21.81; std: 3.94) with no history of hearing problems or neurological disorders according to self-report participated in the

Table 1 | Stimulus parameter space for stream 1 stimuli.

SPECTRAL TEMPLATE (TRUMPET: T; SAXOPHONE: S; CLARINET: C)				
Group 1 [Level 1–4]	1*T	4.5/5.5*T + 1/5.5*S	3.5/5.5*T + 2/5.5*S	2.5/5.5*T + 3/5.5*S
Group 2 [L 5–8]	1.5/5.5*T+4/5.5*S	0.5/5.5*T+5/5.5*S	0.5/5.5*C+5/5.5*S	1.5/5.5*C+4/5.5*S
Group 3 [L 9–12]	2.5/5.5*C+3/5.5*S	3.5/5.5*C+2/5.5*S	4.5/5.5*C+1/5.5*S	1*C
PITCH				
Group 1 [L 1–4]	196 Hz (G3)	208 Hz (G3#)	220 Hz (A3)	233 Hz (A3#)
Group 2 [L 5–8]	247 Hz (B3)	262 Hz (C4)	277 Hz (C4#)	294 Hz (D4)
Group 3 [L 9–12]	311 Hz (D4#)	330 Hz (E4)	349 Hz (F4)	370 Hz (F4#)
DURATION				
Group 1 [L 1–4]	40 ms	50 ms	60 ms	70 ms
Group 2 [L 5–8]	80 ms	90 ms	100 ms	110 ms
Group 3 [L 9–12]	120 ms	130 ms	140 ms	150 ms
INTENSITY				
Group 1 [L 1–4]	60 dB	61.25 dB	62.5 dB	63.75 dB
Group 2 [L 5–8]	65 dB	66.25 dB	67.5 dB	68.75 dB
Group 3 [L 9–12]	70 dB	71.25 dB	72.5 dB	73.75 dB

experiments with informed consent. Due to the length of experimental conditions, subjects were randomly assigned to one of the three experimental conditions described above (12 subjects in each experiment), and compensated for their participation. All experimental procedures were approved by the Homewood Institutional Review Board (HIRB). Subjects sat in a comfortable chair inside a dim lighted sound-dampened chamber. Ambient noise inside sound booth was around 30 dB SPL tested by a sound level meter. A computer screen was placed 1.5 meters in front of participants. Participants wore a pair of ER-3A insert earphones during the experiment. Sound pressure level from the earphones was adjusted to a range between 60 and 80 dB SPL measured by a sound level meter.

EEG recording was done by using a 128-channel Biosemi Active Two system (Biosemi Inc., The Netherlands) with a sampling frequency at 2048 Hz (**Figure 1B**). Offsets of each channel were kept under 40 dB examined after preparation. Online filters were set from 0 to 417 Hz following the default values. Left and right mastoid electrodes were also recorded and used as averaged mastoid references during offline processing. The nose channel was recorded serving as a reference to check for existence of a Mismatch Negativity (MMN) component. Eye artifacts were monitored by recording four EOG channels. Two were placed below left and right eyes and two were put outside of the left and right outer canthus.

The whole experiment lasted about 2.5 h including about 45 min of preparation. Participants were asked to detect the presence of a salient change in each trial and respond after the trial ends by pressing "yes" or "no" buttons on a response box using their left and right index fingers. For half of the participants in each experiment, "yes" button was on the left and "no" answer was on the right, and for the other half it was the opposite. The position of the buttons was randomly assigned to participants. There were 10 practice trials to let participants get familiar with the task. If hit rate was under 60%, participants were given a second 10-trial practice session. A d-prime measure was used to assess the accuracy of detecting a deviant.

DATA PROCESSING

EEG signals were down sampled to 512 Hz and lowpass filtered at 104 Hz using a Decimator software provided by Biosemi, and then further preprocessed using FieldTrip (Oostenveld et al., 2011), MATLAB (MATLAB Release 2013a, the MathWorks, Inc., Natick, Massachusetts, United States), and EEGLab (Delorme and Makeig, 2004). Each data segment included 1065 ms (duration of 5 music notes) immediately before or after the deviant point in the sound stream. Data were re-referenced to the averaged mastoid electrodes as the reference, and a de-mean was applied to each segment to align the average at zero. A denoising procedure was applied by first bandpassing all signals between 0.7 and 30 Hz. Bad channels were marked by an experienced experimenter and flagged as channels with unusual high frequency noise or large drifts. They were replaced by the arithmetic mean of the surrounding channels. No more than two bad channels were replaced for each participant. Eye artifacts were removed by excluding correspondent one or two eye-blink components and one eye-movement component after ICA (Independent Component Analysis) decomposition. Finally, a second 0.7–30 Hz bandpass filter was applied to smooth over minor distortions occasionally introduced by ICA reconstruction. In order to ensure that trial onset effects were not contaminating our analysis, we only analyzed 320 trials out of 400 experimental trials which contained more than one second of the auditory stream.

After data preprocessing, ASSR and ERP responses were analyzed separately. ASSR was derived by first concatenating all trials in each condition for each participant over the 5 notes duration before and after a salient change. A Fourier transform was then applied to obtain the spectral distribution. The amplitude at 4.7 Hz (1/213 ms) was calculated to capture the stimulus-locked steady-state neural response (**Figure 3A**). The ASSR amplitude was confirmed by verifying that a peak at 4.7 Hz was greater than the averaged energy at surrounding frequencies across all participants and conditions (2–6 Hz excluding 4.7 Hz served as baseline). The ASSR analysis was derived from 24 central electrodes (Cz group) around the vertex as shown in **Figure 1B** (including electrodes A1 (Cz), A2, A3, A4, B1, B2, B19, B20, B21, B32, C1, C2, C11, C22, C23, C24, D1, D2, D13, D14, D15, D16, D17, D18 in the layout of 128-channel Biosemi headcap). The lateralization of ASSR was calculated by averaging ASSR peaks in the 56 left and 56 right electrodes excluding the midline electrodes.

Event-related potentials (ERPs) were analyzed by averaging all trials over the duration of two music notes (i.e., 426 ms) immediately before or after the deviant point with baseline correction using 100 ms prior to the segments. This analysis was called 1-note analysis in the results section in order to highlight that the focus is on the early and late components relative to the onset of 1-note. Because of the closeness in timing between successive music notes in the stimulus (SOA 213 ms), the analysis of the late ERP component of a given note was contaminated by onset effects of the following note. The ERP waveform at 0–213 ms was therefore subtracted from the ERP waveform at 213–426 ms in order to analyze any putative late components, including a P300 component. This waveform correction for analysis of the P300 component was repeated using other forms of correction (no correction, or by subtracting onset effects of a standard note instead of deviant one). Though different forms of correction resulted in slightly different shapes of the corrected waveform and peak latencies, all methods of a putative P300 analysis resulted in the same statistical results in terms of significant changes in the P300 component as a function of top-down attention and acoustic saliency.

The analysis based on 1-note required averaging across fewer trials (320 trials) and was repeated using an average of three consecutive music notes (3-notes) with more data (960 trials). In the 3–note analysis, data from 3 notes was averaged then baseline-corrected using 100 ms prior to the onset of the averaged segments. In order to avoid overlapping between the auditory stream and salient change, data starting from the 2nd last music note before the salient event served as standard and the 1st music note after the salient event served as deviant for the 1-note analysis. The data starting from 2nd, 3rd, and 4th last music notes before the salient event served as standard and the 1st, 2nd, and 3rd music notes after the salient event served as deviant for the 3-note analysis. The enlarged N1 was prominently visible only in

the 1-note analysis and concealed changes in the positive P1 peak and the presence of a frontal MMN-like negativity which were only revealed with the 3 note analysis. The late P300 component was consistent across the 1-note and 3-note analysis.

ERP components were defined as follows: (i) An early P1 positivity was defined as a positive peak in standard and deviant ERP curves over the range 30–100 ms from the central Cz electrode groups. We defined the peak latency of the component as the average peak latency across all participants. Once a positive peak was identified, the amplitude over a window of ±15 ms around this peak was averaged and checked for statistical significance relative to the variance in the data. The peak latency for the P1 component was found to be 64 ms for all 3 experiments (1-stream attend, 2-stream attend and 2-stream ignore). The central topography of the P1 peak was confirmed by full head topography to verify that the maximal positivity is indeed localized in the central electrodes. (ii) An early negativity (N1) was analyzed similarly as a significant negative trough in both the standard and deviant ERP curves over the time range 100–213 ms in the Cz electrode group, with analysis window ±15 ms around lowest negativity. The peak latency for the N1 component was found to be at 156 ms for 1-stream attend, 176 ms for 2-stream attend and 166 ms for 2-stream ignore. (iii) An MMN-like component was defined as a significant negativity over 100–213 ms in the difference curve between standard and deviant in the frontal electrodes from the Fz electrode group as shown in **Figure 1B** (electrodes B30, B31, B32, C2, C3, C4, C11, C12, C13, C20, C21, C22, C23, C24, C25, C26, D2, D3, D4, D11, D12, D13). The frontal distribution of the MMN-like component was confirmed by full head topography with either the averaged mastoid reference or the nose reference. The peak latency for this component was found to be at 141 ms for 1-stream attend, 197 ms for 2-stream attend and 182 ms for 2-stream ignore. (iv) A late P300 positivity was defined as a significant positive peak in the difference curve over the time window 300–426 ms in the central Cz electrodes (corrected by ERP curves of the following note at 0–213 ms, as explained earlier). The peak latency for the P300 component was found to be at 360 ms for 1-stream attend, 410 ms for 2-stream attend and 410 ms for 2-stream ignore.

The complex nature of the variables manipulated in the current study required an across-experiment strategy without a full factorial design. Therefore, we conducted a number of statistical tests to alleviate concerns of comparisons across pools of subjects in different experiments. A One-Way ANOVA was performed on the behavioral results in order to detect any differences of performance across experiments. For the analysis of neural responses using ASSR and ERPs, we first conducted pairwise *T*-tests to investigate the effects of the salient event. Since bottom-up attention was one of the most important factors in our study, we reported the effect of salient change in each group as well with Bonforroni correction on multiple comparisons. Apart from the effect of salient change, we also examined the influences of (1) top-down attention or (2) auditory scene complexity on: (a) the auditory stream before the salient change by a univariate analysis on ASSR or MANOVA on ERPs and (b) the increase introduced by the salient change by looking at the interaction between saliency and attention or scene complexity in the Two-Way ANOVA. We also compared the lateralization of auditory stream and the salient change by using pairwise *T*-tests.

RESULTS

Though the complexity of the listening environment differed across experiments (presence of one or two streams), the ability of listeners to indicate the presence of a salient change or deviant in the attended stream was comparable. All three experiments yielded an average d-prime performance between 1.5 and 2 (**Figure 2**). A One-Way ANOVA confirmed that there were no statistical differences in performance between all 3 experiments (ANOVA between the three experiments: $[F_{(2,\ 33)} = 0.259,\ p = 0.733,\ \eta_p^2 = 0.015]$; between one-stream attend and two-stream attend $[t_{(22)} = 0.646,\ p = 0.528]$; or between two-stream attend and two-stream ignore $[t_{(22)} = -0.443,\ p = 0.662)]$. All three tasks were far from ceiling, hence engaging participants' selective attentional processes to a similar degree.

The neural responses reflecting the presence of a repeating pattern in the musical notes stream (whether attended or not) show a strong 4.7 Hz component in the neural signal of individual listeners (**Figure 3A**). **Figure 3A** (top panel) depicts the Fourier transform of the neural response for one participant in the one-stream attend task during the "standard" portion of the music auditory stream revealing a clear peak at 4.7 Hz entrained to the tempo of the music notes. This peak is further enhanced (**Figure 3A**, bottom panel) once a salient change occurs in the musical note stream. It is important to highlight that while its tempo is fixed, the musical notes stream used in this study is not deterministic by nature. It is characterized by small ongoing fluctuations along pitch, timbre, loudness, or duration. Despite this dynamic nature, the presence of a salient-enough change in the stream caused further enhancement to the phase-locked ASSR component. **Figure 3B** shows an analysis of the population response, and reveals that the enhancement was statistically

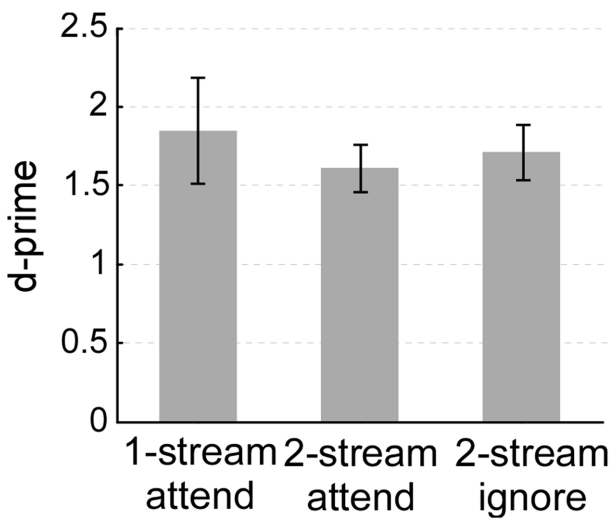


FIGURE 2 | Behavioral performance (d-prime) in the one-stream attend, two-stream attend or two-stream ignore experiments. Error bars denote one standard error.

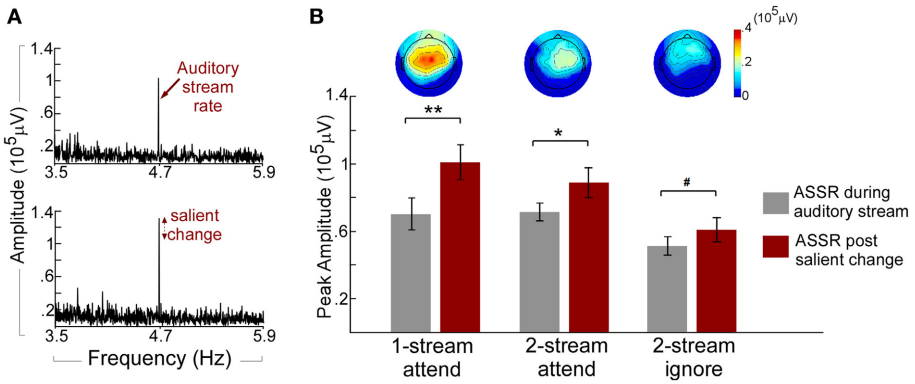


FIGURE 3 | (A) Example ASSR peak during the auditory stream and post salient change. The spectrum of the neural response for one participant in the one-stream attend experiment is obtained by concatenating 320 trials at the Cz electrode. The Top panel shows the spectrum obtained during the auditory stream segment of the trial (prior to any change) and shows a clear peak at 4.7 Hz entrained to the tempo of the music notes. The Bottom panel depicts further enhancement in the 4.7-Hz peak after a salient change is introduced in the stimulus **(B)** ASSR amplitudes of the auditory stream and salient change and topographies of the ASSR increase in the three experiments. Each bar depicts the population average of ASSR peak before (left) and after (right) a salient change in the auditory stream in the Cz electrode group. Error bars denote one standard error. The topographies show the scalp distribution of the ASSR increase (change from before to after salient change). All topographies are shown to the same amplitude range. $^{**}p < 0.01$; $^{*}p < 0.05$; $^{\#}p < 0.1$.

significant irrespective of attentional state (attend or ignore experiments), or complexity of the auditory scene (one stream or two streams); with a central to frontocentral topography across experiments. A pairwise T-test on all participants revealed a significant increase of ASSR during the salient event [$t_{(35)} = 5.239$, $p < 0.001$; Bonferroni adjustment for the following three comparisons $p = 0.017$: one-stream attend: $t_{(11)} = 4.123$, $p < 0.002$; two-stream attend: $t_{(11)} = 3.015$, $p < 0.012$; and marginal significant increase in the two-stream ignore: $t_{(11)} = 2.186$, $p = 0.051$]. This strength in phase-locked responses is akin of previously reported effects of neural representations of salient, rare, unexpected events, particularly at the level of auditory cortex (Gutschalk et al., 2005; Bidet-Caulet et al., 2007; Chait et al., 2012).

