ACTIVE COGNITIVE PROCESSING FOR AUDITORY PERCEPTION

EDITED BY: Shannon Heald, Stephen Charles Van Hedger and Howard Charles Nusbaum

PUBLISHED IN: Frontiers in Neuroscience and Frontiers in Psychology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-147-0
DOI 10.3389/978-2-88976-147-0

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

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

ACTIVE COGNITIVE PROCESSING FOR AUDITORY PERCEPTION

Topic Editors:
Shannon Heald, The University of Chicago, United States
Stephen Charles Van Hedger, Huron University College, Canada
Howard Charles Nusbaum, The University of Chicago, United States

Citation: Heald, S., Van Hedger, S. C., Nusbaum, H. C., eds. (2022). Active Cognitive Processing for Auditory Perception. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-147-0

Table of Contents

frontiers in Psychology

ORIGINAL RESEARCH
published: 07 April 2021
doi: 10.3389/fpsyg.2021.631060

Second Language Experience Facilitates Sentence Recognition in Temporally-Modulated Noise for Non-native Listeners

*Jingjing Guan, Xuetong Cao and Chang Liu**

Department of Speech, Language, and Hearing Sciences, University of Texas at Austin, Austin, TX, United States

Non-native listeners deal with adverse listening conditions in their daily life much harder than native listeners. However, previous work in our laboratories found that native Chinese listeners with native English exposure may improve the use of temporal fluctuations of noise for English vowel identification. The purpose of this study was to investigate whether Chinese listeners can generalize the use of temporal cues for the English sentence recognition in noise. Institute of Electrical and Electronics Engineers (IEEE) sentence recognition in quiet condition, stationary noise, and temporally-modulated noise were measured for native American English listeners (EN), native Chinese listeners in the United States (CNU), and native Chinese listeners in China (CNC). Results showed that in general, EN listeners outperformed the two groups of CN listeners in quiet and noise, while CNU listeners had better scores of sentence recognition than CNC listeners. Moreover, the native English exposure helped CNU listeners use high-level linguistic cues more effectively and take more advantage of temporal fluctuations of noise to process English sentence in severely degraded listening conditions [i.e., the signal-to-noise ratio (SNR) of −12 dB] than CNC listeners. These results suggest a significant effect of language experience on the auditory processing of both speech and noise.

Keywords: sentence recognition in noise, temporally-modulated noise, non-native speakers, masking release from temporal modulation, second language experience

OPEN ACCESS

Edited by:
Stephen Charles Van Hedger, Western University, Canada

Reviewed by:
Dan Zhang, Tsinghua University, China
Yang Zhang, University of Minnesota Health Twin Cities, United States

***Correspondence:**
Chang Liu
changliu@utexas.edu

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

Received: *19 November 2020*
Accepted: *11 March 2021*
Published: *07 April 2021*

Citation:
Guan J, Cao X and Liu C (2021) Second Language Experience Facilitates Sentence Recognition in Temporally-Modulated Noise for Non-native Listeners. Front. Psychol. 12:631060. doi: 10.3389/fpsyg.2021.631060

INTRODUCTION

In daily life, speech sounds are usually perceived in noisy and complex listening environments. A number of studies have demonstrated that speech recognition in noise is much harder for non-native listeners than for native listeners regardless of stimuli are either vowels, consonants, words, or sentences (Mayo et al., 1997; Cooke et al., 2008; Cutler et al., 2008; Shi, 2010; Jin and Liu, 2012; Mi et al., 2013; Rimikis et al., 2013; Guan et al., 2015; Zinszer et al., 2019). Even though non-native listeners could achieve comparable performance with native listeners in quiet, non-native listeners performed much worse than native listeners in long-term speech-shaped noise (LTSSN) and multi-talker babble (MTB; Garcia Lecumberri and Cooke, 2006; Cutler et al., 2008). The difficulties of non-native listeners' speech perception in noise were mainly impacted by their language backgrounds including their native language (L1) and

second language (L2) experience, and listening conditions (Jin and Liu, 2012; Mi et al., 2013; Guan et al., 2015). For example, Jin and Liu (2012) reported that two groups of English non-native listeners (native Chinese and native Korean listeners) had similar scores in English sentence recognition in quiet and in LTSSN and these two non-native groups had significantly lower performance than native English listeners (EN). However, Korean listeners outperformed Chinese listeners in MTB, indicating that listeners' native language (Korean or Chinese) might interfere with speech noise. However, in a follow-up study, Jin and Liu (2014) found no significant difference in English vowel identification in LTSSN and MTB, suggesting that L2 speech recognition in noises was dependent on L1 experience, the type of noise, and speech materials.

Compared with native listeners, non-native listeners with incomplete knowledge of target language and less automatic language processing had to spend more efforts to recognize speech in interfering noise (Kousaie et al., 2019). However, L2 exposure may help non-native listeners' perception to be fine-tuned with sounds of L2. Previous studies showed two groups of native Chinese listeners, one in the US (CNU) with L2 exposure and the other in China (CNC) without L2 exposure, had similar performance of English vowel identification in quiet and LTSS noise, whereas CNU listeners showed significantly higher scores in MTB (Mi et al., 2013) and temporally modulated noise (Guan et al., 2015). Both studies argued that extensive native English (L2) experiences of CNU listener might help improve their capacity to perceive English vowels in temporally fluctuating maskers.

In most listening conditions, background noise is non-stationery, and amplitude of noise fluctuates over the time. It has been well-established that listeners with normal hearing including native and non-native listeners typically performed better in temporally modulated noise than in stationary noise (Festen and Plomp, 1990; Jin and Nelson, 2006; Broersma and Scharenborg, 2010; Stuart et al., 2010; Guan et al., 2015). This phenomenon, called masking release from the temporal modulation of noise, was usually explained as an ability to catch the glimpses of speech information during the short gaps of fluctuating noise and integrate these glimpses to restore the content of speech. Previous studies demonstrated that native listeners benefited more than non-native listeners from temporally fluctuating noise and obtained a larger masking release (Stuart et al., 2010; Guan et al., 2015). Guan et al. (2015) examined English vowel identification for three groups of listeners: native English (EN) listeners, CNU, and CNC listeners. As expected, EN listeners had significantly greater masking releases from temporal modulation in noise at low signal-to-noise ratios (SNRs) than the two groups of Chinese listeners. Moreover, CNU listeners had more masking releases for vowel identification than CNC listeners. Native English experiences for 1–3 years might help CNU listeners become better tuned to the temporal structure of English speech such that they had larger masking release. It was well demonstrated the basic psychophysical capacities to detect basic auditory tasks including temporal processing (Guan et al., 2015) were similar across listeners with different native language (L1) backgrounds and L2 exposure. Nevertheless, compared with basic auditory tasks, higher-level speech recognition in noise not only relied on listeners' basic psychophysical capacities but also associated with listeners' L1 and L2 experiences (Guan et al., 2015; Liu and Jin, 2015; Huo et al., 2016). Thus, this study was to investigate whether the CNU-CNC difference in using temporal dips of noise for English vowel identification could be extended in English sentence recognition, which was more ordinary in daily speech communication. Previous studies showed that speech materials significantly influenced the effect of L1 experience on L2 speech identification, i.e., significant difference in English sentence recognition in MTB between native Chinese and native Korean listeners (Jin and Liu, 2012), but no group difference for English vowel recognition in MTB between the two groups (Jin and Liu, 2014). Thus, the effect of L2 experiences (e.g., with or without L2 exposures) may also rely on speech materials (vowels and sentences).

MATERIALS AND METHODS

Listeners

Three groups of listeners whose ages ranged from 20 to 35 years old were recruited based on their language experience: native EN, CNU, and CNC. The EN group comprised 10 listeners, while each of two native Chinese groups consisted of 14 listeners. All listeners had normal hearing with pure-tone thresholds $\leq$ 15 dB HL at octave intervals between 250 and 8,000 Hz (ANSI, 2010). Listeners in the EN and CNU groups were undergraduate or graduate students from the University of Texas at Austin, and listeners in the CNC group were undergraduate or graduate students from Beijing Normal University. Listeners in the two native Chinese groups started their formal school-based English education at 9–13 years old in China. The CNU group had internet-based Test of English as a Foreign Language (TOEFL) scores at least 80 with the United States residency of 1–3 years. Listeners in the CNC group passed the College English Test Band 4 (CET-4) in China without any residency history in English-speaking countries. The CET-4 is required for most of undergraduate students in China to receive their bachelor degree. All listeners received the consent form at the beginning of this study approved by the Ethics Review Board of Beijing Normal University and the Institutional Review Board of the University of Texas at Austin. They were paid for their participation.