In addition to changes in the steady-state response, the appearance of a salient event in the auditory scene evoked changes in the ERPs. ERP components before and after the salient change in the stream were contrasted using pairwise T-tests. An analysis based on averaging 1-note immediately before and after the salient change (see Materials and Methods) revealed a significant change of the early negativity (possibly a long-latency N1 component) in the Cz central electrodes in the attend experiments (**Figure 4**, left column). A pairwise T-test revealed salient changes increased the amplitude of N1 significantly [$t_{(35)} = -6.058$, $p < 0.001$; Bonferroni adjustment for the following three comparisons $p = 0.017$: one-stream attend: $t_{(11)} = -6.816$, $p < 0.001$; two-stream attend: $t_{(11)} = -4.765$, $p < 0.001$; two-stream ignore: $p = 0.426$]. This component could conceal a mismatch negativity (MMN) component but cannot be clearly labeled as such because of its localization at central regions of the scalp especially in the attend experiments (**Figure 4**, left column). In addition, the analysis revealed a significant late P300 positive component in the attend conditions with a predominantly central distribution (**Figure 4**, middle column, Cz electrode group). A pairwise T-test confirmed the significant P300 peak [$t_{(35)} = 5.074$, $p < 0.001$; Bonferroni adjustment for the following three comparisons $p = 0.017$: one-stream attend: $t_{(11)} = 2.830$, $p < 0.016$; two-stream attend: $t_{(11)} = 3.978$, $p < 0.002$; two-stream ignore experiment: $t_{(11)} = 2.142$, $p = 0.055$].

To reveal any concealed ERP components that were not clearly visible from the 1-note averaging, the same analysis was repeated by averaging data from more trials including three notes before and after the salient change (see Materials and Methods). The 3-note analysis confirmed results observed in the late P300 components, with significant enhancement across all three experiments (data not shown in the Figure). The 3-note analysis also revealed a significant difference in the P1 component in the central Cz electrodes. A pairwise T-test revealed significant saliency induced amplitude increases [$t_{(35)} = 6.009$, $p < 0.001$; Bonferroni adjustment for the following three comparisons $p = 0.017$: one-stream attend: $t_{(11)} = 4.012$, $p < 0.002$; two-stream attend: $t_{(11)} = 4.136$, $p < 0.002$; two-stream ignore: $t_{(11)} = 2.578$, $p < 0.026$]. The analysis also revealed a frontocentrally distributed MMN-like component in the difference deviant/standard curve in one-stream attend and two-stream ignore experiments in the Fz electrode group (**Figure 4**, right column). A pairwise T-test showed significant saliency effect in general [$t_{(35)} = -4.025$, $p < 0.001$; Bonferroni adjustment for the following three comparisons $p = 0.017$: one-stream attend: $t_{(11)} = -2.663$, $p < 0.022$; two-stream ignore: $t_{(11)} = -3.049$, $p < 0.011$; two-stream attend: $p = 0.161$].

TOP-DOWN ATTENTION AND SALIENCY EFFECTS

In order to tease apart the contribution of the attentional state of participants and the presence of a salient change in an acoustic stream, we first compared the 4.7 Hz ASSR component during the acoustic scene (prior to any salient change) in the two-stream attend vs. two-stream ignore experiments. Differences between these neural responses directly reflected the modulatory-effect of task-dependent attention given that the acoustic stimulation

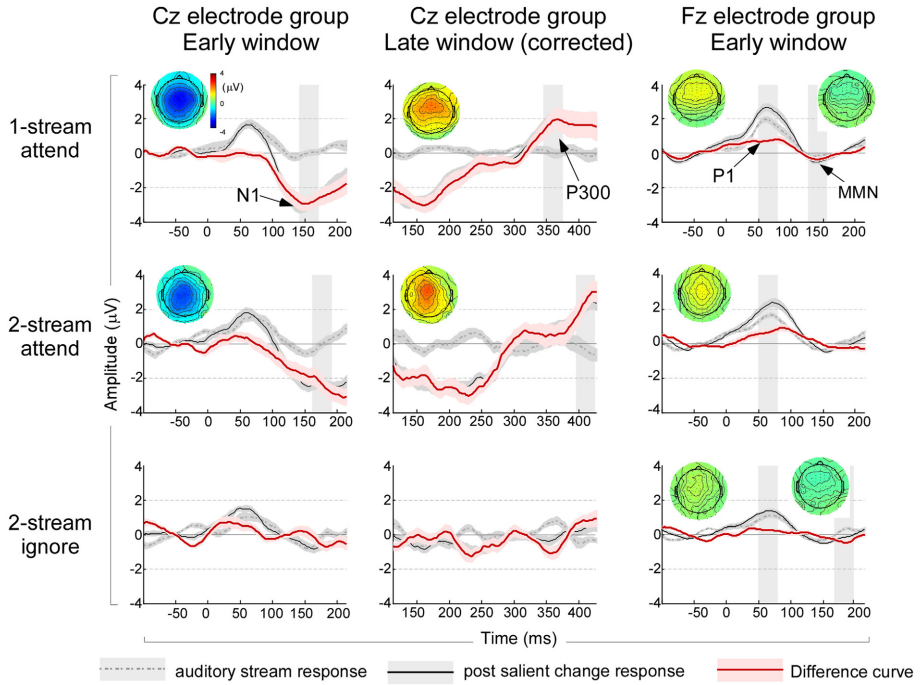


FIGURE 4 | Left column: ERP waveforms and topographies in the early window immediately after onset of the music note from the Cz electrode group. The waveforms are based on analysis of 1-note before (dashed gray line) and 1-note after (solid black line) a salient change, in addition to a difference curve (solid red line). The shaded curves indicate a standard error around the mean waveform. A shaded time interval indicates a 30 ms window where a statistically significant N1 peak is observed, along with the full head scalp topography of the N1 peak in the same analysis window. **Middle column:** ERP waveforms and topographies in the late window after note onset corrected for onset effects of the following note from the Cz electrode group. The analysis follows the same structure shown in the left column and time intervals and topographies indicate significant P300 peaks in a 30 ms time window. **Right column:** ERP waveforms and topographies immediately after onset from analysis of 3-notes before and after a salient change from the Fz electrode group. The first significant time interval indicates timing of the P1 peak (analyzed from the Cz electrode group—not shown); while the second time interval around 160 ms indicates the significant interval where an MMN-like component is observed in the Fz electrode group. Topographies shown correspond to the P1 and MMN intervals respectively. Note that non-significant intervals and topographies are not shown. All topographies and waveforms are kept to the same scale for ease of comparison of amplitudes.

in both cases was identical. In other words, listeners were presented with the same physical sound parameters (2 streams) while we directed their attentional focus toward or away from the music stimuli. In line with previous findings in the literature (Elhilali et al., 2009; Xiang et al., 2010), the attended auditory stream revealed a significant increase in the stimulus phase-locked response with attention relative to the ignore experiment [$F_{(1,\ 22)} = 6.115$, $p < 0.022$, $\eta_p^2 = 0.217$] in the univariate analysis. This result confirms an effect of neural enhancement caused by top-down attention. Once the salient change occurred in the music note stream, both conditions showed significant additional increase in ASSR energy but with no significant interaction between the two experiments and the increase ($p = 0.298$) in the Two-Way ANOVA test. The lack of *differential* change in the phase-locked response suggests that the bottom-up saliency caused by the notable change introduced ASSR energy increase that was independent in the stream notes of the top-down attentional state (**Figure 3B**). This result further corroborates our earlier finding that bottom-up attention induced by salient changes in an auditory stream is likely reflected in enhanced phase-locking of neural responses entrained by the stimulation rate, independent of the top-down attentional state of listeners.

We further examined the lateralization of ASSR of the auditory stream and the change of ASSR due to the salient change of the acoustic scene by averaging all left electrodes and right electrodes in the two-stream attend and ignore experiments. Prior to the salient change, there was a significant left lateralization of the ASSR response in the two-stream attend experiment [pairwise T-test; $t_{(11)} = 2.323$, $p < 0.040$], but no significant lateralization in the two-stream ignore experiment ($p = 0.140$). This result is consistent with previously reported left-hemisphere biases during selective attention in auditory and visual tasks (Coch et al., 2005; Bidet-Caulet et al., 2007). The change in ASSR amplitude before and after the salient change did not show any significant lateralization effects under both of the two-stream attend ($p = 0.141$) and two-stream ignore ($p = 0.310$) experiments suggesting a lack of lateralization during bottom-up attention.

We also examined the ERP components of the acoustic scene prior to a salient change under different attentional states, in the two-stream attend and two-stream ignore experiments. Consistent with the previous analysis, the N1 components were derived from 1-note analysis and P1 components were derived from 3-note analysis. Prior to any salient change in the musical note stream, the MANOVA test with attend vs. ignore conditions

as fixed factor and amplitudes of the two ERP components as dependent variables revealed a significant difference in P1 amplitude [$F_{(1,\ 22)} = 8.578$, $p < 0.008$, $\eta_p^2 = 0.281$], though no differences were observed in the N1 amplitude ($p = 0.198$). Once a salient change occurred in the stream, the Two-Way ANOVA with salient change and attend vs. ignore conditions as two factors on the four ERP components revealed further significant increase denoted by significant interactions between saliency and top-down attention in the P1 amplitude [$F_{(1,\ 22)} = 5.308$, $p < 0.031$, $\eta_p^2 = 0.194$], the N1 peak amplitude [$F_{(1,\ 22)} = 12.401$, $p < 0.002$, $\eta_p^2 = 0.360$] and the P300 component [$F_{(1,\ 22)} = 5.472$, $p < 0.029$, $\eta_p^2 = 0.199$]. No significant interaction were noted in the MMN-like component ($p = 0.528$). *Post-hoc T*-test comparisons of the amplitude changes from auditory stream to salient events under the attend and ignore conditions confirmed an enhanced saliency effect under top-down attention in the three aforementioned ERP components [P1: $t_{(22)} = 2.304$, $p < 0.031$; N1: $t_{(22)} = -3.521$, $p < 0.002$; P300: $t_{(22)} = 2.339$, $p < 0.029$; MMN-like: $t_{(22)} = 0.641$, $p = 0.528$]. This differential change between attend and ignore conditions hints to enhancement of bottom-up effects due to acoustic saliency that are mediated by top-down attention.

THE ACOUSTIC SCENE AND SALIENCY EFFECTS

Next, we explored the effect of the acoustic scene itself on the neural responses to the music notes, by comparing the one-stream attend and two-stream attend experiments. In both conditions, participants were attending to the same auditory stream and performing a similar task of salient change detection, but in presence (or absence) of a competing background stream. Attending to an auditory stream caused similar ASSR phase-locked responses, irrespective of whether the background contained a competing stream or not. The ASSR response was similar under both one-stream attend and two-stream attend conditions during the auditory stream ($p = 0.916$) in the univariate analysis. The introduction of a deviant change in the attended stream caused similar increases in the ASSR power under both one-stream and two-stream conditions ($p = 0.171$) shown in the Two-Way ANOVA test. The change in ASSR energy appeared to be marginally left lateralized in the one-stream attend experiment [$t_{(11)} = 1.805$, $p = 0.098$], but not in the two-stream attend experiment ($p = 0.141$). There was no significant difference between listening to one-stream and two-stream conditions in the N1 and P1 ERP responses (**Figure 4**) during the auditory stream in the MANOVA test. No significant interactions between saliency and scene complexity were found in the Two-Way ANOVA test on all four ERP components.

DISCUSSION

The current study focuses on how bottom-up, stimulus-driven salient changes in a dynamic scene modulate the neural representation of an auditory stream under different states of top-down attentional focus and with different complexities of the acoustic scene. Specifically, we report that a salient change evokes an increase in the phase-locked steady-state response to this salient event, irrespective of whether subjects attended to it or not. Given the complex nature of saliency changes in the current paradigm (across different acoustic dimensions defined along non-deterministic distributions), the observed ASSR changes likely reflect neural generators operating across larger neuronal groups or different neural centers (Escera et al., 2014); rather than mechanisms operating at the single neuron level such as stimulus-specific adaptation (SSA) which modulates the neural response to rare, unexpected or prominent events but is gated by the tuning properties of each neuron (Nelken, 2014). There was an observed left lateralization once attention is directed to the auditory stream (relative to the ignore condition), consistent with previous reports of a functional role of the left hemisphere in selective attention (Zani and Proverbio, 1995; Coch et al., 2005; Bidet-Caulet et al., 2007). An intriguing observation is the time-locked effect caused by the salient change response rather than a global, broad entrainment reflecting more general, non-specific enhancement in the neural response as previously reported with top-down attention (Picton and Hillyard, 1974a; Hari et al., 1980). Such specificity hints to a closer role of sensory cortex in coding the salience and dynamics of acoustic events with high degree of fidelity. The ASSR increase due to salient change in the scene appears to supplement further ASSR increase induced by top-down attention; which could itself be mediated by adaptive changes in the response characteristics of cortical neurons via mechanisms of neuronal plasticity (Fritz et al., 2007b). The current report of ASSR modulation with saliency of acoustic event is in agreement with earlier studies (Elhilali et al., 2009), but goes further in arguing for independent mechanisms underlying bottom-up induced ASSR changes and top-down attentional state. This bottom-up/top-down dissociation is further supported by the comparable size of ASSR increase which builds on top of increases boosted by purely top-down effects. Similar observations were made in the visual literature; whereby an increase in the steady-state visual evoked potential (SSVEP) is evoked by top-down attention; while the degree of change caused by different levels of saliency are generally the same under attend and non-attend states (Andersen et al., 2012).

An earlier report of changes in the ASSR energy due to salient event noted an opposite effect to the one reported here. A study by Rockstroh et al. (1996) found a reduction of ASSR power at 40 Hz after a deviant in a typical oddball paradigm, accompanied by a P300 component. A plausible explanation for the discrepancy between these two reports is that our definition of an auditory stream was characterized within tempi directly matched to time constants where auditory streaming is often observed, of the order of a few Hertz (Moore and Gockel, 2002). These are believed to have direct biophysical underpinnings matched to response dynamics of cortical neurons, whose selectivity to temporal rates is also of the order of a few Hertz (Miller et al., 2002; Liégeois-Chauvel et al., 2004). Rockstroh et al. (1996) examined deviance in a range where no reported cortical phase-locking exists, and is likely evoking different mechanisms than those involved in auditory stream formation and tracking.

An interesting aspect of the experimental paradigm used in the current study is that it did not follow a classic oddball paradigm, and was not defined by any repetitive tokens or specific sequence. Rather, the musical note stream of interest consisted of a series of complex tones (synthesized music instrument sound) with

varying acoustic parameters randomly following a restricted distribution along a given acoustic dimension, e.g., timbre, pitch, duration, or intensity. The auditory tokens used were dynamic but restricted to a range of parameters that likely groups them into a cohesive auditory stream. For instance, the pitch variations were confined within three semitone range, largely within natural contours expected in natural speech (Nooteboom, 1997) and also consistent with frequency variations tested in auditory streaming with varying frequency components (Bregman, 1990). Despite the fluctuating patterns in the stream, the ability of listeners to detect big deviations from its underlying statistical distribution strongly suggests that the auditory system is collecting statistics about the inherent variations in the stream, and either forming unconfined templates of the average distribution, representative statistics and parameters of the acoustic space or estimates of the underlying dynamical system of the stimulus that can be used to predict fluctuations in the scene (Sussman, 2007; D'Antona et al., 2013; Simpson et al., 2014).

Our analysis of event-related potentials reveals that salient change-induced ASSR changes are accompanied by attention-dependent modulation of the early P1, N1, and P300 components. Earlier reports corroborate an increase in the amplitude of the N1 component by involuntary attention (Näätänen et al., 1978). The observed change in both P1 and N1 peaks is likely reflecting the salient nature of the change introduced in the ongoing stream. It could mark the reset of the grouping process, a flag of aberrant events within the existing stream or initiation of a new auditory stream which does not fit within the expected fluctuations of the ongoing stream (Winkler et al., 2009). The differential effect of top-down attention on both P1 and N1 components with salient change implies a complex interaction of bottom-up deviance detection with top-down attention, though only the P1 component was modulated by purely top-down attention when comparing the attend and ignore cases prior to any salient change in the stream. Moreover, in agreement with most deviance detection studies, an MMN-like component was also observed with the salient change under attend and ignore conditions though not in the two-stream attend (Näätänen and Kreegipuu, 2012). This was accompanied by a late P300 component that was associated with discrimination in all three experiments (Martin et al., 1999; Martin and Stapells, 2005; Näätänen et al., 2007). It is important to note that dissociating the observed N1 and MMN-like component given the complexity of the task was non-trivial and mostly based on diverging scalp topographies for the two components. The observed mismatch negativity reported here was called "MMN-like" to guardedly reflect the peculiarity of this component in the current study. However, as previously noted by competing hypotheses of mismatch negativity generation, the distinction between MMN and N1 components always warrants caution and careful methodological strategies that are not easily attainable in the current design (Näätänen and Winkler, 1999; Garrido et al., 2009; May and Tiitinen, 2010; Lozano-Soldevilla et al., 2012).

Finally, the manifestation of a salient event was investigated as a function of the acoustic properties of the scene. A salient change occurring in an auditory stream that exists in isolation induces a notable MMN-like component that is not visible in presence of a competing stream under top-down attention, though both cases were associated with similar increases in the ASSR amplitude and involved comparable attentional loads as indicated by similar behavioral performances. The association between ASSR power increases and saliency is in agreement with effects reported earlier in both auditory and visual studies (Elhilali et al., 2009; Andersen et al., 2012), though earlier works have also observed a decrease of the P1 amplitude with delays of P1 latency (Billings et al., 2009) as well as other ERP components (Martin et al., 1999; Martin and Stapells, 2005; Lagemann et al., 2010) under a masking noise stream.

Overall, the current study reinforces the notion that auditory perception of complex acoustic scenes engages complex processes of statistical inference and dynamical tracking that define an auditory stream as an amorphous structure defined in a statistical sense. It is relative to this baseline that the system is able to flag any salient, deviant or unexpected events in a seemingly automatic, sensory-driven fashion. However, top-down attention reinforces the bottom-up saliency detection with mechanisms that are manifested mostly in amplitude of both early and late ERP components. Both forms of attention induce enhanced phase-locked responses to the ongoing rhythm of the auditory stream and may underlie the brain's ability to not only switch attention amongst simultaneous sounds streams, but also adjust its interpretation of the auditory stimulus depending on the acoustic parameters of the sensory input.

ACKNOWLEDGMENT

This work is supported by grants from the NSF CAREER IIS-0846112, AFOSR FA9550-09-1-0234, NIH 1R01AG036424-01, and ONR N000141010278 and N00014-12-1-0740.

REFERENCES

Andersen, S. K., Müller, M. M., and Martinovic, J. (2012). Bottom-Up biases in feagure-selective attention. *J. Neurosci.* 32, 16953–16958. doi: 10.1523/JNEUROSCI.1767-12.2012

Bidet-Caulet, A., Fischer, C., Besle, J., Aguera, P. E., Giard, M. H., and Bertrand, O. (2007). Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. *J. Neurosci.* 27, 9252–9292. doi: 10.1523/JNEUROSCI.1402-07.2007

Billings, C. J., Tremblay, K. L., Stecker, G. C., and Tolin, W. M. (2009). Human evoked cortical activity to signal-to-noise ratio and absolute signal level. *Hear. Res.* 254, 15–24. doi: 10.1016/j.heares.2009.04.002

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: The MIT Press.

Buschman, T. J., and Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science* 315:1860. doi: 10.1126/science.1138071

Capilla, A., Pazo-Alvarez, P., Darriba, A., Campo, R., and Gross, J. (2011). Steady-state visual evoked potentials can be explained by temporal superposition of transient event-related responses. *PLoS ONE* 6:e14543. doi: 10.1371/journal.pone.0014543

Chait, M., Ruff, C. C., Griffiths, T. D., and McAlpine, D. (2012). Cortical responses to changes in acoustic regularity are differentially modulated by attentional load. *Neuroimage* 59, 1932–1941. doi: 10.1016/j.neuroimage.2011.09.006

Chi, T., Gao, Y., Guyton, M. C., Ru, P., and Shamma, S. A. (1999). Spectro-temporal modulation transfer functions and speech intelligibility. *J. Acoust. Soc. Am.* 106, 2719–2732.

Coch, D., Sanders, L. D., and Neville, H. J. (2005). An event-related potential study of selective auditory attention in children and adults. *J. Cogn. Neurosci.* 17, 605–606. doi: 10.1162/0898929053467631

D'Antona, A. D., Perry, J. S., and Geisler, W. S. (2013). Humans make efficient use of natural image statistics when performing spatial interpolation. *J. Vis.* 13:11. doi: 10.1167/13.14.11

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Meth.* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Elhilali, M., Xiang, J., Shamma, S. A., and Simon, J. Z. (2009). Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biol.* 7:e1000129. doi: 10.1371/journal.pbio.1000129

Escera, C., Leung, S., and Grimm, S. (2014). Deviance detection based on regularity encoding along the auditory hierarchy: electrophysiological evidence in humans. *Brain Topogr.* 27, 527–538. doi: 10.1007/s10548-013-0328-4

Fritz, J. B., Elhilali, M., David, S. V., and Shamma, S. A. (2007a). Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in A1? *Hear. Res.* 229, 186–203. doi: 10.1016/j.heares.2007.01.009

Fritz, J. B., Elhilali, M., David, S. V., and Shamma, S. A. (2007b). Auditory attention–focusing the searchlight on sound. *Curr. Opin. Neurobiol.* 17, 437–455. doi: 10.1016/j.conb.2007.07.011

Garrido, M. I., Kilner, J. M., Stephan, K. E., and Friston, K. J. (2009). The mismatch negativity: a review of underlying mechanisms. *Clin. Neurophysiol.* 120:453. doi: 10.1016/j.clinph.2008.11.029

Goto, M. (2004). "Development of the RWC music database," in *Proceedings of the 18th International Congress on Acoustics (ICA 2004)* (Kyoto), I-552–I-556.

Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R. (2003). "RWC music database: music genre database and musical instrument sound database," in *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2003)* (Baltimore, MD), 229–230.

Gutschalk, A., Micheyl, C., Melcher, J. R., Rupp, A., Scherg, M., and Oxenham, A. J. (2005). Neuromagnetic correlates of streaming in human auditory cortex. *J. Neurosci.* 25, 5382–5388. doi: 10.1523/JNEUROSCI.0347-05.2005

Hari, R., Aittoniemi, K., Jarvinen, M. L., Katila, T., and Varpula, T. (1980). Auditory evoked transient and sustained magnetic fields of the human brain. localization of neural generators. *Exp. Brain Res.* 40, 237–240. doi: 10.1007/BF00237543

Katsuki, F., and Constantinidis, C. (2014). Bottom-Up and top-down attention: different processes and overlapping neural systems. *Neuroscientist*. doi: 10.1177/1073858413514136. (in press).

Kiefer, M. (2007). Top-down modulation of unconscious "automatic" processes: a gating framework. *Acv. Cogn. Psychol*. 3, 289–306. doi: 10.2478/v10053-008-0032-2

Kim, H. (2014). Involvement of the dorsal and ventral attention networks in oddball stimulus processing: a meta-analysis. *Hum. Brain Mapp.* 35, 2265–2284. doi: 10.1002/hbm.22326

Lagemann, L., Okamoto, H., Teismann, H., and Pantev, C. (2010). Bottom-up driven involuntary attention modulates auditory signal in noise processing. *BMC Neurosci.* 11:156. doi: 10.1186/1471-2202-11-156

Liégeois-Chauvel, C., Lorenzi, C., Trebuchon, A., Regis, J., and Chauvel, P. (2004). Temporal envelope processing in the human left and right auditory cortices. *Cereb. Cortex* 14, 731–740. doi: 10.1093/cercor/bhh033

Lozano-Soldevilla, D., Marco-Pallares, J., Fuentemilla, L., and Grau, C. (2012). Common N1 and mismatch negativity neural evoked components are revealed by independent component model-based clustering analysis. *Psychophysiology* 49, 1454–1463. doi: 10.1111/j.1469-8986.2012.01458.x

Luck, S. J., Woodman, G. F., and Vogel, E. K. (2000). Event-related potential studies of attention. *Trends Cogn. Sci.* 4, 432–440. doi: 10.1016/S1364-6613(00)01545-X

Martin, B. A., Kurtzberg, D., and Stapells, D. R. (1999). The effects of decreased audibility produced by high-pass noise masking on N1 and the mismatch negativity to speech sounds /ba/ and /da/. *J. Speech Lang. Hear. Res.* 42, 271–286.

Martin, B. A., and Stapells, D. R. (2005). Effects of low-pass noise masking on auditory event-related potentials to speech. *Ear Hear.* 26, 195–213. doi: 10.1097/00003446-200504000-00007

May, P. J., and Tiitinen, H. (2010). Mismatch negativity (MMN), the deviance-elicited auditory deflection, explained. *Psychophysiology* 47, 66–122. doi: 10.1111/j.1469-8986.2009.00856.x

Miller, L. M., Escabí, M. A., Read, H. L., and Schreiner, C. E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *J. Neurophysiol.* 87, 516–527. doi: 10.1152/jn.00395.2001

Moore, B. C. J., and Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acustica* 88, 320–333.

Müller, N., Schlee, W., Hartmann, T., Lorenz, I., and Weisz, N. (2009). Top-down modulation of the auditory steady-state response in a task-switch paradigm. *Front. Human Neurosci.* 3:1. doi: 10.3389/neuro.09.001.2009

Näätänen, R. (1992). *Attention and Brain Function.* Hillsdale, NJ: Lawrence Erlbaum, 307–313.

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

Näätänen, R., and Kreegipuu, K. (2012). "The mismatch negativity (MMN)," in *The Oxford Handbook of Event-Related Potential Components*, eds S. J. Luck and E. Kappenman (New York, NY: Oxford University Press), 143–158.

Näätänen, R., and Michie, P. T. (1979). Early selective attention effects on the evoked potential. a critical review and reinterpretation. *Biol. Psychol.* 8, 81–136. doi: 10.1016/0301-0511(79)90053-X

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., and Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. *Psychol. Bull.* 125, 826–859. doi: 10.1037/0033-2909.125.6.826

Nelken, I. (2014). Stimulus-specific adaptation and deviance detection in the auditory system: experiments and models. *Biol. Cybern.* doi: 10.1007/s00422-014-0585-7. [Epub ahead of print]

Nooteboom, S. (1997). "The prosody of speech: melody and rhythm," in *The Handbook of Phonetic Sciences*, eds W. J. Hardcastle and J. Laver (Oxford: Blackwell), 640–673.

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011:156869. doi: 10.1155/2011/156869

Picton, T. W., and Hillyard, S. A. (1974a). Human auditory evoked potentials. I. the nature of the response. *Electroencephalogr. Clin. Neurophysiol.* 36, 186–197.

Picton, T. W., and Hillyard, S. A. (1974b). Human auditory evoked potentials. II. effects of attention. *Electroencephalogr. Clin. Neurophysiol.* 36, 191–199. doi: 10.1016/0013-4694(74)90156-4

Pinto, Y., van der Leji, A. R., Sligte, I. G., Lamme, V. A. F., and Scholte, H. S. (2013). Bottom-up and top-down attention are independent. *J. Vis.* 13, 1–14. doi: 10.1167/13.3.16

Rockstroh, B., Müller, M., Heinz, A., Wagner, M., Berg, P., and Elbert, T. (1996). Modulation of auditory responses during oddball tasks. *Biol. Psychol.* 43, 41–55. doi: 10.1016/0301-0511(95)05175-9

Simpson, A. J., Harper, N. S., Reiss, J. D., and McAlpine, D. (2014). Selective adaptation to oddball sounds by the human auditory system. *J. Neurosci.* 34, 1963–1969. doi: 10.1523/JNEUROSCI.4274-13.2013

Sussman, E. S. (2007). A new view on the MMN and attention debate. *J. Psychophysiol.* 21, 164–175. doi: 10.1027/0269-8803.21.34.164

Sussman, E. S., and Steinschneider, M. (2001). Attention modifies sound level detection in young children. *Dev. Cogn. Neurosci.* 1, 351–360. doi: 10.1016/j.dcn.2011.01.003

Winkler, I., Denham, S., 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

Xiang, J., Simon, J., and Elhilali, M. (2010). Competing streams at the cocktail party: exploring the mechanisms of attention and temporal integration. *J. Neurosci.* 30, 12084–12093. doi: 10.1523/JNEUROSCI.0827-10.2010

Zani, A., and Proverbio, A. M. (1995). ERP signs of early selective attention effects to check size. *Electroencephalogr. Clin. Neurophysiol.* 95, 277–292. doi: 10.1016/0013-4694(95)00078-D

Zhang, L., Peng, W., Zhang, Z., and Hu, L. (2013). Distinct features of auditory steady-state responses as compared to transient event-related potentials. *PLoS ONE* 8:e69164. doi: 10.1371/journal.pone.0069164

Zhang, X., Zhaoping, L., Zhou, T., and Fang, F. (2012). Neural activities in V1 create a bottom-up saliency map. *Neuron* 73, 183–192. doi: 10.1016/j.neuron.2011.10.035

Zhaoping, L. (2008). Attention capture by eye of origin singletons even without awareness—a hallmark of a bottom-up saliency map in the primary visual cortex. *J. Vis.* 8, 1–18. doi: 10.1167/8.5.1

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

Received: 02 February 2014; accepted: 27 June 2014; published online: 21 July 2014.