Stimuli

Sentence recognition performance was measured by using the Institute of Electrical and Electronics Engineers (IEEE) sentences, which was recorded from a female native English speaker. The IEEE sentences include 72 lists with each list composed of 10 sentences. Each sentence contains five key words. For a given listening condition, one sentence list was randomly selected, and was presented only once across all conditions.

Sentences were presented in either quiet or noise. Two types of noise, stationary LTSSN and temporally modulated noise, were presented at 70 dB SPL. The LTSSN was generated from

Gaussian noise that was shaped by a filter with an average spectrum of a 12-talker babble (Kalikow et al., 1977). SNRs were manipulated at −12, −9, −6, −3, 0, and 3 dB. In addition, the temporally modulated noise was generated by applying a temporal-modulated envelope on the stationary LTSSN with the modulation frequency at 4, 16, and 64 Hz, and the modulation depth of 100%. The phase of the temporal modulation of noise was randomized between 0 and 2π. Speech and noise levels were calibrated with an AEC201-A IEC 60318-1 ear simulator by a Larson-Davis sound-level meter (Model 2800) with a linear weighting band.

Procedure

For each listener, there were 25 experimental conditions (quiet, six SNRs × four modulation frequencies including the stationary noise). For a given condition, a randomly selected IEEE list was presented. Speech signals and noise, digitized at 12,207 Hz, were presented *via* SONY MDR-7506 headphones to the right ear of the listeners who were seated in a quiet test room. Stimulus presentation was controlled by the Tucker-Davis Technologies mobile processor (RM1). Listeners were seated in front of an LCD monitor and asked to type what they heard in a text box. After the sentence presentation, listeners were required to respond within 30 s. Sentence recognition score in percent correct for each sentence was evaluated offline and calculated according to the number of words correctly typed out of the five key words. The final sentence recognition performance for one listening condition was based on the average score of all 10 sentences. Before data collection, listeners were trained with a 5-min practice session of sentence recognition in a quiet listening condition to familiarize them with the experimental procedure.

After the practice session, listeners were examined with the test session in quiet first followed by the noise conditions. The order of the 24 noise conditions was randomized, which was manipulated by the software Sykofizx®.

RESULTS

Sentence Recognition in Quiet

In quiet, the average sentence recognition score was 99.6% for the EN group, 70.7% for the CNU group, and 41.0% for the CNC group. In order to examine the effect of listener group on sentence recognition performance in quiet, a generalized linear model (GLM) was conducted with the scores in quiet as the dependent variable and listener group as a fixed effect. Results revealed that the main effect of listener group was significant [$\chi^2(2) = 136.35$, $p < 0.001$]. Pairwise comparisons demonstrated that the EN group performed significantly better than the two Chinese groups, while the CNU group significantly outperformed the CNC group (all $p < 0.001$).

Sentence Recognition in Noise and Masking Release From Temporal Modulation

As shown in **Figure 1**, sentence recognition scores in noise conditions were lower relative to scores in quiet for all the three groups. The native group outperformed the two non-native groups for all noise conditions. The data were analyzed using a generalized linear mixed model (GLMM) in SPSS software (version 26) where the performance of sentence recognition in noise as a dependent variable. In the model, fixed effects included listener group, SNR, modulation frequency, their two-way interactions, and the three-way interaction. To account for baseline differences in sentence recognition performance across subjects, we also included subject as a random effect. The GLMM analyses showed significant main effects of listener group, SNR, and modulation frequency (all $p < 0.001$). The two-way interactions between listener group and SNR [$F(10,838) = 8.62$, $p < 0.001$], SNR, and modulation frequency [$F(15,838) = 4.67$, $p < 0.001$], and the three-way interaction were significant [$F(30,838) = 1.59$, $p < 0.05$].

In order to further explore how the impact of listener group on sentence recognition in noise changed as the SNR varied, planned comparisons at each SNR indicated that native English listeners performed significantly better than two non-native Chinese groups regardless of modulation frequencies (all $p < 0.05$), while the CNU group obtained significantly higher scores of sentence recognition in noise than the CNC group across all the SNR conditions from −12 to +3 dB (all $p < 0.05$).

In order to examine the amount of masking release of sentence recognition from the temporal modulation of noise, the difference between the scores with temporally modulated noise (i.e., modulation frequency of 4, 16, and 64 Hz) and the scores with stationary noise were calculated (see **Figure 2**). With the masking release of sentence recognition in temporally modulated noise being a dependent variable, a GLMM model with fixed effects (i.e., listener group, SNR, modulation frequency, their two-way interactions, and the three-way interaction) were run. Subject was entered as a random effect factor. The GLMM model yielded significant main effects of modulation frequency [$F(2,628) = 6.46$, $p < 0.05$] and SNR [$F(5,628) = 13.47$, $p < 0.001$], a significant interaction between listener group and SNR [$F(10,628) = 5.18$, $p < 0.001$], and a significant interaction between SNR and modulation frequency [$F(10,628) = 2.40$, $p < 0.05$]. Follow-up analysis showed a significant effect of listener group only at the SNR of −12 dB [$F(2,628) = 17.65$, $p < 0.001$], indicating the EN group received significantly greater masking release than the CNU group whose masking release was significantly larger than that of the CNC group only at the SNR of −12 dB. No significant group differences were observed at other SNRs (all $p > 0.05$).

Normalized Sentence Recognition Performance in Noise and Related Masking Release

In order to rule out the effect of the language proficiency across the three listener groups, normalized scores of sentence recognition were computed by the percentage scores in noise listening conditions divided by the percentage scores in quiet for each listener. As shown in **Figure 3**, for example, the EN, CNU, and CNC groups had 84.4, 38.6, and 19.4% scores for sentence recognition in the stationary noise at the SNR of −9 dB.