*Citation: Shuai L and Elhilali M (2014) Task-dependent neural representations of salient events in dynamic auditory scenes. Front. Neurosci. **8**:203. doi: 10.3389/fnins.2014.00203*

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

frontiers in NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 12 September 2014
doi: 10.3389/fnins.2014.00277

Auditory stream segregation using bandpass noises: evidence from event-related potentials

Yingjiu Nie[1]*, Yang Zhang[2,3] and Peggy B. Nelson[2]

[1] *Department of Communication Sciences and Disorders, James Madison University, Harrisonburg, VA, USA*
[2] *Department of Speech-Language-Hearing Sciences, University of Minnesota, Twin-Cities, MN, USA*
[3] *Center for Neurobehavioral Development, University of Minnesota, Twin-Cities, MN, USA*

Edited by:
Susann Deike, Leibniz Institute for Neurobiology, Germany

Reviewed by:
Elyse S. Sussman, Albert Einstein College of Medicine, USA
Alexandra Bendixen, Carl von Ossietzky University of Oldenburg, Germany

****Correspondence:***
Yingjiu Nie, Department of Communication Science and Disorders, James Madison University, 801 Carrier Drive-MSC 4304, Harrisonburg, VA 22807, USA
e-mail: nieyx@jmu.edu

The current study measured neural responses to investigate auditory stream segregation of noise stimuli with or without clear spectral contrast. Sequences of alternating A and B noise bursts were presented to elicit stream segregation in normal-hearing listeners. The successive B bursts in each sequence maintained an equal amount of temporal separation with manipulations introduced on the last stimulus. The last B burst was either delayed for 50% of the sequences or not delayed for the other 50%. The A bursts were jittered in between every two adjacent B bursts. To study the effects of spectral separation on streaming, the A and B bursts were further manipulated by using either bandpass-filtered noises widely spaced in center frequency or broadband noises. Event-related potentials (ERPs) to the last B bursts were analyzed to compare the neural responses to the delay vs. no-delay trials in both passive and attentive listening conditions. In the passive listening condition, a trend for a possible late mismatch negativity (MMN) or late discriminative negativity (LDN) response was observed only when the A and B bursts were spectrally separate, suggesting that spectral separation in the A and B burst sequences could be conducive to stream segregation at the pre-attentive level. In the attentive condition, a P300 response was consistently elicited regardless of whether there was spectral separation between the A and B bursts, indicating the facilitative role of voluntary attention in stream segregation. The results suggest that reliable ERP measures can be used as indirect indicators for auditory stream segregation in conditions of weak spectral contrast. These findings have important implications for cochlear implant (CI) studies—as spectral information available through a CI device or simulation is substantially degraded, it may require more attention to achieve stream segregation.

Keywords: attention, auditory stream segregation, bandpass noise, MMN, P300, P3b, spectral separation, temporal pattern

INTRODUCTION

Auditory stream segregation (also referred to as auditory streaming) is an auditory process that occurs naturally in daily life. When listening to a talker in a party or following a melody played by an instrument in an orchestra, listeners with normal hearing interpret the mixture of sounds in such a way that sounds from different sources are allocated to individual sound generators. Research has demonstrated that auditory stream segregation may operate on various physical properties, such as the sound spectrum (Bregman and Campbell, 1971; Warren and Obusek, 1972; van Noorden, 1975; Dannenbring and Bregman, 1976a,b; Bregman et al., 1999) and the temporal envelopes (Singh and Bregman, 1997; Vliegen et al., 1999; Vliegen and Oxenham, 1999; Grimault et al., 2000, 2001, 2002; Roberts et al., 2002). Behavioral (van Noorden, 1975; Botte et al., 1997; Brochard et al., 1999) and neurophysiological (Sussman et al., 1998a, 2005; Sussman and Steinschneider, 2009) studies have further indicated that listeners' voluntary attention facilitates stream segregation.

Behavioral laboratory studies (e.g., van Noorden, 1975, for a review, see Moore and Gockel, 2002) have indicated that frequency separation and stimulus presentation rate are critical for the formation of auditory streams. The identification of the temporal coherence boundary (TCB) and the fission boundary (FB) (van Noorden, 1975) are amongst the earlier suggestions of ways in which voluntary attention may influence stream segregation. In the van Noorden study, when a frequency separation of two potential tonal streams was larger than the TCB or smaller than the FB, a listener would perceive two streams and one stream, respectively, regardless of their focused attention. When the frequency separation fell in between TCB and FB, offering an ambiguous cue for segregation/integration, a listener could hold either the integrated or the segregated perception depending on directed attention. Brochard et al. (1999) further investigated the role of attention by presenting interleaved subsequences (streams) of tones to normal-hearing listeners. They evaluated attentional effort for stream segregation by measuring the threshold of a temporal offset of a given tone in a focused subsequence (stream)

for a listener to detect an irregular rhythm in that subsequence (stream). Their findings showed that more attentional effort was needed for stream segregation when the frequency separation between the subsequences was smaller.

The effect of voluntary attention in auditory stream segregation has also been studied using neurophysiological methods. One important measure was the mismatch negativity (MMN) response. The MMN is typically elicited by automatic change detection in a passive listening oddball paradigm, in which a frequent auditory stimulus (i.e., the standard) is repeatedly presented, while an infrequent stimulus (i.e., the deviant) occasionally replaces the standard (Ford and Hillyard, 1981; Nordby et al., 1988a,b). The MMN is topographically represented by a negative centro-frontal scalp distribution in the temporal window of 100–300 ms post the onset of the change, reflecting pre-attentive detection of the deviant irrespective of attentional efforts (for reviews, see Näätänen, 1995; Näätänen et al., 2007). The MMN can also be elicited during attentive listening (for reviews, see Picton et al., 2000a; Näätänen et al., 2007), but tends to overlap with a negative component N2b that also shows a negative centro-frontal scalp distribution (Näätänen et al., 1982; Novak et al., 1990, for a review, see Folstein and Van Petten, 2008). Previous studies have demonstrated that the MMN can be measured as an indirect index of stream segregation in passive listening when the concurrent auditory streams are sufficiently different in frequency (Sussman et al., 1998a, 1999, 2001; Winkler et al., 2001, 2003; Yabe et al., 2001). For example, Sussman et al. (1999) presented normal-hearing listeners with standard stimulus sequences comprising two interleaved subsequences (streams) of tones. Deviants were inserted into either subsequence. In the unattended condition, MMNs to the deviants were elicited only when the tones were presented at a fast rate wherein stream segregation was induced. Studies have further demonstrated that the MMN measure could reflect the facilitative role of voluntary attention in stream segregation (Sussman et al., 1998a, 2005).

Another important neurophysiological measure to examine the role of voluntary attention in stream segregation is the P300 response, which is known to index attentional shift to novelty detection (Sutton et al., 1965). While P3a (an earlier component of the P300 family with a frontal distribution) reflects obligatory processes (e.g., involuntary attention shift) in the passive or attentive listening condition, the P3b with a posterior parietal distribution is associated with voluntary attentional orientation to the deviants in the attentive listening condition (Knight and Nakada, 1998; Knight and Scabini, 1998; Corbetta and Shulman, 2002), or with classifying initially uncategorized events into a discrete group (Friedman et al., 2001). Sussman and Steinschneider (2009) studied how two sets of tones with various frequency separations were segregated into different auditory streams or integrated into one stream in adults and children. They assessed P3b in the attentive condition and found that in adults, higher P3b amplitudes were associated with larger inter-stream frequency separation, which presumably led to better performance in stream segregation.

Overall, the published ERP data have shown that the MMN and P300 measures can reflect segregation success or failure along with behavioral tests. In cases of children with hearing loss, and especially children with cochlear implants (CIs), ERP measures of stream segregation could serve as indirect indicators of signal-processing or rehabilitative program success. However, previous ERP studies have only used tonal stimuli with clear spectral separation between the auditory streams. The current investigation extended this line of research by examining whether the MMN and P300 measures could be reliably elicited for auditory stream segregation based on noise stimuli with or without clear spectral contrast. The revelation of normal-hearing listeners' ability to pre-attentively segregate the two noise streams may offer a baseline against which future research on CI users can be compared.

Behavioral studies on normal-hearing adult listeners showed that bandpass-filtered noises with separated (Dannenbring and Bregman, 1976b) and overlapped spectra (Bregman et al., 1999) could be perceived as from different auditory streams. The bandpass-filtered noise stimuli may simulate the stimulations with less salient spectral differences through a CI electrode array, whose users only demonstrate 8–10 effective auditory filters (Fu et al., 1998; Friesen et al., 2001; Nelson et al., 2003). However, it remains inconclusive how CI users perform stream segregation based on input with degraded spectral information (Chatterjee et al., 2006; Hong and Turner, 2006; Cooper and Roberts, 2009). While Hong and Turner (2006) showed CI users could segregate streams of pure tones at different frequencies, Cooper and Roberts (2009) argued that they were unable to segregate streams based on electrode-separation which represents spectral separation in the electrical stimulation. Chatterjee et al. (2006) noted that only one out of five CI subjects showed definitive electrode-separation based stream segregation. Most recently, the work of Böckmann-Barthel et al. (2014) suggested that CI users were able to segregate different streams of tones adequately differing in frequency with a processing time course comparable to normal-hearing listeners. Studying how the reduced spectral separation operates for stream segregation at various processing stages (i.e., pre-attentive and attentive) in normal-hearing listeners may provide a foundation to better understand stream segregation based on spectral separation in CI users.

One of the speech cues that is well preserved for CI users is rhythm (e.g., McDermott, 2004). A recent behavioral study (Micheyl and Oxenham, 2010) has indicated that the steady rhythm of a tone series can be a cue for listeners to segregate the tone series from another. In that study, reiterated triplets of tones were presented in an ABA pattern with A tones quasi-randomly occurring temporally and B tones occurring at a fixed B-to-B interval. The listeners were able to identify the temporal displacement of the last B tone to some extent even when A and B tones were set to the same frequency. Thus, detecting the deviant B-to-B interval at the end of a sequence would require the perception of the organized B tones into one stream on the basis of the built-in temporal pattern.

The current study followed Micheyl and Oxenham (2010) in the use of a deviant in a rhythmic stream to assess the stream segregation process. Unlike the previous study, sequences of two interleaved subsequences of noise bursts, namely A and B bursts, were presented. While the temporal position of each A burst was quasi-randomized between the two adjacent B bursts, the B

bursts were presented steadily (i.e., with a constant B-to-B onset-to-onset interval) except that the last B burst was delayed (i.e., presented with a longer B-to-B onset-to-onset interval), in half of the stimulus sequences. It was anticipated that, if the steadily presented B bursts were perceived as one stream distinct from the stream comprising the unsteadily-presented A bursts, a delayed B burst would be perceived as a deviant breaking the steady rhythm of the B-to-B gaps. In contrast, subjects may have also integrated the A and B bursts as one stream consisting of arrhythmic elements. The jittered timing of A bursts would introduce uncertainty to an A-to-B gap, making an A-to-B gap an ineffective cue to detect a delayed B burst. Therefore, the detection of the deviant in the rhythmic stream can be used as an indirect indicator for stream segregation.

The goal of the current study was to examine neural correlates of stream segregation for stimuli with weak spectral information in passive listening as well as in attentive listening conditions. We were specifically interested in addressing three questions. First, would the noise-based stimulus paradigm yield measurable neurophysiological responses to the delay of the last B bursts to indirectly index stream segregation? Second, is the selected spectral separation between two sets of bandpass-filtered noise, simulating a moderate electrode-separation through a CI processor, necessary to show differences in stream segregation at the pre-attentive level? Third, does voluntary attention facilitate stream segregation for the noise stimuli? The MMN responses in a passive listening condition were measured to indirectly index pre-attentive stream segregation, and the P3b responses in an attentive listening condition were measured to indirectly index stream segregation with involvement of voluntary attention.

MATERIALS AND METHODS

SUBJECTS

Nine right-handed adult listeners (5 females and 4 males; 19–37 years old) participated in the study. Their hearing thresholds were no greater than 20 dB HL at audiometric frequencies of 250, 500, 1000, 1500, 2000, 3000, 4000, 6000, and 8000 Hz on the right side. The research procedure was approved by the Institutional Review Board at the University of Minnesota for experiments with human subjects. Each subject gave written informed consent.

STIMULI AND TEST PROCEDURE

Sequences of 12 pairs of A and B noise bursts were presented (**Figure 1**). The duration of an A or B burst was 80 ms including 8-ms rise/fall time. The B burst sequences maintained an onset-to-onset interval of 340 ms except for the last (12th) B burst whose onset was either delayed or not delayed. In the delayed sequences, the 12th B bursts were delayed from their nominal temporal positions by 30 ms. In the no-delay sequences, the 12th B bursts were presented at the nominal temporal location. The total duration was 3.1 s for the delayed sequences and 3.07 s for the no-delay sequences. The A bursts (expect the first one) were pseudo-randomly placed between two successive B bursts. The stimulus-onset asynchrony (SOA)—defined as an interval between the onsets of two consecutive bursts (i.e., the onsets of an A burst and its adjacent B bursts, or the onsets of a B burst and its adjacent A bursts)—was assigned with a nominal mean value of 130 ms. The real SOAs varied between 90 and 170 ms due to the jittering of the A bursts. Specifically, the onset of any given non-initial A burst was advanced or delayed by an amount of time ranging from 0 to 40 ms. The amount of jitter was determined based on a pilot study.

Manipulation of spectral separation between the burst sequences used two types of noises, broadband noise and bandpass filtered noise. For the broadband noise stimuli with no spectral separation (no-SSEP), an independent Gaussian noise was created for each element of a given stimulus sequence for the A and B burst sequences. To introduce spectral separation (SSEP), the A bursts were generated with a 4th-order Butterworth filter (cutoff frequencies at 2149 and 7000 Hz), and a bandpass filter between 200 and 1426 Hz was utilized to generate the B bursts. The slope of the filters was set at 12 dB/octave to resemble the shallow filter slope in CI users (Anderson et al., 2011). Altogether, the bandpass filtering procedure resulted in a 2.86-octave distance between the center frequencies of the A and B bursts with a 0.59-octave distance between the lower edge frequency of A bursts and the higher edge frequency of the B bursts.

Listeners were seated comfortably in a chair in an acoustically-attenuated and electrically-isolated chamber (ETS-Lindgren Acoustic Systems). Stimulus presentation used the EEvoke software (ANT Inc., The Netherlands). The sounds were presented through an insert earphone (Etymotic Research ER-3A) monaurally to the right ear. The sound level was set at 60 dB above the subject's hearing threshold for a 1000 Hz sine wave tone.

Both passive listening and attentive listening conditions were administered for the two types of stimulus setup (no-SSEP and SSEP), which resulted in four stimulus blocks lasting approximately 2 hours. All subjects were presented the same stimuli; the order of the stimulus blocks was counterbalanced across the subjects. The subjects repeated the same four stimulus blocks on a different day to allow sufficient number of trials for ERP data analysis. In the passive listening condition, subjects were directed to ignore the acoustic stimuli while watching a muted movie with subtitles. In the attentive listening condition, the subjects were instructed to respond only to the delayed sequences by pushing a key on a computer keyboard. Each stimulus block contained 120 independently generated stimulus sequences with 60 delayed and 60 no-delay sequences arranged in a random order. The offset-to-onset inter-stimulus-sequence interval was randomized between 900 and 1000 ms. The inter-block break interval was 5–10 min.

Prior to the ERP experiments, all subjects had participated in a behavioral experiment with the same stimulus paradigm and more spectral separations between A and B bursts for 10 days spread over 1–2 months for a total of approximately 15 hours of listening task to detect the delayed 12th B bursts[1].

[1]It would be of interest to learn the amount of time needed for training to reach a stable performance level prior to the neurophysiological experiment using the current stimulus paradigm. However, this experiment and the behavioral study prior to it did not provide sufficient data to address this question. The subjects in the previous behavioral study had completed the experiments before the question of training arose, and thus this project did not address the training effect over time.

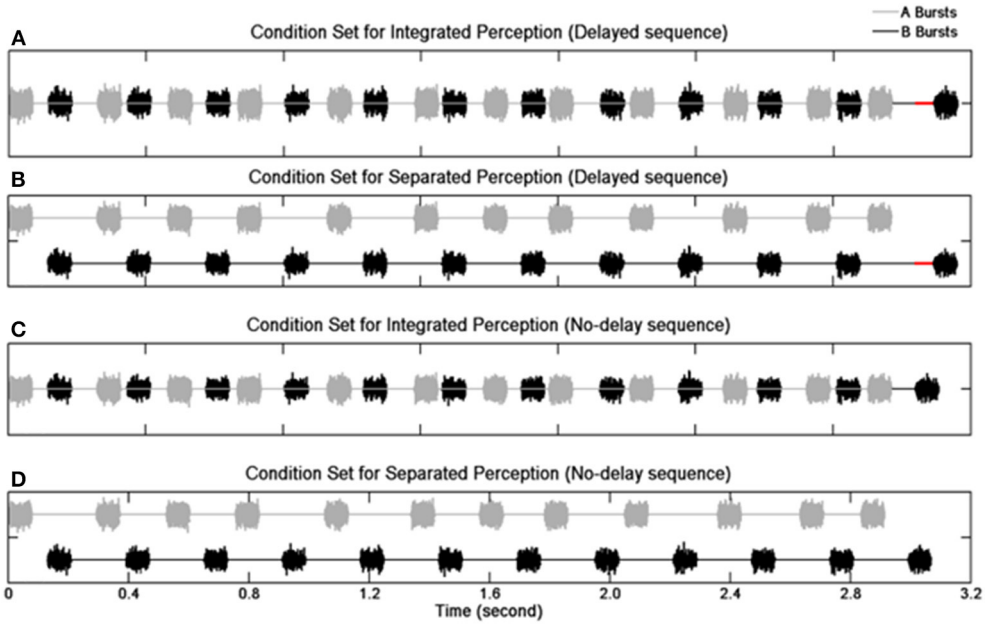


FIGURE 1 | Schematic illustration of the stimulus paradigm. Waveforms of noise bursts are plotted. Panels **(A,B)** show the delayed sequences with the red solid lines indicating the duration of the delay for the 12th B burst. Panels **(C,D)** show the no-delay sequences. Panels **(A,C)** depict the condition when individual sequence elements result in an integrated perception, and panels **(B,D)** depict the condition when individual elements result in a segregated perception.

EEG RECORDING AND ANALYSIS

Continuous EEG was recorded (bandwidth = 0.016–200 Hz; sampling rate = 512 Hz) using the ASA-Lab system with a REFA-72 amplifier (TMS International BV) and a 64-channel WaveGuard cap (ANT Inc., The Netherlands). The ground was positioned at AFz, and the default reference for the REFA-72 amplifier was the common average of all connected unipolar electrode inputs. Impedances of individual electrodes were at or below 5 kΩ.

ERP averaging was performed offline in BESA (Version 6.0, MEGIS Software GmbH, Germany) following recommended guidelines (Picton et al., 2000b; DeBoer et al., 2007). The automated EOG correction algorithm in BESA was applied to the EEG data with the threshold parameters at 150 μV for HEOG and 250 μV for VEOG. The ERP epoch length was 700 ms, including a pre-stimulus baseline of 100 ms. The ERP data were bandpass filtered at 0.53–40 Hz and re-referenced to the average of the recordings at the mastoid channels. Trials with potentials exceeding ±50 μV were rejected. For each subject, the ERP waveforms recorded in the two separate sessions for the same condition were first analyzed individually and then combined with weighted averaging in BESA. Unweighted averaging was calculated for the grand mean at the subject group level.

The standard stimuli were the ending (i.e., 12th) B bursts of the no-delay sequences, and the deviant stimuli were those ending B bursts of the delayed sequences. Difference ERP waves were obtained by subtracting the standard ERPs from the deviant ERPs in each of the four conditions (i.e., no-SSEP passive listening, SSEP passive listening, no-SSEP attentive listening, and SSEP attentive listening). In the passive listening condition, the individual subjects had the total number of accepted trials in the range of 101–118 for either the delayed or no-delay stimulus sequences. In the attentive listening condition, the trials with a behaviorally incorrect response were excluded from the ERP analysis; the total number of correct responses was in the range of 59–114 for the SSEP stimuli, and the correct responses were in the range of 39–101 for the no-SSEP stimuli. The counts of false alarms were 0–19 for the SSEP stimuli and 4–46 for the no-SSEP stimuli. Given the insufficient number of trials, no ERP analysis was performed on the false-alarm trials for the no-delay 12th B bursts.

Global field power (GFP) was calculated by computing the standard deviation of the amplitude data across the 64 electrodes at each sampling point as an unbiased estimate independent of electrode selection (Lehmann and Skrandies, 1980; Hamburger and Burgt, 1991). The time windows for ERP component analysis were selected based on the grand-mean GFP plots and scalp topography maps. The ERP analysis windows were confirmed by visual inspection of the grand-mean ERP waveform overlay plots using single channels. Temporal windows of 200, 50, and 50 ms around the GFP peaks for P3b, MMN, and MMN/N2b responses were chosen, respectively. Specifically, for the attentive listening conditions, the windows were 244–444 ms for the SSEP stimuli and 300–500 ms for the no-SSEP stimuli to assess the P3b component, respectively. For the passive listening conditions, the window of 280–330 ms was selected for both SSEP and no-SSEP stimuli. To assess the MMN/N2b component in attentive listening, the windows were 97–147 ms for the SSEP stimuli and 174–224 ms for the no-SSEP stimuli. The CPz electrode was used for P3b analysis, and the Fz was used for MMN analysis.

Given the small amplitude of the N2b response, six centro-frontal channels (FC1, FC2, FCz, C1, C2, and Cz) were grouped in our analysis. For each subject, mean amplitude values in the selected time windows for the ERP components were calculated for the deviant-to-standard difference waveforms under the four conditions (i.e., two spectral contrasts by two attentional conditions) (Luck, 2005; Clayson et al., 2013).

One-tailed one-sample *t*-tests were conducted on these average amplitudes independently under each of the four conditions. The criterion for statistical significance was divided by six (i.e., $\alpha = 0.0083$) to correct for the number (i.e., six) of comparisons. A paired two-tailed *t*-test was carried out to compare the average amplitudes obtained for the SSEP stimuli against those obtained for the no-SSEP stimuli for the attentive listening condition.

To assess the strength of the ERP activities relative to the baseline independent of electrode selection, *z*-scores were calculated with Bonferroni correction for the post-stimulus GFP at each sampling point (Rao et al., 2010; Zhang et al., 2011). Sustained latency intervals of at least 20 ms or longer (Rao et al., 2010; Zhang et al., 2011) were highlighted where *z*-scores indicated the GFP was significant.

ANALYSIS ON BEHAVIORAL DATA DURING ATTENTIVE LISTENING

For each subject, the hit rate for the delayed sequences and the false alarm rate for the no-delay sequences were combined across the two recording sessions for the no-SSEP and SSEP stimuli. The behavioral scores were then converted to d' scores (Macmillan and Creelman, 2005). A paired two-tailed *t*-test was performed to compare the d' scores between the SSEP and no-SSEP stimulus conditions.

CORRELATIONAL ANALYSIS FOR P3b RESPONSE AND BEHAVIORAL PERFORMANCE

A *post-hoc* Pearson correlation was conducted for the P3b measure (separate tests for amplitude and latency) and the behavioral d' measure in the attentive listening condition. Considering the small number of subjects in our study, we pooled the SSEP and no-SSEP stimulus conditions together.

RESULTS

P3b IN THE ATTENTIVE CONDITION

Significant P3b responses were observed for both the SSEP and no-SSEP stimuli in the attentive listening condition (**Table 1** and **Figure 2**). Grand-mean ERP waveforms at the CPz channel and topographic maps showed a strong positive posterior parietal distribution, which peaked earlier for the SSEP stimuli (at 345 ms) than for the no-SSEP stimuli (at 417 ms) (**Figure 3**). But there was no significant difference in the P3b amplitude between the SSEP and no-SSEP stimulus conditions [$t_{(1,\ 8)} = 0.89$, $p = 0.395$].

Sequential topographic maps exhibited a positive potential maximum moving from the frontal area to the posterior parietal area in the time window starting from around 255 ms post stimulus to around 355 ms for the SSEP stimuli, and from 364 to 434 ms for the no-SSEP stimuli (**Figure 4**). The earlier frontal positive distributions indicated a possible occurrence of the P3a component before the occurrence of the posterior-parietal P3b for both SSEP and no-SSEP stimuli.

MMN/N2b COMPONENT IN THE ATTENTIVE CONDITION

Visual inspection of the grand-mean ERP difference waveforms (**Figure 5**) indicated the presence of a small MMN/N2b component at the centro-frontal sites preceding the P3b response during attentive listening. However, this N2b component was not significantly different from the zero baseline in either the SSEP or the no-SSEP stimulus condition (**Table 1**).

MMN IN THE PASSIVE CONDITION

Both GFP data and Fz data showed no presence of MMN during passive listening when there was no spectral separation between the A and B burst sequences (i.e., no-SSEP stimulus condition) (**Figures 2** and **6**). In the SSEP stimulus condition, however, a discrepancy was observed between the GFP waveform analysis and the Fz electrode waveform analysis on the MMN data. The GFP data showed significant MMN activities in the time windows of 209–241 and 290–362 ms for the SSEP stimuli during

Table 1 | Mean amplitudes of the ERP components calculated from the difference waveforms (i.e., delayed minus no-delay) for the 12th B bursts.

	Attentive condition (P3b)		Attentive condition (MMN/N2b)		Passive condition (MMN)	
	SSEP	no-SSEP	SSEP	no-SSEP	SSEP	no-SSEP
Average amplitude of difference waveforms (μV)	1.70 (0.90)	1.25 (1.45)	−0.44 (1.30)	−0.16 (1.25)	−0.85 (1.60)	0.45 (1.41)
t-value (one-tailed one-sample *t*-test for the difference waveforms)	$t_{(1,\ 8)} = 5.34$	$t_{(1,\ 8)} = 2.93$	$t_{(1,\ 8)} = -1.01$	$t_{(1,\ 8)} = -0.39$	$t_{(1,\ 8)} = -1.59$	$t_{(1,\ 8)} = 0.89$
p-value (statistical significance level $\alpha < 0.0083$)	0.0003	0.0022	0.1707	0.3534	0.0749	0.2005

Standard deviations are shown in the parentheses. The t- and p-values were calculated from one-tailed one-sample t-tests comparing the amplitudes of the ERP component with the zero baseline.

passive listening (**Figure 2**). Based on the GFP data, the MMN peak latency in the SSEP stimulus condition was found to be at 315 ms. In contrast, the MMN amplitudes at the Fz channel failed to reach statistical significance for the SSEP stimuli (**Figure 6**). Inspection of the individual data showed that one subject had a mismatch positivity response at Fz, which could have contributed to the discrepancy between the GFP and MMN statistical results.

BEHAVIORAL DATA

Subjects showed greater d' values for the SSEP stimuli than for the no-SSEP stimuli [$t_{(1,\ 8)} = 5.18,\ p < 0.001$] (**Table 2**). Thus, the presence of spectral contrast facilitated the detection of the delayed 12th B bursts in the SSEP stimulus condition.

Table 2 | Hit rates, false alarm rates, and *d'*-values calculated from subjects' behavioral responses when attempting to identify the delayed 12th B bursts during attentive listening.

	SSEP		no-SSEP	
	Range	Average (SD)	Range	Average (SD)
Hit rate (%)	49–95	74 (15)	32–84	65 (15)
False alarm rate (%)	0–16	5 (5)	3–38	17 (11)
d'	1.46–3.33	2.56 (0.62)	0.57–2.29	1.45 (0.58)

BRAIN-BEHAVIORAL CORRELATION DURING ATTENTIVE LISTENING

In the attentive listening condition, a significant correlation was found between the P3b latency values and the d' scores (**Figure 7**). Higher d' scores for detecting the change in the final position of the B sequences were associated with earlier P3b responses ($r = -0.549$, $p < 0.05$). However, the d' scores were not significantly correlated with the P3b amplitude data.

DISCUSSION

Overall, results in both attentive and passive listening conditions confirm that ERP measures may be reliable indirect indicators of stream segregation based on less distinctive spectral separation cues. In the passive condition, significant MMN activities were observed in the GFP data for the SSEP stimuli but not for the no-SSEP stimuli, indicating that spectral separation between A and B bursts was necessary for stream segregation at the pre-attentive level. In the attentive condition, spectral separation between the A and B bursts contributed to better performance in stream segregation. Conscious perception of stream segregation was indirectly indicated by the presence of significant P3b activities for both SSEP and no-SSEP stimuli. Moreover, better behavioral performance was correlated with earlier P3b peak response.