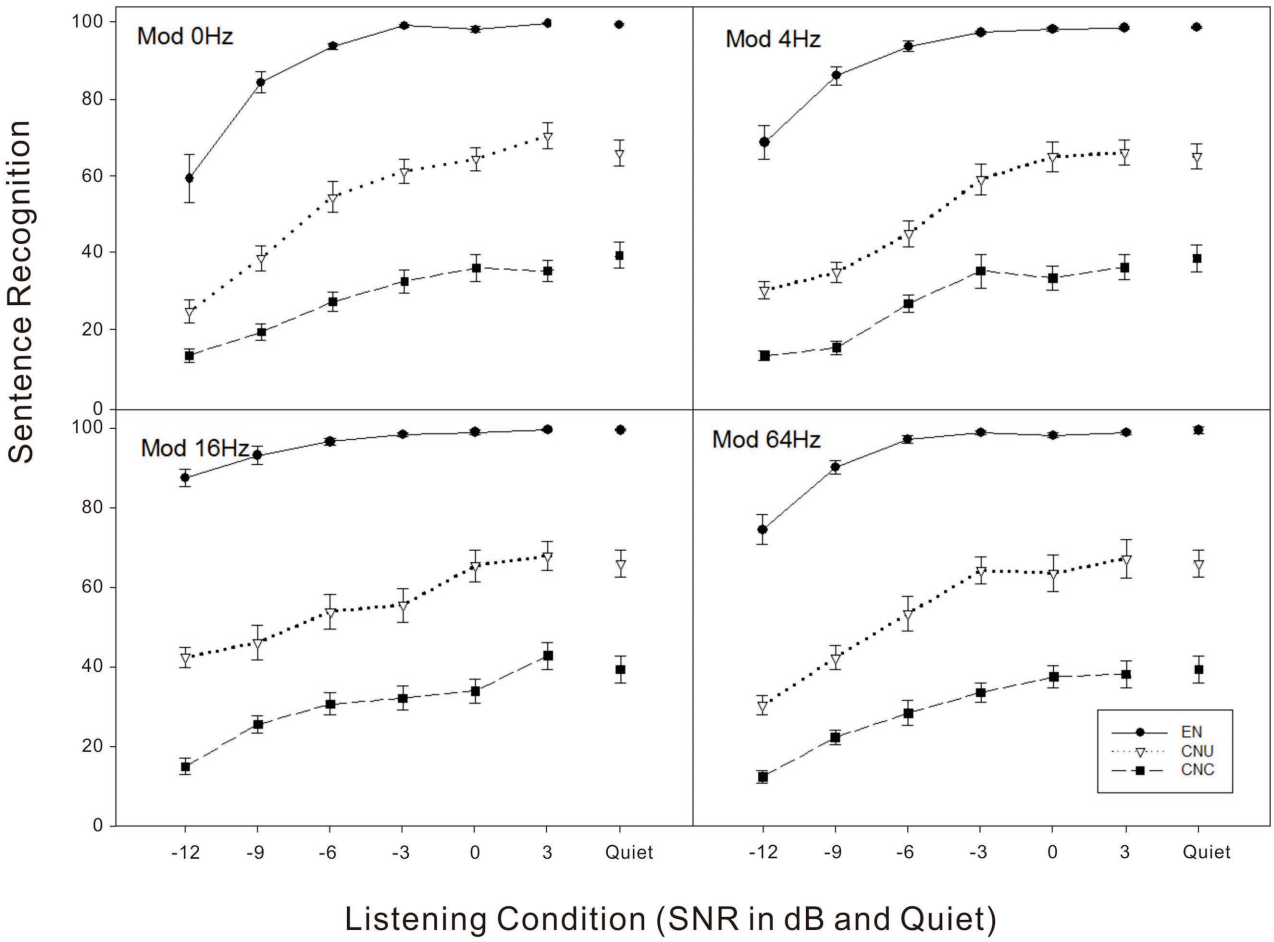

FIGURE 1 | Sentence recognition scores in percent correct as a function of listening condition (quiet and signal-to-noise ratio in dB) for four noise conditions (top left - Mod 0 Hz: stationary noise; top right - Mod 4 Hz: noise temporally modulated at 4 Hz; bottom left - Mod 16 Hz: noise temporally modulated at 16 Hz; bottom right - Mod 64 Hz: noise temporally modulated at 64 Hz) for three groups of listeners: English-native (EN) listeners, Chinese-native listeners in the United States (CNU), and Chinese-native listeners in China (CNC).

As the baseline, the scores of sentence recognition in quiet for the three groups were 99.6, 70.7, and 41.0%, respectively. Therefore, the normalized scores of sentence recognition were then 0.85 (84.4/99.6%), 0.55 (38.6/70.7%), and 0.47 (19.4/41.0%) for the EN, CNU, and CNC groups, respectively. The GLMM model on normalized sentence recognition performance revealed main effects of listener group, SNR, and modulation frequency were significant (all $p < 0.05$). The interaction between listener group and SNR [$F(10,838) = 10.33$, $p < 0.001$] and the interaction between SNR and modulation [$F(15,838) = 2.64$, $p < 0.05$] were also significant. *Post hoc* analysis showed that normalized performance of EN group was significantly better than the two non-native Chinese groups (all $p < 0.05$) while two Chinese groups had no significant difference at the SNRs of −12, −9, and −6 dB. No other significant differences among three listener groups were found at the SNRs of −3, 0, and 3 dB.

The amount of masking release from the temporal modulation of noise for normalized sentence recognition scores was also calculated by subtracting the normalized scores with stationary noise from the normalized scores with temporally modulated noise (i.e., modulation frequency of 4, 16, and 64 Hz; see **Figure 4**). The GLMM model was applied to examine the effects of listener group, SNR, modulation frequency, and their interactions on masking release for normalized sentence recognition performance. Results demonstrated that normalized masking release was significantly contributed by the main effects of modulation frequency [$F(2,628) = 4.17$, $p < 0.05$] and SNR [$F(5,628) = 4.46$, $p < 0.05$], and the interaction of listener group and SNR [$F(10,628) = 2.10$, $p < 0.05$] but not by the other main effect and interaction effects (all $p > 0.05$). *Post hoc* analysis suggested listeners achieved significantly greater normalized masking release at the modulation frequency of 16 Hz than at 4 Hz, while no significant differences were observed for other comparisons among modulation frequencies. Furthermore, a significant group difference on normalized masking release was only found at the SNR of −12 dB [$F(2,628) = 5.85$, $p < 0.05$] but not at other SNR conditions. Specifically, as shown in **Figure 4**, normalized masking release in severely degraded

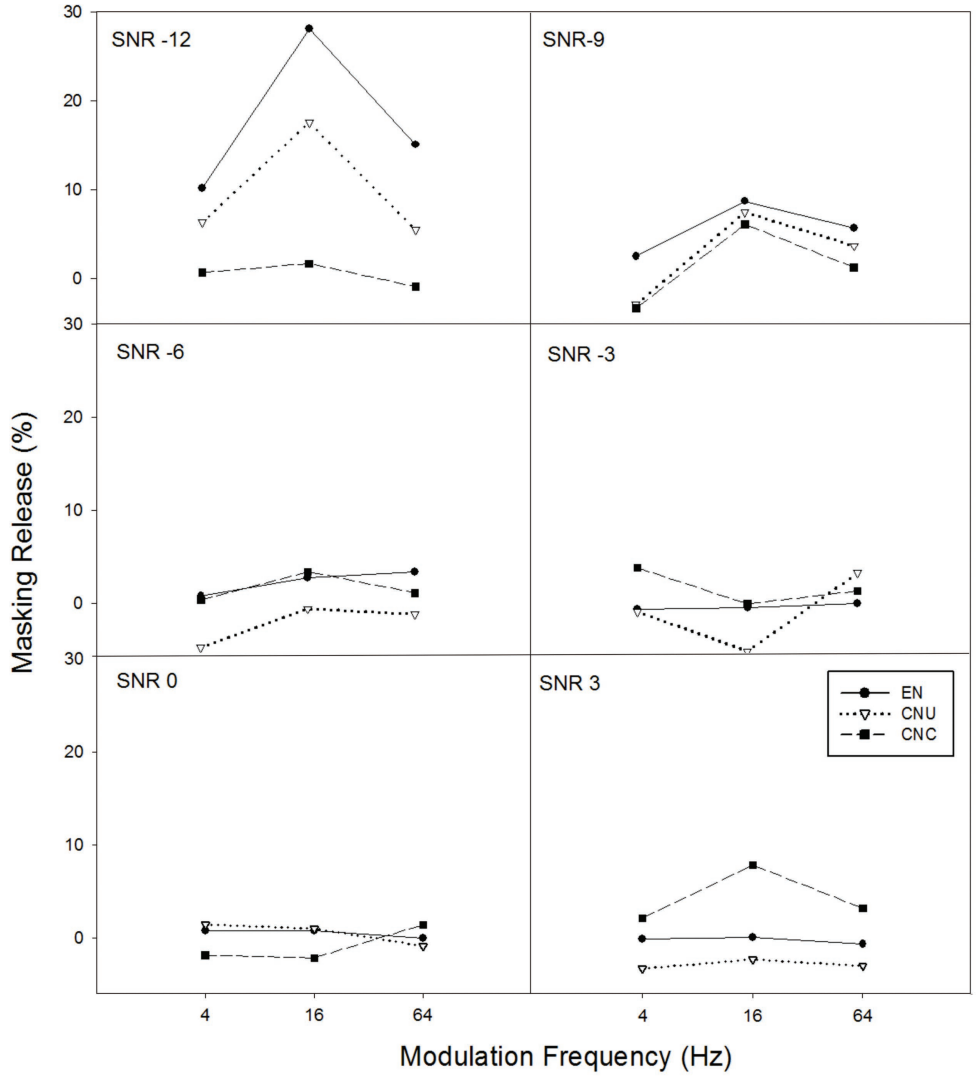

FIGURE 2 | The masking release from the temporal modulation of noise in percent (e.g., scores in temporally-modulated noise minus scores in stationary noise) as a function of temporal modulation frequency of noise for six SNRs (−12, −9, −6, −3, 0, and 3 dB for each panel) for three groups of listeners: EN listeners, CNU, and CNC.

listening conditions (i.e., the SNR of −12 dB) for the EN and CNU groups had significantly greater masking release than the CNC group (both $p < 0.05$), while the EN and CNU achieved similar masking release ($p > 0.05$).

Correlation Between Sentence Recognition Performance and Masking Release

In order to examine whether English proficiency level of Chinese-native speakers significantly affected the masking release from the temporal modulation of noise, the correlation between sentence recognition scores in quiet and the masking release at each listening condition (e.g., at a given SNR and one temporal modulation frequency of noise) was analyzed. Results indicated a significant correlation for Chinese-native speakers including both CNU and CNC listeners at the SNR of −12 dB for the temporal modulation frequency of 16 Hz for both original (Pearson $r = 0.546$, $p < 0.05$) and normalized (Pearson $r = 0.392$, $p < 0.05$) data. That is, for Chinese-native speakers, better sentence recognition in quiet and larger masking release from the temporal modulation in noise.