STIMULUS PARADIGM AND EFFECT OF RHYTHM ON STREAM SEGREGATION

The first goal of the current study was to investigate whether the stimulus paradigm used in our behavioral study (Nie and Nelson, 2010) could elicit reliable ERP measures to

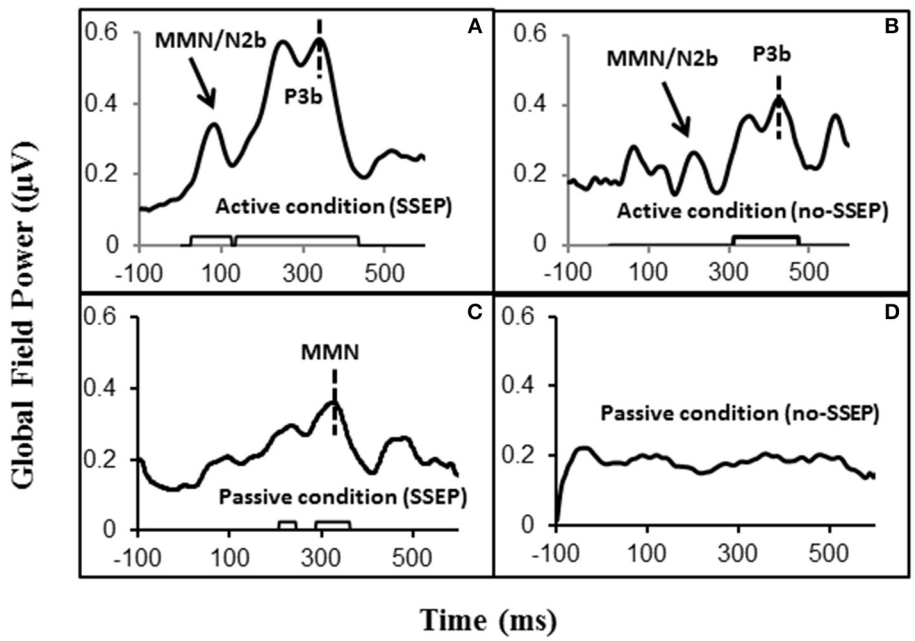


FIGURE 2 | Global field power (GFP) data obtained from the grand-mean ERP deviant-to-standard difference waveforms. Panels **(A,B)** respectively show GFPs for the SSEP and no-SSEP stimuli during attentive listening. Panels **(C,D)** show GFPs for the SSEP and no-SSEP stimuli during passive listening. The bars along the x-axes in **(A–C)** show significant activities as determined from *z*-scores relative to the 100-ms pre-stimulus baseline. Panel **(D)** shows an absence of significant GFP activities. The dashed vertical lines in **(A,B)** mark the GFP peaks of P3b. The dashed vertical line in **(C)** indicates the GFP peak of MMN for the SSEP stimuli. The GFP peaks falling in the MMN/N2b time window are indicated for the SSEP **(A)** and no-SSEP stimuli **(B)** by the arrows.

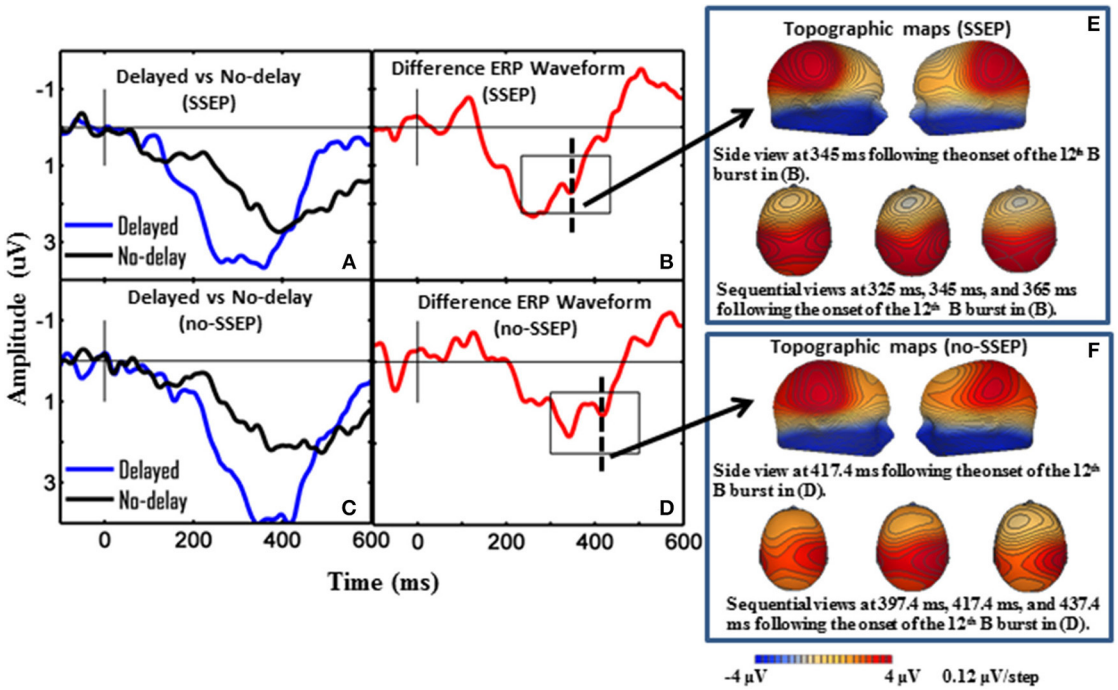


FIGURE 3 | Grand-mean ERP waveforms and grand-mean ERP deviant-to-standard difference waveforms from CPz. Panels **(A–D),** respectively show attentive listening for the SSEP stimuli, attentive listening for the no-SSEP stimuli, passive listening for the SSEP stimuli, and passive listening for the no-SSEP stimuli. The topographic maps for the P3b peaks are shown in **(E)** (SSEP stimuli) and panel **(F)** (no-SSEP stimuli). The solid vertical lines depict the onsets of the 12th B bursts. The thick dashed vertical lines in **(B,D)** mark the P3b peaks. The boxes surrounding the peaks in **(B,D)** show the analysis windows for the P3b activity.

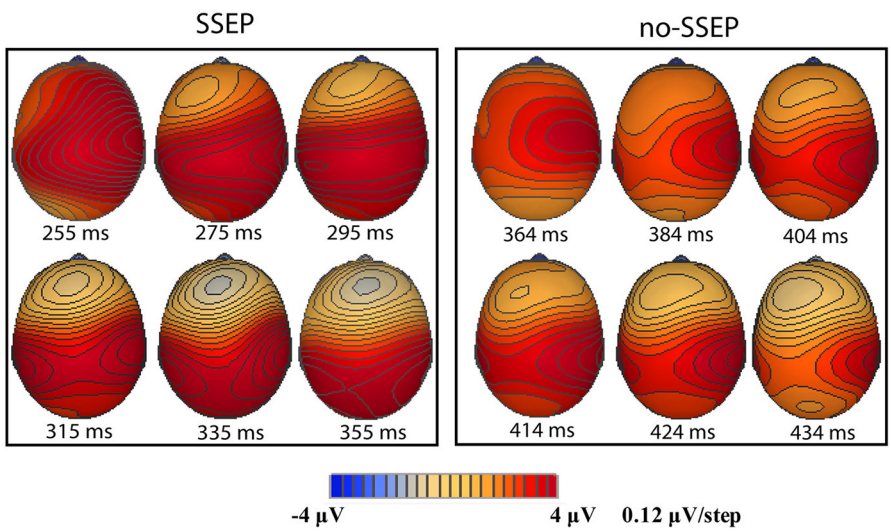


FIGURE 4 | Sequential topographic maps illustrating a frontal-to-parietal shift in the positive potential maximum at selected time points post the onset of the 12th B bursts for the SSEP (left panel) and no-SSEP (right panel) stimuli, respectively.

indirectly index stream segregation with noise bursts. Similar paradigms with tonal stimuli have been used in behavioral psychophysical studies on stream segregation (e.g., Roberts et al., 2002; Hong and Turner, 2006; Micheyl and Oxenham, 2010). Our behavioral data showed feasibility of this paradigm with noise stimuli, and our ERP data here further demonstrated that a clear P3b response and a late MMN response could be respectively, elicited for the attentive and passive listening conditions in the perceptual/cognitive processing of the temporally displaced element using such a paradigm. It should be

noted that the ERP measures (i.e., MMN, P3b, MMN/N2b, and P3a components) reported here are indirect indicators for stream segregation that are specific to the stimulus paradigm. The presence of the ERP components signifies the pre-attentive or conscious detection of the delayed 12th B bursts, which is facilitated by segregation of the A and B sequences based on temporal pattern, spectral separation as well as voluntary attention. We speculate that the stimulus paradigm involving noise bursts (as opposed to tonal stimuli) and rhythmic B sequences intermixed with arrhythmic A sequences may have jointly affected the latency of the MMN responses in individual subjects and increased the inter-subject variability during passive listening.

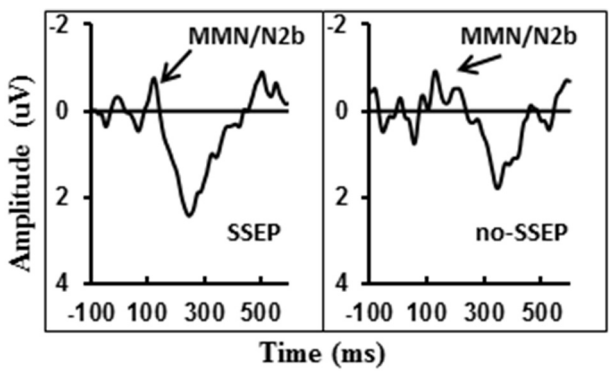


FIGURE 5 | Grand-mean ERP deviant-to-standard difference waveforms grouped across six centro-frontal channels (FC1, FC2, FCz, C1, C2, and Cz) in attentive listening for SSEP (left panel) and no-SSEP (right panel) stimuli. The MMN/N2b-like peaks are indicated by the arrows.

Our ERP results reveal that the rhythm cue itself is not adequate to generate stream segregation pre-attentively; however, when listeners focus attention on following the rhythm, they may be able to utilize this cue to form stream segregation. In the no-SSEP stimulus condition, the sole cue available for listeners to segregate the streams of A and B bursts was the steady rhythm in the B stream. The presence of P3b for the both SSEP and no-SSEP stimuli suggests that the steady rhythm of B bursts can contribute to stream segregation in addition to spectral separation. This is consistent with the weak segregation effect reported by Micheyl and Oxenham (2010) for ABA tone sequences that differed in rhythm but not in frequency.

Together, the results suggest that the current stimulus paradigm can be a potential tool for future ERP studies on stream segregation in normal as well as clinical populations with degraded auditory signal. We would like to stress that the benefit of this paradigm may be limited by the substantial amount of time (approximately 15 hours) spent on the training/behavioral experiment prior to the ERP study. Further study of the effects of training on ERP results may help address this limitation.

EFFECT OF SPECTRAL SEPARATION ON STREAM SEGREGATION WITH BANDPASS NOISE STIMULI

The second goal of the study was to investigate whether weak (noisy) spectral separation cues using bandpass noises in the current design could generate pre-attentive stream segregation. A series of studies using tonal stimuli (e.g., Sussman et al., 1998b, 1999, 2005) have systematically demonstrated that MMN could be used as an indirect index for stream segregation providing the stimulus paradigm is suitably designed.

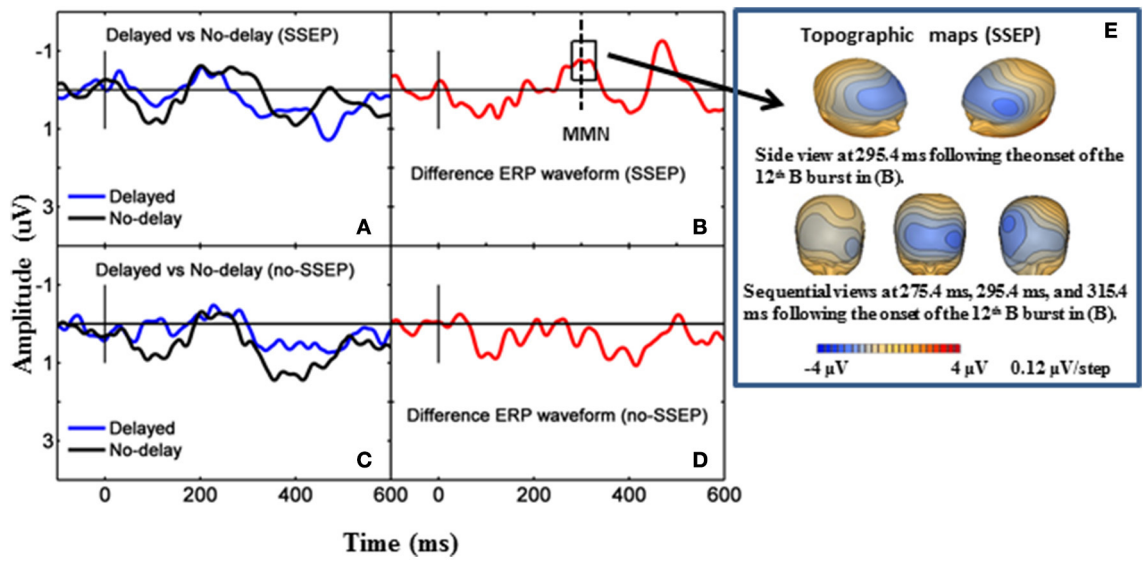


FIGURE 6 | Grand-mean ERP waveforms and grand-mean ERP deviant-to-standard difference waveforms from the Fz electrode. Panels **(A,B)** stand for passive listening in the SSEP stimulus condition, and panels **(C,D)** stand for passive listening in the no-SSEP stimulus condition. The topographic maps for the MMN peaks are shown in **(E)** for the SSEP stimuli. The solid vertical lines mark the onsets of the 12th bursts. The dashed vertical line in **(B)** mark the MMN peak. The box around the peak in **(B)** represents the analysis window for the MMN.

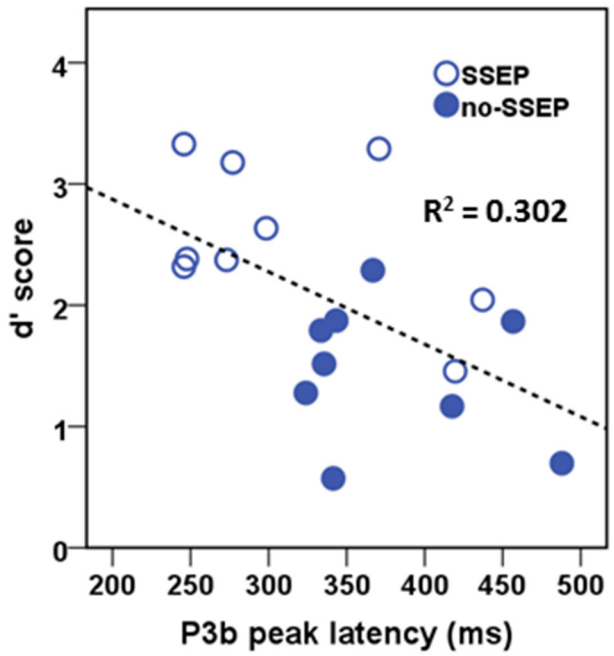


FIGURE 7 | Correlational scatterplot illustrating the relationship between behavioral *d′* score and the P3b peak latency in attentive listening. The data were pooled across the SSEP and no-SSEP stimulus conditions.

The absence of MMN in the no-SSEP stimulus condition in our study indicated the need for spectral separation in addition to the temporal rhythmic cue for the auditory streaming of the noise sequences at the pre-attentive level. In the SSEP stimulus condition, the presence of MMN in response to the delayed 12th B bursts may not have been a result of integrating A and B bursts into one stream. If the A and B bursts were integrated, a stronger integration would be anticipated for the no-SSEP stimuli, which could presumably lead to a stronger MMN to deviants for the no-SSEP stimuli than for the SSEP stimuli. The current findings are contrary to this prediction. Therefore, the MMN elicitation in the delayed sequences would depend on the pre-attentive detection of rhythmic disruption in the B sequences (the deviant B-to-B gaps in contrast to the standard B-to-B gaps stored in the auditory memory), which was aided by the presence of spectral separation between the A and B sequences.

The elicitation of the late MMN activity for the SSEP stimuli suggests that the listeners were able to extract the temporal patterns in the B sequence from the intermixed A-B series. With an SOA of 130 ms, the amount of frequency separation between the two alternating bandpass noises can be allocated into segregated streams at the pre-attentive auditory processing stage. Specifically, the center frequencies of the A and B bursts in the SSEP stimulus condition were separated by 2.86 octaves. This amount of frequency separation would be an unambiguous cue to generate stream segregation for tonal stimuli with a SOA between 90 and 130 ms (van Noorden, 1975; Sussman et al., 1999, 2005). However, the frequency separation between the two sets of bandpass noises was effectively reduced as a result of the 0.59-octave difference between the edge frequencies of the A and B bursts in addition to the use of a shallow filtering slope of 12 dB/octave. Even with the degraded frequency separation of the bandpass-filtered stimuli, there was evidence for pre-attentive stream segregation in the current paradigm.

EFFECT OF VOLUNTARY ATTENTION ON STREAM SEGREGATION

The third goal was to examine the effect of voluntary attention on stream segregation. Due to different ERP components analyzed for the passive and attentive listening conditions, no statistical comparisons were performed between the ERP measures obtained in the two listening conditions in the current study. Nevertheless, consistent with previous studies (Sussman et al., 1998a, 2005), voluntary attention has been shown to facilitate stream segregation as P3b was consistently elicited for both SSEP and no-SSEP stimuli during attentive listening in our study. In particular, the absence of MMN during passive listening and the presence of P3b during attentive listening for the no-SSEP stimuli clearly demonstrate that when there is no spectral separation between the A and B noise sequences, attentive listening is necessary to allow stream segregation. Despite the lack of spectral cues, voluntarily directing attention toward the built-in temporal patterns allowed listeners to segregate the two auditory streams.

The P3b response has been demonstrated to reflect uncertainty resolution involving the integrative processing of stimulus evaluation and response execution (Verleger, 1997, 2010; Dien et al., 2004). Previous research has shown that shorter P3b latencies and larger P3b amplitudes tend to be associated with greater stimulus salience (e.g., Sussman and Steinschneider, 2009). Despite a strong effect of spectral separation (SSEP vs. no-SSEP) in the behavioral data, our P3b amplitude data did not show significant differences between the SSEP and no-SSEP stimulus conditions. Neither was the P3b amplitude correlated with behavioral accuracy. Thus, the P3b amplitude might not be a proper measure to assess the strength of stream segregation with the current experimental design. In contrast, the P3b latency measure was found to be significantly correlated with behavioral performance in the *post-hoc* analysis. Spectral separation allows increased certainty accompanied by greater behavioral accuracy in evaluating the B sequences with either delay or no delay in the last noise burst. As our study did not incorporate experimental manipulations to differentiate contributions of stimulus evaluation and response selection to P3b activity, further studies would be necessary to explore how P3b amplitude and latency may be related to behavioral accuracy and response times under different processing strategies (e.g., requiring listeners to prioritize response accuracy over speed or vice versa).

TECHNICAL CONSIDERATIONS ON DISCREPANCIES IN MMN RESULTS

The GFP data and ERP waveform data (at Fz) did not yield consistent results in the MMN analysis for the SSEP stimuli during passive listening. As GFP calculation uses all electrodes and ERP waveform analysis uses only selected electrodes, some minor differences are not unexpected (Miller and Zhang, 2014). The GFP data in our study clearly indicated the occurrence of significant

neural activities within the typical MMN time window (100–300 ms post-stimulus) followed by a larger late response. The topographic maps further indicated a frontal distribution with a negative polarity, and the peak MMN activity as shown in GFP occurred at 315 ms after the onset of the delayed 12th B burst (**Figure 2C**), which was equivalent to 345 ms after the expected onset of a no-delay 12th B burst. Our GFP data are consistent with previous MMN studies on the extraction of complex tone patterns or abstract regularities/rules, which also have reported a similar response pattern with the regular MMN followed by an additional late MMN (Korpilahti et al., 1995; Zachau et al., 2005). Late MMN responses have also been reported in adult subjects for studies involving linguistic stimuli (Hill et al., 2004; Zhang et al., 2005, 2009; Takegata et al., 2008). This late negativity is sometimes referred to as late discriminative negativity (LDN), which has been found mainly in MMN studies on children using linguistic stimuli (Ceponiene et al., 1998, 2003, 2005; Korpilahti et al., 2001; Maurer et al., 2003). Some researchers suggest that the late MMN activity may be an index of attentional reorienting or auditory learning and development (Putkinen et al., 2012) and that it diminishes in the course of child development (Ceponiene et al., 2003).

Unlike the GFP results, the waveform data analysis at the Fz site failed to show the elicitation of significant MMNs at the subject group level. Inspection of the individual data indicated that one subject showed a positive mismatch response (p-MMR). The p-MMR data point was found to be a statistical outlier in terms of its *z*-score, and its inclusion in the group level analysis affected the statistical outcome. The existence of the p-MMR data point only affected the waveform analysis but not the GFP analysis because GFP calculation relies on the absolute amount of deviation irrespective of polarity. The p-MMR responses in the passive listening oddball paradigm have been reported in a number of child studies (e.g., Maurer et al., 2003; Shafer et al., 2010). As there has been no systematic report about adult p-MMR, it is unclear what might have caused the occurrence of one aberrant p-MMR data point in our study. We suspect that it could be related to attentional processing in this individual subject.

IMPLICATIONS FOR COCHLEAR IMPLANT USERS

Results from the current study have important implications for the investigation of stream segregation in CI users. The spectral separation used in our SSEP stimuli corresponds to the frequency regions of the lower four channels and the upper three channels of a simulated 8-channel CI (Fu and Nogaki, 2005). With such a large frequency difference, stimulation through CIs may lead to pre-attentive stream segregation. Furthermore, when the frequency cue is limited, temporal rhythm may be an important cue for CI users to form stream segregation, which can be facilitated by voluntary attention.

ACKNOWLEDGMENTS

This work was partially supported by the Doctoral Dissertation Fellowship from the University of Minnesota and funding from the Department of CSD at James Madison University to Yingjiu Nie, the CLA startup fund and Brain Imaging Research Project Award from the University of Minnesota to Yang Zhang, and, the NIH grant No. DC 008306 to Peggy B. Nelson. The authors are indebted to Sharon Miller for her help with graphing for the revision. We are appreciative to the two reviewers for their insights and helpful suggestions.

REFERENCES

Anderson, E. S., Nelson, D. A., Kreft, H., Nelson, P. B., and Oxenham, A. J. (2011). Comparing spatial tuning curves, spectral ripple resolution, and speech perception in cochlear implant users. *J. Acoust. Soc. Am.* 130, 364–375. doi: 10.1121/1.3589255

Böckmann-Barthel, M., Deike, S., Brechmann, A., Ziese, M., and Verhey, J. L. (2014). Time course of auditory streaming: do CI users differ from normal-hearing listeners? *Front. Psychol.* 5:775. doi: 10.3389/fpsyg.2014.00775

Botte, M. C., Drake, C., Brochard, R., and McAdams, S. (1997). Perceptual attenuation of nonfocused auditory streams. *Percept. Psychophys.* 59, 419–425. doi: 10.3758/BF03211908

Bregman, A. S., and Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *J. Exp. Psychol.* 89, 244–249. doi: 10.1037/h0031163

Bregman, A. S., Colantonio, C., and Ahad, P. A. (1999). Is a common grouping mechanism involved in the phenomena of illusory continuity and stream segregation? *Percept. Psychophys.* 61, 195–205. doi: 10.3758/BF03206882

Brochard, R., Drake, C., Botte, M. C., and McAdams, S. (1999). Perceptual organization of complex auditory sequences: effect of number of simultaneous subsequences and frequency separation. *J. Exp. Psychol. Hum. Percept. Perform.* 25, 1742–1759. doi: 10.1037/0096-1523.25.6.1742

Ceponiene, R., Alku, P., Westerfield, M., Torki, M., and Townsend, J. (2005). ERPs differentiate syllable and nonphonetic sound processing in children and adults. *Psychophysiology* 42, 391–406. doi: 10.1111/j.1469-8986.2005.00305.x

Ceponiene, R., Cheour, M., and Naatanen, R. (1998). Interstimulus interval and auditory event-related potentials in children: evidence for multiple generators. *Electroencephalogr. Clin. Neurophysiol.* 108, 345–354. doi: 10.1016/S0168-5597(97)00081-6

Ceponiene, R., Lepisto, T., Alku, P., Aro, H., and Naatanen, R. (2003). Event-related potential indices of auditory vowel processing in 3-year-old children. *Clin. Neurophysiol.* 114, 652–661. doi: 10.1016/S1388-2457(02)00436-4

Chatterjee, M., Sarampalis, A., and Oba, S. I. (2006). Auditory stream segregation with cochlear implants: a preliminary report. *Hear. Res.* 222, 100–107. doi: 10.1016/j.heares.2006.09.001

Clayson, P. E., Baldwin, S. A., and Larson, M. J. (2013). How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study. *Psychophysiology* 50, 174–186. doi: 10.1111/psyp.12001

Cooper, H. R., and Roberts, B. (2009). Auditory stream segregation in cochlear implant listeners: measures based on temporal discrimination and interleaved melody recognition. *J. Acoust. Soc. Am.* 126, 1975–1987. doi: 10.1121/1.3203210

Corbetta, M., and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3, 201–215. doi: 10.1038/nrn755

Dannenbring, G. L., and Bregman, A. S. (1976a). Effect of silence between tones on auditory stream segregation. *J. Acoust. Soc. Am.* 59, 987–989. doi: 10.1121/1.380925

Dannenbring, G. L., and Bregman, A. S. (1976b). Stream segregation and the illusion of overlap. *J. Exp. Psychol. Hum. Percept. Perform.* 2, 544–555. doi: 10.1037/0096-1523.2.4.544

DeBoer, T., Scott, L. S., and Nelson, C. A. (2007). "Methods for acquiring and analyzing infant event-related potentials," in *Infant EEG and Event-Related Potentials,* ed M. D. Haan (Hove: Psychology Press), 5–37.

Dien, J., Spencer, K. M., and Donchin, E. (2004). Parsing the late positive complex: mental chronometry and the ERP components that inhabit the neighborhood of the P300. *Psychophysiology* 41, 665–678. doi: 10.1111/j.1469-8986.2004.00193.x

Folstein, J. R., and Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: a review. *Psychophysiology* 45, 152–170. doi: 10.1111/j.1469-8986.2007.00602.x

Ford, J. M., and Hillyard, S. A. (1981). Event-related potentials (ERPs) to interruptions of a steady rhythm. *Psychophysiology* 18, 322–330. doi: 10.1111/j.1469-8986.1981.tb03043.x

Friedman, D., Cycowicz, Y. M., and Gaeta, H. (2001). The novelty P3: an event-related brain potential (ERP) sign of the brain's evaluation of novelty. *Neurosci. Biobehav. Rev.* 25, 355–373. doi: 10.1016/S0149-7634(01)00019-7

Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. *J. Acoust. Soc. Am.* 110, 1150. doi: 10.1121/1.1381538

Fu, Q. J., and Nogaki, G. (2005). Noise susceptibility of cochlear implant users: the role of spectral resolution and smearing. *J. Assoc. Res. Otolaryngol.* 6, 19–27. doi: 10.1007/s10162-004-5024-3

Fu, Q. J., Shannon, R. V., and Wang, X. (1998). Effects of noise and spectral resolution on vowel and consonant recognition: acoustic and electric hearing. *J. Acoust. Soc. Am.* 104, 3586–3596. doi: 10.1121/1.423941

Grimault, N., Bacon, S. P., and Micheyl, C. (2002). Auditory stream segregation on the basis of amplitude-modulation rate. *J. Acoust. Soc. Am.* 111, 1340. doi: 10.1121/1.1452740

Grimault, N., Micheyl, C., Carlyon, R. P., Arthaud, P., and Collet, L. (2000). Influence of peripheral resolvability on the perceptual segregation of harmonic complex tones differing in fundamental frequency. *J. Acoust. Soc. Am.* 108, 263–271. doi: 10.1121/1.429462

Grimault, N., Micheyl, C., Carlyon, R. P., Arthaud, P., and Collet, L. (2001). Perceptual auditory stream segregation of sequences of complex sounds in subjects with normal and impaired hearing. *Br. J. Audiol.* 35, 173–182.

Hamburger, H., and Burgt, M. G. (1991). Global field power measurement versus classical method in the determination of the latency of evoked potential components. *Brain Topogr.* 3, 391–396. doi: 10.1007/BF01129642

Hill, P. R., McArthur, G. M., and Bishop, D. V. (2004). Phonological categorization of vowels: a mismatch negativity study. *Neuroreport* 15, 2195–2199. doi: 10.1097/00001756-200410050-00010

Hong, R. S., and Turner, C. W. (2006). Pure-tone auditory stream segregation and speech perception in noise in cochlear implant recipients. *J. Acoust. Soc. Am.* 120, 360. doi: 10.1121/1.2204450

Knight, R. T., and Nakada, T. (1998). Cortico-limbic circuits and novelty: a review of EEG and blood flow data. *Rev. Neurosci.* 9, 57–70. doi: 10.1515/REVNEURO.1998.9.1.57

Knight, R. T., and Scabini, D. (1998). Anatomic bases of event-related potentials and their relationship to novelty detection in humans. *J. Clin. Neurophysiol.* 15, 3–13. doi: 10.1097/00004691-199801000-00003

Korpilahti, P., Krause, C. M., Holopainen, I., and Lang, A. H. (2001). Early and late mismatch negativity elicited by words and speech-like stimuli in children. *Brain Lang.* 76, 332–339. doi: 10.1006/brln.2000.2426

Korpilahti, P., Lang, H., and Aaltonen, O. (1995). Is there a late-latency mismatch negativity (MMN) component? *Electroencephalogr. Clin. Neurophysiol.* 95, P96. doi: 10.1016/0013-4694(95)90016-G

Lehmann, D., and Skrandies, W. (1980). Reference-free identification of components of checkerboard-evoked multichannel potential fields. *Electroencephalogr. Clin. Neurophysiol.* 48, 609–621. doi: 10.1016/0013-4694(80)90419-8

Luck, S. J. (2005). *An Introduction to the Event-Related Potential Technique*. Cambridge, MA: MIT Press.

Macmillan, N. A., and Creelman, C. D. (2005). *Detection Theory: A User's Guide*. Mahwah, NJ: Lawrence Erlbaum Associates.

Maurer, U., Bucher, K., Brem, S., and Brandeis, D. (2003). Development of the automatic mismatch response: from frontal positivity in kindergarten children to the mismatch negativity. *Clin. Neurophysiol.* 114, 808–817. doi: 10.1016/S1388-2457(03)00032-4

McDermott, H. J. (2004). Music perception with cochlear implants: a review. *Trends Amplif.* 8, 49–82. doi: 10.1177/108471380400800203

Micheyl, C., and Oxenham, A. J. (2010). Objective and subjective psychophysical measures of auditory stream integration and segregation. *J. Assoc. Res. Otolaryngol.* 11, 709–724. doi: 10.1007/s10162-010-0227-2

Miller, S., and Zhang, Y. (2014). Neural coding of phonemic fricative contrast with and without hearing aid. *Ear Hear.* 35:0000000000000025. doi: 10.1097/AUD.0000000000000025

Moore, B. C. J., and Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acust. United Acust.* 88, 320–333.