DISCUSSION

Disadvantage for Non-native Listeners on Sentence Recognition in Noise

Consistent with previous studies (Mayo et al., 1997; Shi, 2010; Jin and Liu, 2012; Zhong et al., 2019), EN listeners' performance

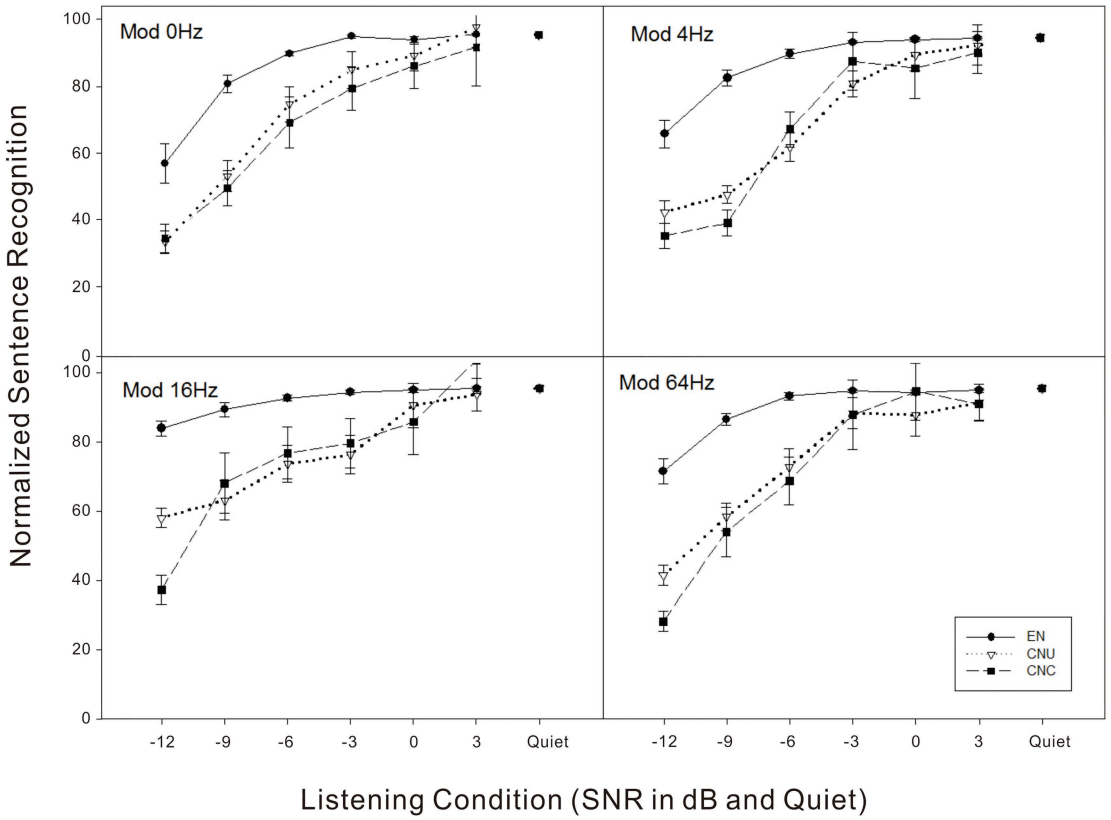

FIGURE 3 | Normalized sentence recognition scores (e.g., recognition scores in noise/quiet divided by scores in quiet) as a function of listening condition (quiet and signal-to-noise ratio in dB) for four noise conditions (top left – Mod 0 Hz: stationary noise; top right – Mod 4 Hz: noise temporally modulated at 4 Hz; bottom left – Mod 16 Hz: noise temporally modulated at 16 Hz; and bottom right – Mod 64 Hz: noise temporally modulated at 64 Hz) for three groups of listeners: EN listeners, CNU, and CNC.

in quiet (99.6%) showed remarkable native language advantages over the two non-native groups (CNU: 70.7%; CNC: 41.0%). However, CNU listeners with native English experiences of 1–3 years significantly outperformed CNC listeners in sentence recognition in quiet, in contrast with the findings of our previous studies (Mi et al., 2013; Guan et al., 2015), which reported there was no difference in English vowel identification in quiet between CNU (accuracy: 70%) and CNC listeners (accuracy: 64%). It was possibly because that sentences contain redundant cues (e.g., acoustic-phonetic, semantic, syntactic, and prosodic cues), while vowels have limited acoustic-phonetic cues. Compared with vowels, sentences were more related with language experience, especially when the IEEE sentences were contextually more complicated than other sentence materials (Nilsson et al., 1994). Zhong et al. (2019) found that for English sentence recognition in quiet and four-talker babble, EN listeners weighted speech cues including semantic and syntactic cues equally, whereas CNU listeners primarily relied on semantic information. That is, the gap between EN and CNU listeners was larger as the sentences became from high to low predictability. The findings of this study combined with those of previous studies in our laboratories suggested that the non-native disadvantage in English sentence perception was significantly dependent on the complexity of sentences, e.g., for sentences with high predictability, there was only 5–10% gap between EN and CNU listeners (Jin and Liu, 2012; Zhong et al., 2019), while for sentences in low predictability, the gap was enlarged to more than 20% in this study and study of Zhong et al.(2019). Altogether, these results indicate that non-native listeners have significant challenges of processing L2 speech signals when primarily using a bottom-up approach, i.e., semantic information as a top-down cue was missing in low-predictability sentences.

Moreover, the comparison between CNU and CNC listeners indicated no difference in vowel identification (Guan et al., 2015), but a significant difference in sentence recognition in this study suggest that native English experience of 1–3 years may improve non-native listeners' processing of high-level linguistic cues of English speech sounds, but not the processing of phonetic cues, the improvement of which may need specific phonetic training.

In this study, the two non-native groups started to learn English around 9–13 years old, who were later learners of the second language. The non-native disadvantage is consistent with previous findings that late bilingual speakers (e.g., age of acquisition about 10–13 years old) were not able to

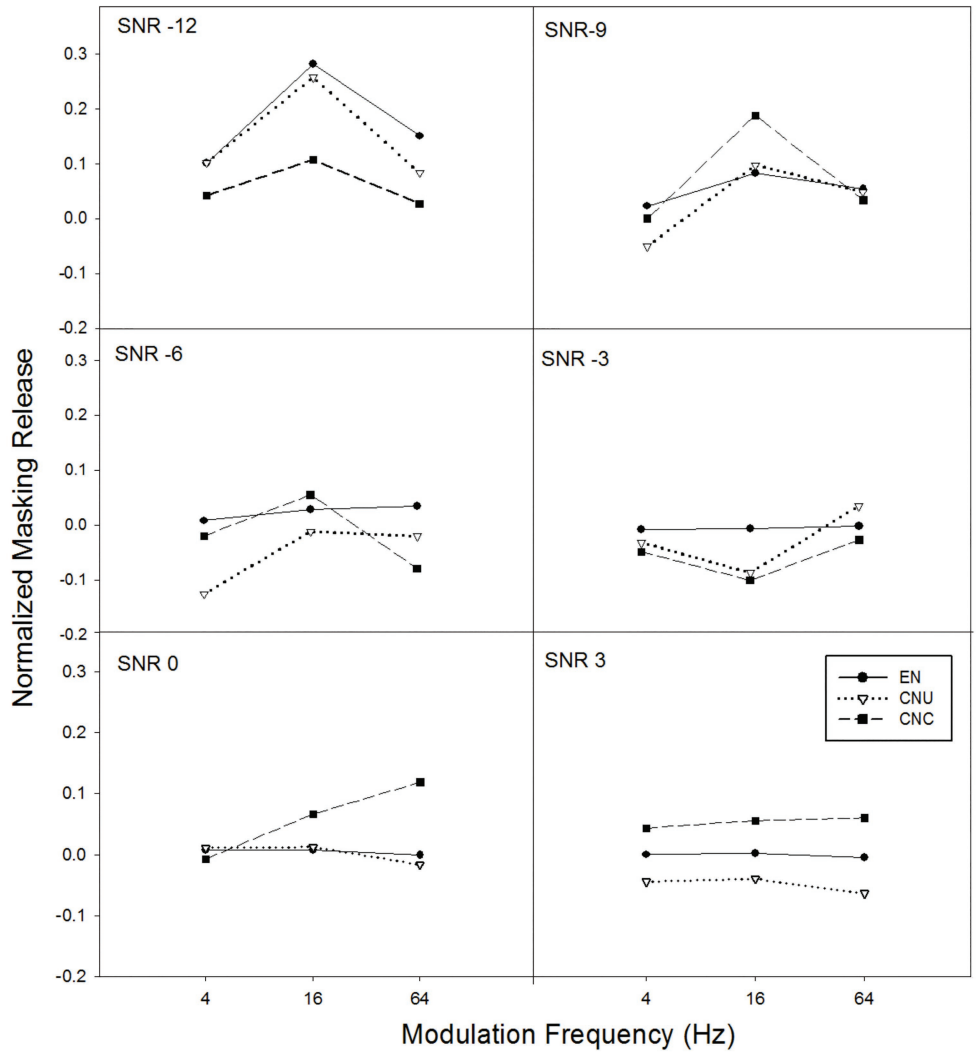

FIGURE 4 | The normalized masking release from the temporal modulation of noise (e.g., normalized scores in temporally-modulated noise minus normalized scores in stationary noise) as a function of temporal modulation frequency of noise for six SNRs (–12, –9, –6, –3, 0, and 3 dB for each panel) for three groups of listeners: EN listeners, CNU, and CNC.

recognize sentences in noise as efficiently as monolingual English speakers (Mayo et al., 1997; Shi, 2010). Further study with more participant variables, such as L2 learning environment, motivation, L2 learning styles, and cognitive abilities, needs to be included and investigated in the future, which may also contribute to the challenges of perceiving L2 speech in non-optimal listening conditions for nonnative listeners (Yang et al., 2017; Kousaie et al., 2019). In order to further discuss sentence recognition in noise by minimizing the effect of English proficiency for Chinese-native listeners, normalized scores were calculated, and the noise effect was then analyzed based on the normalized scores (see **Figure 3**). Results indicated that only in severely degraded listening conditions (i.e., SNRs of −12 to −6 dB,), the native English group had significantly better scores than the two non-native groups. There was no significant group difference for normalized sentence recognition scores at more favorable listening conditions (SNRs from −3 to +3 dB). Results suggested that when the language proficiency was controlled (e.g., the sentence recognition scores were normalized), noise seemed to affect the three groups of listeners similarly at medium and high SNRs, but influenced non-native listeners more negatively than native listeners at low SNRs.

In this study, in acoustically degraded conditions, non-native listeners showed disadvantages in sentence recognition in noise from the SNR of −12 to +3 dB. It should be noted that whether background noise causes greater difficulty for non-native listeners than for native listeners depends on a variety of factors, such as SNR, speech materials, and noise type. For example, for English vowel identification, LTSSN did not enlarge the gap between EN and Chinese-native listeners, while MTB did for high and middle SNRs (Mi et al., 2013). On the other

hand, MTB made no changes in the non-native disadvantages in quiet for CNU listeners for broad SNRs, but resulted in larger non-native advantage for CNC listeners at very low SNRs (e.g., −15 and −20 dB SNR). For sentence recognition, the gap between native and non-native listeners became larger for sentence recognition in MTBs for CNU listeners than in quiet (Jin and Liu, 2012; Zhong et al., 2019), particularly at high and medium SNRs. Thus, the effect of SNRs on non-native disadvantage in speech recognition is quite complicated, depending on a number of factors, such as speech materials, noise type, listeners' L2 experience, and SNR.

Masking Release From Temporal Modulation for Non-native Listeners Mediated With Language Experience

As expected, both native and non-native listeners benefited from the temporal modulation of noise, especially at the low SNRs (e.g., −12 dB; see **Figure 2**). With regard to the masking release from temporal modulation of noise across the three groups, EN and CNU listeners obtained larger masking releases than CNC listeners at the SNR of −12 dB, consistent with previous findings in English vowel identification (Guan et al., 2015) and Chinese vowel identification (Li et al., 2016). Such group difference in the masking release at low SNRs was also observed in **Figure 4** when the effect of English proficiency was minimized (e.g., the scores of sentence recognition were normalized). Thus, it is likely that as for either vowel or sentence testing, smaller benefit from the temporal modulation of noise for CNC listeners may be due to the lack of native English experience with the temporal structure of the target language.

The amount of masking release from temporal modulation at the low SNRs for native and nonnative listeners depended on the temporal modulation frequency of background noise. It was found that in this study, the masking release obtained by listeners in both **Figures 2, 4** was the significantly highest at the modulation frequency of 16 Hz, which was consistent with previous studies (Guan et al., 2015; Li et al., 2016). Compared with performance of speech perception in temporally-modulated noise at either slower (e.g., 4 Hz) or faster modulation frequency (e.g., 64 Hz), higher scores were found at the modulation frequency between 10 and 20 Hz (Gustafsson and Arlinger, 1994). The less masking release at the modulation frequency of 4 Hz was likely due to phonemes, and syllables were more easily interfered in background noise with similar temporal characteristics. Moreover, 64-Hz modulation dips might be too short for auditory responses to recover so as to aid speech recognition (Wojtczak et al., 2011).

Altogether, the previous studies (Stuart et al., 2010; Mi et al., 2013; Guan et al., 2015; Li et al., 2016) including the current one suggest a significant difference in the effectiveness of using temporal dips in noise for speech perception between native English listeners and native Chinese listeners, and between native Chinese listeners with different English experiences (e.g., CNU and CNC listeners). However, it should be also noted that Zhang et al. (2011) reported that there was no difference in the temporal dip listening for English word recognition between EN and CNU listeners. The discrepancy in the language experience effect on temporal dip listening could be due to the difference in speech materials (e.g., vowels in Guan et al., 2015, sentences in this study and Stuart et al., 2010, and words in Zhang et al., 2011) and temporal features of background noise (e.g., temporally modulated noise at 4, 16, and 64 Hz in Guan et al., 2015 and this study, and interrupted noise in Stuart et al., 2010 and Zhang et al., 2011). Another possibility is the different SNRs used across these studies. As suggested in study of Guan et al. (2015) and this study, the group difference in temporal dip listening was found only at quite negative SNRs (e.g., −12 dB), while the SNR of study of Zhang et al. (2011) was at −10 dB. Thus, more systematical studies are needed to fully examine the language experience impact on listening in temporal glimpses with a focus on the two factors: speech materials and acoustic features of noise.

The group difference in temporal dip listening in this and previous studies mentioned above may be due to the difference in the temporal structures of speech between English and Mandarin Chinese, i.e., English speech is more temporally dynamic than Mandarin Chinese (Calandruccio et al., 2010). Regular exposures to English speech environment seem to benefit native Chinese listeners by taking better advantage in using temporal glimpses of noise for speech recognition. These results suggest that speech training in temporally-varying noise may be needed and beneficial for non-native listeners if one's study is to improve their speech perception in dynamic noise backgrounds.

Limitation and Future Directions

One limitation of the current study was several variables of Chinese-native speakers were not well controlled, particularly English proficiency. As indicated above, the sentence recognition scores in quiet were significantly correlated with the masking release from the temporal modulation of noise at the SNR of −12 dB and modulation frequency of 16 Hz for Chinese-native listeners. This implies that Chinese-native listeners' English proficiency may play an import role in temporal dip listening of English speech perception. Thus, a standardized and comprehensive speech perception test including the perception of English phonemes, words, and sentences is needed in future studies to quantify English proficiency level.

Another limitation was that several variables of L2 learners were not well manipulated, such as listeners' cognitive function (e.g., working memory), learning style, L2 learning duration, and perceptual weights on speech redundant cues. Systematic studies are needed to examine the effects of these factors on L2 listeners' English speech recognition in quiet and noisy listening conditions, especially in temporally-fluctuating noise.

CONCLUSION

English listeners generally recognized sentences in quiet and noise more accurately than native Chinese listeners, while CNU listeners had better scores of sentence recognition than CNC

listeners. The native English experience helped CNU listeners process high-level linguistic cues more effectively and take greater advantage of the temporal fluctuations of noise in severely degraded listening conditions (i.e., SNR of −12 dB) compared to their counterparts in China (CNC).

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Review Board of Beijing Normal University and the Institutional Review Board of the University of Texas at Austin. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JG contributed to experimental design, data collection, data analysis, and manuscript writing. XC contributed to experimental design, data analysis, and manuscript writing, while CL made contributions to experimental design, data analysis, and manuscript writing. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the University of Texas at Austin Research Grant.

REFERENCES

ANSI (2010). S3.6, Specification for Audiometers. (New York: Acoustical Society of America)

Broersma, M., and Scharenborg, O. (2010). Native and non-native listeners' perception of English consonants in different types of noise. *Speech Comm.* 52, 980–995. doi: 10.1016/j.specom.2010.08.010

Calandruccio, L., Dhar, S., and Bradlow, A. R. (2010). Speech-on-speech masking with variable access to the linguistic content of the masker speech. *J. Acoust. Soc. Am.* 128, 860–869. doi: 10.1121/1.3458857

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

Cutler, A., Garcia Lecumberri, M. L., and Cooke, M. (2008). Consonant identification in noise by native and non-native listeners: effects of local context. *J. Acoust. Soc. Am.* 124, 1264–1268. doi: 10.1121/1.2946707

Festen, J., and Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing. *J. Acoust. Soc. Am.* 88, 1725–1736. doi: 10.1121/1.400247

Garcia Lecumberri, M., and Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *J. Acoust. Soc. Am.* 119, 2445–2454. doi: 10.1121/1.2180210

Guan, J., Liu, C., Tao, S., Mi, L., Wang, W., and Dong, Q. (2015). Vowel identification in temporal-modulated noise for native and non-native listeners: effect of language experience. *J. Acoust. Soc. Am.* 138, 1670–1677. doi: 10.1121/1.4929739

Gustafsson, H. A., and Arlinger, S. D. (1994). Masking of speech by amplitude-modulated noise. *J. Acoust. Soc. Am.* 95, 518–529. doi: 10.1121/1.408346

Huo, S., Tao, S., Wang, W., Li, M., Dong, Q., and Liu, C. (2016). Auditory detection of non-speech and speech stimuli in noise: native speech advantage. *J. Acoust. Soc. Am.* 139, EL161–EL166. doi: 10.1121/1.4951705

Jin, S.-H., and Liu, C. (2012). English sentence recognition in speech-shaped noise and multi- talker babble for English-, Chinese-, and Korean-native listeners. *J. Acoust. Soc. Am.* 132, EL391–EL397. doi: 10.1121/1.4757730

Jin, S.-H., and Liu, C. (2014). English vowel identification in quiet and noise: effects of listeners' native language background. *Front. Neurosci.* 8:305. doi: 10.3389/fnins.2014.00305

Jin, S.-H., and Nelson, P. (2006). Speech perception in gated noise: the effects of temporal resolution. *J. Acoust. Soc. Am.* 119, 3097–3108. doi: 10.1121/1.2188688

Kalikow, D. N., Stevens, K. M., and Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J. Acoust. Soc. Am.* 61, 1337–1351. doi: 10.1121/1.381436

Kousaie, S., Baum, S., Phillips, N. A., Gracco, V., Titone, D., Chen, J. K., et al. (2019). Language learning experience and mastering the challenges of perceiving speech in noise. *Brain Lang.* 196:104645. doi: 10.1016/j.bandl.2019.104645

Li, M., Wang, W., Tao, S., Dong, Q., Guan, J., and Liu, C. (2016). Mandarin Chinese vowel- plus-tone identification in noise: effect of language experience. *Hear. Res.* 331, 109–118. doi: 10.1016/j.heares.2015.11.007

Liu, C., and Jin, S. H. (2015). Auditory detection of non-speech and speech stimuli in noise: effects of listeners' native language background. *J. Acoust. Soc. Am.* 138, 2782–2790. doi: 10.1121/1.4934252

Mayo, L. H., Florentine, M., and Buus, S. (1997). Age of second-language acquisition and perception of speech in noise. *J. Speech Lang. Hear. Res.* 40, 686–693. doi: 10.1044/jslhr.4003.686

Mi, L., Tao, S., Wang, W., Dong, Q., Jin, S.-H., and Liu, C. (2013). English vowel identification in long-term speech-shaped noise and multi-talker babble for English and Chinese listeners. *J. Acoust. Soc. Am.* 133, EL391–EL397. doi: 10.1121/1.4800191

Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.* 95, 1085–1099. doi: 10.1121/1.408469

Rimikis, S., Smiljanic, R., and Calandruccioa, L. (2013). Nonnative English speaker performance on the basic english lexicon (BEL) sentences. *J. Speech Lang. Hear. Res.* 56, 792–804. doi: 10.1044/1092-4388(2012/12-0178)

Shi, L.-F. (2010). Perception of acoustically degraded sentences in bilingual listeners who differ in age of English acquisition. *J. Speech Lang. Hear. Res.* 53, 821–835. doi: 10.1044/1092-4388(2010/09-0081)

Stuart, A., Zhang, J., and Swink, S. (2010). Reception thresholds for sentence in quiet and noise for monolingual English and bilingual mandarin-English listeners. *J. Am. Acad. Audiol.* 21, 239–248. doi: 10.3766/jaaa.21.4.3

Wojtczak, M., Nelson, P. C., Viemeister, N. F., and Carney, L. H. (2011). Forward masking in the amplitude-modulation domain for tone carriers: psychophysical results and physiological correlates. *J. Assoc. Res. Otolaryngol.* 12, 361–373. doi: 10.1007/s10162-010-0251-2

Yang, X., Jiang, M., and Zhao, Y. (2017). Effects of noise on English listening comprehension among Chinese college students with different learning styles. *Front. Psychol.* 8:1764. doi: 10.3389/fpsyg.2017.01764

Zhang, J., Stuart, A., and Swink, S. (2011). Word recognition in continuous and interrupted noise by monolingual and bilingual speakers. *Can. J. Speech-Lang. Pathol. Audiol.* 35, 322–331.

Zhong, L., Liu, C., and Tao, S. (2019). Sentence recognition for native and non-native English listeners in quiet and babble: effects of contextual cues. *J. Acoust. Soc. Am.* 145, EL297–EL302. doi: 10.1121/1.5097734

Zinszer, B. D., Riggs, M., Reetzke, R., and Chandrasekaran, B. (2019). Error patterns of native and non-native listeners' perception of speech in noise. *J. Acoust. Soc. Am.* 145, EL129–EL135. doi: 10.1121/1.5087271

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

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 17 June 2021
doi: 10.3389/fnins.2021.664985

Common Brain Substrates Underlying Auditory Speech Priming and Perceived Spatial Separation

Junxian Wang[1†], Jing Chen[2,3†], Xiaodong Yang[1], Lei Liu[1], Chao Wu[4], Lingxi Lu[5], Liang Li[1,3,6] and Yanhong Wu[1,3]*

[1] School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China, [2] Department of Machine Intelligence, Peking University, Beijing, China, [3] Speech and Hearing Research Center, Key Laboratory on Machine Perception (Ministry of Education), Peking University, Beijing, China, [4] School of Nursing, Peking University, Beijing, China, [5] Center for the Cognitive Science of Language, Beijing Language and Culture University, Beijing, China, [6] Beijing Institute for Brain Disorders, Beijing, China

Under a "cocktail party" environment, listeners can utilize prior knowledge of the content and voice of the target speech [i.e., auditory speech priming (ASP)] and perceived spatial separation to improve recognition of the target speech among masking speech. Previous studies suggest that these two unmasking cues are not processed independently. However, it is unclear whether the unmasking effects of these two cues are supported by common neural bases. In the current study, we aimed to first confirm that ASP and perceived spatial separation contribute to the improvement of speech recognition interactively in a multitalker condition and further investigate whether there exist intersectant brain substrates underlying both unmasking effects, by introducing these two unmasking cues in a unified paradigm and using functional magnetic resonance imaging. The results showed that neural activations by the unmasking effects of ASP and perceived separation partly overlapped in brain areas: the left pars triangularis (TriIFG) and orbitalis of the inferior frontal gyrus, left inferior parietal lobule, left supramarginal gyrus, and bilateral putamen, all of which are involved in the sensorimotor integration and the speech production. The activations of the left TriIFG were correlated with behavioral improvements caused by ASP and perceived separation. Meanwhile, ASP and perceived separation also enhanced the functional connectivity between the left IFG and brain areas related to the suppression of distractive speech signals: the anterior cingulate cortex and the left middle frontal gyrus, respectively. Therefore, these findings suggest that the motor representation of speech is important for both the unmasking effects of ASP and perceived separation and highlight the critical role of the left IFG in these unmasking effects in "cocktail party" environments.

Keywords: auditory speech priming, perceived spatial separation, speech recognition, speech motor system, common brain substrate, unmasking

OPEN ACCESS

Edited by:
Howard Charles Nusbaum, The University of Chicago, United States

Reviewed by:
Xing Tian, New York University Shanghai, China
Iiro P. Jääskeläinen, Aalto University, Finland

***Correspondence:**
Yanhong Wu
wuyh@pku.edu.cn

†These authors share first authorship

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

Received: *06 February 2021*
Accepted: *10 May 2021*
Published: *17 June 2021*

Citation:
Wang J, Chen J, Yang X, Liu L, Wu C, Lu L, Li L and Wu Y (2021) Common Brain Substrates Underlying Auditory Speech Priming and Perceived Spatial Separation. Front. Neurosci. 15:664985. doi: 10.3389/fnins.2021.664985

Abbreviations: ACC, anterior cingulate cortex; ANSP, auditory non-speech priming; ASP, auditory speech priming; ASSN, amplitude-modulated speech-spectrum noise; fMRI, functional magnetic resonance imaging; IFG, inferior frontal gyrus; IPL, inferior parietal lobule; MFG, middle frontal gyrus; nPS, ANSP and separation condition; nPC, ANSP and co-location condition; OrbIFG, pars orbitalis of IFG; PC, ASP and co-location condition; PS, ASP and separation condition; PPI, psychophysiological interaction; SMG, supramarginal gyrus; SMR, signal-to-masker ratio; TriIFG, par triangularis of IFG.

INTRODUCTION

How do our brains deal with a complex scene where listeners need to selectively detect, follow, and recognize a speaker's words (target) when multiple people are talking (masker) at the same time (i.e., the "cocktail party" problem) (Cherry, 1953; Schneider et al., 2007; McDermott, 2009; Bronkhorst, 2015)? Previous studies have shown that listeners can take advantage of diverse perceptual and/or cognitive cues, such as prior knowledge of the contents of the speech and/or the speaker's voice (Freyman et al., 2004; Yang et al., 2007), information obtained from lip reading (Wu et al., 2017b), and perceived spatial separation (Freyman et al., 1999), to improve their recognition of target speech masked by non-target sounds (i.e., release from masking or unmasking). In real-life conditions, these unmasking cues are not alone, and several cues belonging to an auditory object influence the perception at the same time. However, a majority of relevant studies usually focused on the cognitive and neural mechanisms of a single one among these unmasking cues (Zheng et al., 2016; Wu et al., 2017a,b), and very few have investigated dual unmasking cues that are closer to a real-life condition (Du et al., 2011; Lu et al., 2018). Therefore, this study employed two unmasking cues in a unified paradigm to investigate neural bases across different unmasking effects.

Among multiple unmasking cues, the prior knowledge of the content and voice of the target speech [i.e., auditory speech priming (ASP)] and perceived spatial separation are two typical and effective cues. The ASP refers to a segment of the target phrase spoken by the target speaker, which is presented without interferences before a target–masker mixture, and it can improve recognition of the last keyword in the target speech, even though the last keyword does not appear in the segment (Freyman et al., 2004; Wu et al., 2012a,b, 2017a). The perceived spatial separation represents spatially separating sound images of target speech from those of competing speech, which can cause even larger unmasking effect on speech recognition (Freyman et al., 1999, 2004; Li et al., 2004; Bronkhorst, 2015). Studies of patients with schizophrenia suggest that ASP and perceived separation share some common features. Even though patients with schizophrenia exhibit deficiencies in speech perception and increased vulnerability to masking stimuli, they retain the ability of using ASP and perceived separation to improve their recognition of the target speech masked by two-talker speech (Wu et al., 2012, 2017a). Similarly, despite age-related declines in hearing, older adults can utilize ASP and perceived separation to improve their recognition of the targets under a noisy environment just as well as do younger adults (Li et al., 2004; Ezzatian et al., 2011). These studies propose that certain top-down auditory mechanisms underlying the unmasking effects of ASP and perceived separation are preserved in patients with schizophrenia and older adults (Li et al., 2004; Ezzatian et al., 2011; Wu et al., 2012, 2017a). Researchers have claimed that the ASP helps listeners maintain the target's voice and content in working memory and may facilitate grouping the target speech to enhance the listener's selective attention to the target (Freyman et al., 2004; Schneider et al., 2007; McDermott, 2009; Ezzatian et al., 2011; Wu et al., 2012a,b; Carlile, 2015; Wang et al., 2019). The perceived separation mainly works via the head-shadowing effect increasing the signal-to-masker ratio (SMR) in the ear close to the target, and/or the neurophysiological effect of disparity in interaural time between targets and maskers, both of which enhance selective attention to the target speech (Freyman et al., 1999; Li et al., 2004). Accordingly, the unmasking effects of ASP and perceived separation may both contribute to the high-level processing, such as enhanced selective attention to the target speech. Moreover, the selective attention might allocate more cognitive resources to the motor representation of speech that is beneficial for speech recognition under "cocktail party" listening conditions (Wu et al., 2014). The intersecting processing of the unmasking effects induced by ASP and perceived separation is predicted to occur in the high-level processing.

Previous studies have no consensus on whether two unmasking cues are processed independently or interdependently. If an additive effect was observed, i.e., the combined effect of two cues is equivalent to the sum of their individual effects, researchers conclude that these two cues are processed independently in separate brain regions (Bronkhorst, 2015). Otherwise, the non-additive effect indicates intersecting processing of these cues in overlapping brain regions. Studies have shown varied additives of two cues; for instance, Du et al. (2011) found an additive effect of the differences in fundamental frequency and spatial location during speech segregation. Lu et al. (2018) found an additive effect of emotional learning (of the target voice) and perceived separation on improving speech recognition. Despite this, Darwin et al. (2003) found that the effect of the differences in fundamental frequency and vocal-tract length is larger than the sum of their individual effects during speech recognition. Freyman et al. (2004) found that the benefit of ASP was more significant when the target and maskers were perceived to be co-located than when they were perceived to be separated, indicating the non-additive effect of ASP and perceived separation. These suggest that the combinations of different cues result in varied additive effects. In the case of ASP and perceived separation, their benefits to speech recognition cannot be added, suggesting the existence of intersections of their neural underpinnings. The present study aims to investigate how ASP and perceived separation improve the recognition of the target speech in a complex scene and reveal the neural bases underlying their unmasking effects.

Despite a lack of studies on neural mechanisms underlying dual unmasking effects, some researchers have separately investigated neural bases of the unmasking effects of ASP and perceived spatial separation. They have found that ASP and perceived separation mainly activated the ventral and dorsal pathways in the auditory system, respectively. Notably, both of them activate the left inferior frontal gyrus (IFG) (Zheng et al., 2016; Wu et al., 2017a). In addition, ASP and perceived separation both enhance the functional connectivity between the IFG and specific brain substrates, such as the left superior temporal gyrus and left posterior middle temporal gyrus for the ASP, and the superior parietal lobule for the perceived separation (Zheng et al., 2016; Wu et al., 2017a). This suggests that the speech motor system, especially the IFG (Du et al., 2014), may be the common brain area that is activated by both the

unmasking effects of ASP and perceived separation. The left IFG is thought to be the speech-related area involved in both speech production and perception (Hickok and Poeppel, 2004; Liakakis et al., 2011). It exchanges information with other brain substrates located in the ventral and dorsal pathways in the auditory system (Friederici, 2012), and it is involved in unification operations during speech comprehension (Acheson and Hagoort, 2013). Moreover, the par triangularis of the IFG (TriIFG) is a junction in the IFG (Friederici, 2012; Frühholz and Grandjean, 2013). A dual stream model for auditory processing suggests that auditory information transferred via both ventral and dorsal streams terminates in the TriIFG (Frühholz and Grandjean, 2013). The TriIFG is thought to be the main locus of speech intelligibility in the IFG (Abrams et al., 2013), and it is especially important in the retrieval or selection of semantic information under adverse listening conditions (Skipper et al., 2007). The present study predicts that the speech motor system plays a critical role in the unmasking effects of ASP and perceived separation. In particular, the TriIFG is the interested brain substrate, and it is predicted to correlate with recognition accuracy of the target speech in a multitalker condition.

In summary, previous studies inconsistently suggest the existence of intersecting processing of dual unmasking cues. In the present study, we introduced ASP and perceived spatial separation in a "cocktail party" listening condition and combined them in a unified paradigm. We aim to first verify that the unmasking effects of ASP and perceived separation are not processed independently and further investigate the existence of their common neural bases, by using functional magnetic resonance imaging (fMRI). The existence of overlapping brain substrates is thought to be a primary investigation of the commonality of neural mechanisms underlying two unmasking effects. In the light of past results, we hypothesized that the unmasking effects of ASP and perceived separation intersect in the high-level processing, and the left IFG is the shared brain substrate critically involved in these unmasking effects.

MATERIALS AND METHODS

Participants

Thirty-six participants (21 females, 15 males; mean age = 22.06 years, SD = 1.94 years; Min_{age} = 19 years, Max_{age} = 27 years) took part in the behavioral testing, and 27 of them (15 females, 12 males; mean age = 22 years, SD = 1.94 years; Min_{age} = 19 years, Max_{age} = 27 years) voluntarily participated in the follow-up fMRI testing. All participants were right-handed university students who spoke Mandarin. All of them had normal pure-tone hearing thresholds ($\leq$25 dB HL) in each ear and had bilaterally symmetric hearing ($\leq$15 dB HL). The hearing thresholds were measured by an audiometer (Aurical, 60645-1, Danmark) at frequencies of 125, 250, 500, 1,000, 2,000, 4,000, 6,000, and 8,000 Hz for 22 participants. The hearing thresholds of the remaining participants were measured by Apple iOS-based automated audiometry (Xing et al., 2016) at frequencies of 250, 500, 1,000, 2,000, 3,000, 4,000, 6,000, and 8,000 Hz in a sound booth. All participants gave their informed consent before the experiment and received a certain monetary compensation. The experimental procedures were approved by the Committee for Protecting Human and Animal Subjects in the School of Psychological and Cognitive Sciences, Peking University.

Stimuli and Procedures

The mean duration of a sound stimulus was 3,711 $\pm$ 341 ms. Each speech stimulus comprised three parts: a priming stimulus [of the ASP condition or auditory non-speech priming (ANSP) condition], a target phrase, and a two-talker speech masker. It started with a priming stimulus, following which a target phrase mixed with a two-talker speech masker was presented. The speech stimuli were processed by head-related transfer functions (Qu et al., 2009) to simulate sounds from one of three azimuth angles (i.e., −90°, 0°, 90°) 30 cm from the center of the listeners' heads in the horizontal plane. The target phrases and the two-talker speech masker were "non-sense" sentences in Chinese that were syntactically, but not semantically, correct. The target phrases, spoken by a young female (talker A), were three-word phrases. Each word of the target phrase contained two syllables. For example, one target phrase translated into English was "contest this employee" (keywords are underlined). The structure of these target phrases did not support any context for recognizing the keywords. There was no overlap in target phrases between behavioral and fMRI tests for each participant. In the fMRI testing, we used a minority of response trials as probes to monitor whether participants were doing a speech-recognition task. Target phrases of these response trials were modified by substituting the last keyword of the original target phrase with the first keyword of it, so that the first and the last keywords were identical.

Considering that the benefits of ASP and perceived separation are more significant when masked by the two-talker speech masker than other types of maskers (Freyman et al., 2004; Lu et al., 2018), we used the two-talker speech masker as the distractive sounds. The masker was a 47-s loop of a digitally combined continuous recording of Chinese non-sense sentences spoken by two young females (talkers B and C). No keyword in the speech masker appeared in target phrases. In the behavioral testing, the sound pressure level of the target phrases was fixed at 56 dBA SPL, and the sound levels of the maskers were adjusted to produce four SMRs: −12, −8, −4, and 0 dB. In the fMRI testing, the SMR was fixed at −4 dB (Wu et al., 2017a,b). All sound pressure levels were measured by an Audiometer Calibration and Electroacoustic Testing System (AUDit and System 824; Larson Davis, Provo, UT, United States). The SMRs were calibrated before applying the head-related transfer function.

The priming stimuli were manipulated differently in the ASP and the ANSP conditions (see the illustration in **Figure 1**). In the ASP condition, a priming stimulus was identical to a target phrase except that the last keyword of the target was replaced by a piece of white noise. The duration of the white noise matched that of the longest last keyword across all target phrases. The sound pressure level of the white noise was 10 dB lower than that of the corresponding target phrase (following Freyman et al., 2004) to ensure its perceived loudness being consistent. In the ANSP

condition, a priming stimulus was produced as follows: The target segment containing the first two words of the target phrase was time reversed and connected to the same white noise as described above. The acoustic property (i.e., long-term spectrum, mean pitch) of this priming stimulus was close to that used in the ASP condition, but without semantic information.

Because the continuity and certainty of location help identify a target object and release it from masking (Kidd et al., 2005; Best et al., 2008), the perceived locations of the priming stimuli and the target phrases were fixed across conditions at the azimuthal 0° to prevent listeners from switching their attention along the spatial dimension. In the perceived separation condition, the masker was simulated from an azimuth of −90° or 90°. In the perceived co-location condition, all of the priming stimuli, target phrases, and maskers were simulated from the azimuthal 0°.

During the fMRI testing, we employed amplitude-modulated speech-spectrum noises (ASSNs) as the baseline condition. The ASSNs were obtained by (1) a fast Fourier transform algorithm returning the discrete Fourier transform of the original speech signals and then randomizing phases of all spectral components, (2) an inverse discrete Fourier transform to convert these modified Fourier outputs back into the time domain, (3) normalized amplitudes of these speech-spectrum noises, and (4) modulated speech-spectrum noises according to the amplitudes of the original speech signals extracted by a Hilbert transformation. As a result, the ASSN of the same length as the original speech had temporal fluctuations and spectra close to those of the speech stimuli, but sounded like noise.

fMRI Testing

The fMRI testing featured two intraparticipant variables: (1) priming type (ASP, ANSP) and (2) perceived laterality relationship (separation, co-location). The combination of them led to four conditions: ASP and separation condition (i.e., the PS condition), ASP and co-location condition (i.e., the PC condition), ANSP and separation condition (i.e., the nPS condition), and ANSP and co-location condition (i.e., the nPC condition).

In functional runs, we used sparse temporal sampling to reduce the influence of machine noises (Hall et al., 1999; Zheng et al., 2016; Wu et al., 2017a,b). Consequently, the sounds were presented between scans. In each trial of the functional runs (see the illustration in **Figure 1**), a sound stimulus was presented 800 ms after the end of scanning of the previous trial so that the midpoint of presentation of the stimulus occurred approximately 4,300 ms before the onset of scanning for the given trial. This was done to ensure that the sound stimulus–evoked hemodynamic responses peaked during the scanning period (Wild et al., 2012a; Zheng et al., 2016; Wu et al., 2017a,b). A speech stimulus started with a priming stimulus that presented without interferences, following which a target phrase mixed with a two-talker speech was presented (the target phrase occurred 1,000 ± 200 ms after the onset of the masker, and they were terminated simultaneously). The participants were instructed to carefully listen to the sounds and press a key immediately when they determined that the first and the last keywords of the target phrase were identical (i.e., when they encountered response trials, they were required to make a key press).

The SMR was fixed at the −4 dB as the ASP was significantly advantageous in both the perceived separation and perceived co-location conditions in this SMR (see *Results*). We used the ASSN baseline condition as the controlling condition for four speech stimuli conditions, and the response trials as probes to monitor whether participants were doing a speech-recognition task. In each functional run, there were 10 trials for each of the five listening conditions (PS, PC, nPS, nPC, and baseline), six response trials, and 10 blank trials (no sound was presented between scans). These trials were randomly presented. A whole-course scanning for each participant consisted of an 8-minute structural scanning and four 10-minute functional runs. Before the formal testing, a training session was conducted outside the scanning room to ensure that participants understood their tasks in the fMRI testing. The stimuli were delivered via a visual and audio stimulation system for fMRI (Sinorad, Shenzhen, China), driven by MATLAB (MathWorks Inc., Natick, MA, United States). The sounds were bilaterally presented to listeners using a pair of MRI-compatible headphones (Sinorad, Shenzhen, China). Foam earplugs (Noise Reduction Rating: 29 dB; 3 M, 1100, Maplewood, United States) were provided to reduce scanner noises.

fMRI Equipment

A 3.0-T Siemens Prisma MRI scanner (Siemens, Erlangen, Germany) was used to acquire blood oxygenation level–dependent gradient echo-planar images (spatial resolution: 112 × 112 × 62 matrix with a voxel size of 2 mm × 2 mm × 2 mm; acquisition time: 2,000 ms; time to repeat: 9,000 ms; echo time: 30 ms; flip angle: 90°; FoV: 224 mm × 224 mm). It provided high-resolution T1-weighted structural images (448 × 512 × 192 matrix with a voxel size of 0.5 mm × 0.5 mm × 1 mm; repetition time: 2,530 ms; echo time: 2.98 ms; flip angle: 7°).

fMRI Data Processing and Analysis

Preprocessing

The fMRI data were analyzed by Statistical Parametric Mapping software (SPM12[1]). The preprocessing of the fMRI data consisted of the following steps: (1) the functional images were realigned to correct head movements. (2) T1-weighted structural images were coregistered with the mean realigned functional images. (3) T1-weighted images were segmented and normalized to the ICBM space template. (4) All functional images were warped by deformation parameters generated in the segmentation process. (5) All functional images were smoothed by a Gaussian kernel with 6-mm full width at half maximum. Because the repetition time was relatively long in such a sparse imaging paradigm, no slice timing was conducted in preprocessing (Wild et al., 2012a; Zheng et al., 2016; Wu et al., 2017a,b).

[1]http://www.fil.ion.ucl.ac.uk/spm/