Näätänen, R. (1995). The mismatch negativity: a powerful tool for cognitive neuroscience. *Ear Hear.* 16, 6–18. doi: 10.1097/00003446-199502000-00002

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., Simpson, M., and Loveless, N. E. (1982). Stimulus deviance and evoked potentials. *Biol. Psychol.* 14, 53–98. doi: 10.1016/0301-0511(82)90017-5

Nelson, P. B., Jin, S.-H., Carney, A. E., and Nelson, D. A. (2003). Understanding speech in modulated interference: cochlear implant users and normal-hearing listeners. *J. Acoust. Soc. Am.* 113, 961. doi: 10.1121/1.1531983

Nie, Y., and Nelson, P. (2010). Auditory stream segregation using amplitude modulated vocoder bandpass noise. *J. Acoust. Soc. Am.* 127, 1809–1809. doi: 10.1121/1.3384104

Nordby, H., Roth, W. T., and Pfefferbaum, A. (1988a). Event-related potentials to breaks in sequences of alternating pitches or interstimulus intervals. *Psychophysiology* 25, 262–268. doi: 10.1111/j.1469-8986.1988.tb01239.x

Nordby, H., Roth, W. T., and Pfefferbaum, A. (1988b). Event-related potentials to time-deviant and pitch-deviant tones. *Psychophysiology* 25, 249–261. doi: 10.1111/j.1469-8986.1988.tb01238.x

Novak, G. P., Ritter, W., Vaughan, H. G. Jr., and Wiznitzer, M. L. (1990). Differentiation of negative event-related potentials in an auditory discrimination task. *Electroencephalogr. Clin. Neurophysiol.* 75, 255–275. doi: 10.1016/0013-4694(90)90105-S

Picton, T. W., Alain, C., Otten, L., Ritter, W., and Achim, A. (2000a). Mismatch negativity: different water in the same river. *Audiol. Neurootol.* 5, 111–139. doi: 10.1159/000013875

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R. Jr., et al. (2000b). Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. *Psychophysiology* 37, 127–152. doi: 10.1111/1469-8986.3720127

Putkinen, V., Niinikuru, R., Lipsanen, J., Tervaniemi, M., and Huotilainen, M. (2012). Fast measurement of auditory event-related potential profiles in 2-3-year-olds. *Dev. Neuropsychol.* 37, 51–75. doi: 10.1080/87565641.2011.615873

Rao, A., Zhang, Y., and Miller, S. (2010). Selective listening of concurrent auditory stimuli: an event-related potential study. *Hear. Res.* 268, 123–132. doi: 10.1016/j.heares.2010.05.013

Roberts, B., Glasberg, B. R., and Moore, B. C. J. (2002). Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. *J. Acoust. Soc. Am.* 112, 2074. doi: 10.1121/1.1508784

Shafer, V. L., Yu, Y. H., and Datta, H. (2010). Maturation of speech discrimination in 4- to 7-yr-old children as indexed by event-related potential mismatch responses. *Ear Hear* 31, 735–745. doi: 10.1097/AUD.0b013e3181e5d1a7

Singh, P. G., and Bregman, A. S. (1997). The influence of different timbre attributes on the perceptual segregation of complex-tone sequences. *J. Acoust. Soc. Am.* 102, 1943–1952. doi: 10.1121/1.419688

Sussman, E., Ceponiene, R., Shestakova, A., Näätänen, R., and Winkler, I. (2001). Auditory stream segregation processes operate similarly in school-aged children and adults. *Hear. Res.* 153, 108–114. doi: 10.1016/S0378-5955(00)00261-6

Sussman, E., Ritter, W., and Vaughan, H. G. Jr. (1998a). Attention affects the organization of auditory input associated with the mismatch negativity system. *Brain Res.* 789, 130–138. doi: 10.1016/S0006-8993(97)01443-1

Sussman, E., Ritter, W., and Vaughan, H. G. Jr. (1998b). Predictability of stimulus deviance and the mismatch negativity. *Neuroreport* 9, 4167–4170. doi: 10.1097/00001756-199812210-00031

Sussman, E., Ritter, W., and Vaughan, H. G. (1999). An investigation of the auditory streaming effect using event-related brain potentials. *Psychophysiology* 36, 22–34. doi: 10.1017/S0048577299971056

Sussman, E. S., Bregman, A. S., Wang, W. J., and Khan, F. J. (2005). Attentional modulation of electrophysiological activity in auditory cortex for unattended sounds within multistream auditory environments. *Cogn. Affect. Behav. Neurosci.* 5, 93–110. doi: 10.3758/CABN.5.1.93

Sussman, E., and Steinschneider, M. (2009). Attention effects on auditory scene analysis in children. *Neuropsychologia* 47, 771–785. doi: 10.1016/j.neuropsychologia.2008.12.007

Sutton, S., Braren, M., Zubin, J., and John, E. R. (1965). Evoked-potential correlates of stimulus uncertainty. *Science* 150, 1187–1188. doi: 10.1126/science.150.3700.1187

Takegata, R., Tervaniemi, M., Alku, P., Ylinen, S., and Näätänen, R. (2008). Parameter-specific modulation of the mismatch negativity to duration decrement and increment: evidence for asymmetric processes. *Clin. Neurophysiol.* 119, 1515–1523. doi: 10.1016/j.clinph.2008.03.025

van Noorden, L. (1975). *Temporal Coherence in the Perception of Tone Sequences*. Unpublished Ph.D. dissertation.

Verleger, R. (1997). On the utility of P3 latency as an index of mental chronometry. *Psychophysiology* 34, 131–156. doi: 10.1111/j.1469-8986.1997.tb02125.x

Verleger, R. (2010). Popper and P300: can the view ever be falsified that P3 latency is a specific indicator of stimulus evaluation? *Clin. Neurophysiol.* 121, 1371–1372. doi: 10.1016/j.clinph.2010.01.038

Vliegen, J., Moore, B. C., and Oxenham, A. J. (1999). The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. *J. Acoust. Soc. Am.* 106, 938–945. doi: 10.1121/1.427140

Vliegen, J., and Oxenham, A. J. (1999). Sequential stream segregation in the absence of spectral cues. *J. Acoust. Soc. Am.* 105, 339–346. doi: 10.1121/1.424503

Warren, R., and Obusek, C. (1972). Identification of temporal order within auditory sequences. *Percept. Psychophys.* 12, 86–90. doi: 10.3758/BF03212848

Winkler, I., Schroger, E., and Cowan, N. (2001). The role of large-scale memory organization in the mismatch negativity event-related brain potential. *J. Cogn. Neurosci.* 13, 59–71. doi: 10.1162/089892901564171

Winkler, I., Sussman, E., Tervaniemi, M., Horváth, J., Ritter, W., and Näätänen, R. (2003). Preattentive auditory context effects. *Cogn. Affect. Behav. Neurosci.* 3, 57–77. doi: 10.3758/CABN.3.1.57

Yabe, H., Koyama, S., Kakigi, R., Gunji, A., Tervaniemi, M., Sato, Y., et al. (2001). Automatic discriminative sensitivity inside temporal window of sensory memory as a function of time. *Brain Res. Cogn. Brain. Res.* 12, 39–48. doi: 10.1016/S0926-6410(01)00027-1

Zachau, S., Rinker, T., Korner, B., Kohls, G., Maas, V., Hennighausen, K., et al. (2005). Extracting rules: early and late mismatch negativity to tone patterns. *Neuroreport* 16, 2015–2019. doi: 10.1097/00001756-200512190-00009

Zhang, Y., Koerner, T., Miller, S., Grice-Patil, Z., Svec, A., Akbari, D., et al. (2011). Neural coding of formant-exaggerated speech in the infant brain. *Dev. Sci.* 14, 566–581. doi: 10.1111/j.1467-7687.2010.01004.x

Zhang, Y., Kuhl, P. K., Imada, T., Iverson, P., Pruitt, J., Stevens, E. B., et al. (2009). Neural signatures of phonetic learning in adulthood: a magnetoencephalography study. *Neuroimage* 46, 226–240. doi: 10.1016/j.neuroimage.2009.01.028

Zhang, Y., Kuhl, P. K., 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

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

Received: 15 February 2014; accepted: 18 August 2014; published online: 12 September 2014.

Citation: Nie Y, Zhang Y and Nelson PB (2014) Auditory stream segregation using bandpass noises: evidence from event-related potentials. Front. Neurosci. ***8****:277. doi: 10.3389/fnins.2014.00277*

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

frontiers in
NEUROSCIENCE

GENERAL COMMENTARY
published: 04 September 2014
doi: 10.3389/fnins.2014.00280

Corrigendum: The effect of stimulus context on the buildup to stream segregation

Elyse S. Sussman*

Department of Neuroscience, Albert Einstein College of Medicine, Bronx, NY, USA
**Correspondence: elyse.sussman@einstein.yu.edu*

Edited and reviewed by:
Susann Deike, Leibniz Institute for Neurobiology, Germany

Keywords: auditoryperception, mismatchnegativity, streamsegregation, event-relatedpotentials, auditoryscene analysis

A commentary on

The effect of stimulus context on the buildup to stream segregation
by Sussman-Fort, J., and Sussman, E. (2014) Front. Neurosci. 8:93. doi: 10.3389/fnins.2014.00093

One of the funding sources was omitted from the Acknowledgments list. The corrected list is as follows. This research was supported by the National Institutes of Health (R01DC004263), the Army Research Office (58760LS, Elyse S. Sussman) and the MSTP training grant (T32-GM007288, Jonathan Sussman-Fort).

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

Received: 04 August 2014; accepted: 18 August 2014; published online: 04 September 2014.

*Citation: Sussman ES (2014) Corrigendum: The effect of stimulus context on the buildup to stream segregation. Front. Neurosci. **8**:280. doi: 10.3389/fnins.2014.00280*
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in NEUROSCIENCE

GENERAL COMMENTARY
published: 22 August 2014
doi: 10.3389/fnins.2014.00264

Corrigendum: Independent impacts of age and hearing loss on spatial release in a complex auditory environment

Frederick J. Gallun[1,2]*, Anna C. Diedesch[3], Sean D. Kampel[1] and Kasey M. Jakien[2]

[1] *Department of Veterans Affairs, National Center for Rehabilitative Auditory Research, Portland VA Medical Center, Portland, OR, USA*
[2] *Otolaryngology / Head and Neck Surgery, Oregon Health and Science University, Portland, OR, USA*
[3] *Hearing and Speech Sciences, Vanderbilt University, Nashville, TN, USA*
**Correspondence: frederick.gallun@va.gov*

Edited and reviewed by:
Elyse S. Sussman, Albert Einstein College of Medicine, USA

Keywords: spatial hearing, aging, hearing loss, virtual spatial array, sensation level

A corrigendum on

Independent impacts of age and hearing loss on spatial release in a complex auditory environment
by Gallun, F. J., Diedesch, A. C., Kampel, S. D., and Jakien, K. M. (2014). Front. Neurosci. 8:264. doi: 10.3389/fnins.2014.00264

In Gallun et al. (2013), the sensation levels of the stimuli were incorrectly calculated. The calculation described on page 6 was implemented correctly, but the conversion from HL (hearing level) to SPL (sound pressure level in standardized units) was not correct for the headphones used in the experiment. Rather than adding 22 dB to the HL value, it would have been correct to add 12.5 dB to the HL value (ANSI, 2004). Consequently, the value reported in sensation level (SL), which is the level in dB relative to the speech reception level, was reported as 9.5 dB lower than is correct.

The following changes should be made to clarify this issue. Bold text indicates where changes have occurred.

(1) The sentence that begins on the end of page 5 and continues onto page 6 should read: This was achieved by first measuring the level at which each listener could just identify speech presented by the audiologist over the audiometer, transforming that value from hearing level (HL) to dB SPL by adding **12.5 dB**, and then adding **39.5 dB** to that level to obtain the level of the target sentence, which was always fixed during the adaptive tracking procedure.

(2) The final sentence in the first paragraph of page 6 should read: No listeners were tested for whom the **39.5 dB** SL level would have resulted in maskers that exceeded 85 dB SPL.

(3) The second sentence in the Results section on the top of page 6 should read: Target levels, which were set **39.5 dB** above the SRT, varied from 47 dB SPL to 72 dB SPL, reflecting the 25 dB range of SRT values in the sample.

(4) The third sentence in the second column of page 7 should read: This is likely due to two factors: the better overall hear- ing of the participants (PTAs for the frequencies 0.5, 1, and 2 kHz were 15.1 dB in Experiment One and 9.4 dB in Experiment Two) and the use of an equal SL target set at **39.5 dB** above individual SRTs for each ear.

REFERENCE

ANSI. (2004). *American National Standard Specification for Audiometers*. New York, NY: ANSI S3.6-2004.

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

Received: 05 August 2014; accepted: 05 August 2014; published online: 22 August 2014.
*Citation: Gallun FJ, Diedesch AC, Kampel SD and Jakien KM (2014) Corrigendum: Independent impacts of age and hearing loss on spatial release in a complex auditory environment. Front. Neurosci. **8**:264. doi: 10.3389/fnins.2014.00264*
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.

frontiers in
NEUROSCIENCE

GENERAL COMMENTARY
published: 13 November 2014
doi: 10.3389/fnins.2014.00367

Corrigendum: Probing auditory scene analysis

Susann Deike[1]*, Susan L. Denham[2,3] and Elyse S. Sussman[4,5]

[1] *Special Lab Non-invasive Brain Imaging, Leibniz Institute for Neurobiology, Magdeburg, Germany*
[2] *Cognition Institute, University of Plymouth, Plymouth, UK*
[3] *School of Psychology, University of Plymouth, Plymouth, UK*
[4] *Department of Neuroscience, Albert Einstein College of Medicine, Bronx, NY, USA*
[5] *Department of Otorhinolaryngology-Head and Neck Surgery, Albert Einstein College of Medicine, Bronx, NY, USA*
**Correspondence: sdeike@lin-magdeburg.de*

Edited and reviewed by:
Isabelle Peretz, Université de Montréal, Canada

Keywords: auditory scene analysis, multistable perception, ambiguity, realistic auditory scenes, stream segregation

A corrigendum on

Probing auditory scene analysis
by Deike, S., Denham, S. L., and Sussman, E. (2014). Front. Neurosci. 8:293. doi: 10.3389/fnins.2014.00293

One of the funding sources was omitted from the Acknowledgments list and one funding source was incorrectly assigned. The corrected list is as follows. We thank the authors for their contributions and the reviewers for their useful comments. Susann Deike was funded by the "Deutsche Forschungsgemeinschaft" [SFB/TRR31]. Elyse S. Sussman was funded by the National Institutes of Health (DC004263).

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

Received: 26 September 2014; accepted: 27 October 2014; published online: 13 November 2014.

*Citation: Deike S, Denham SL and Sussman ES (2014) Corrigendum: Probing auditory scene analysis. Front. Neurosci. **8**:367. doi: 10.3389/fnins.2014.00367*
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience.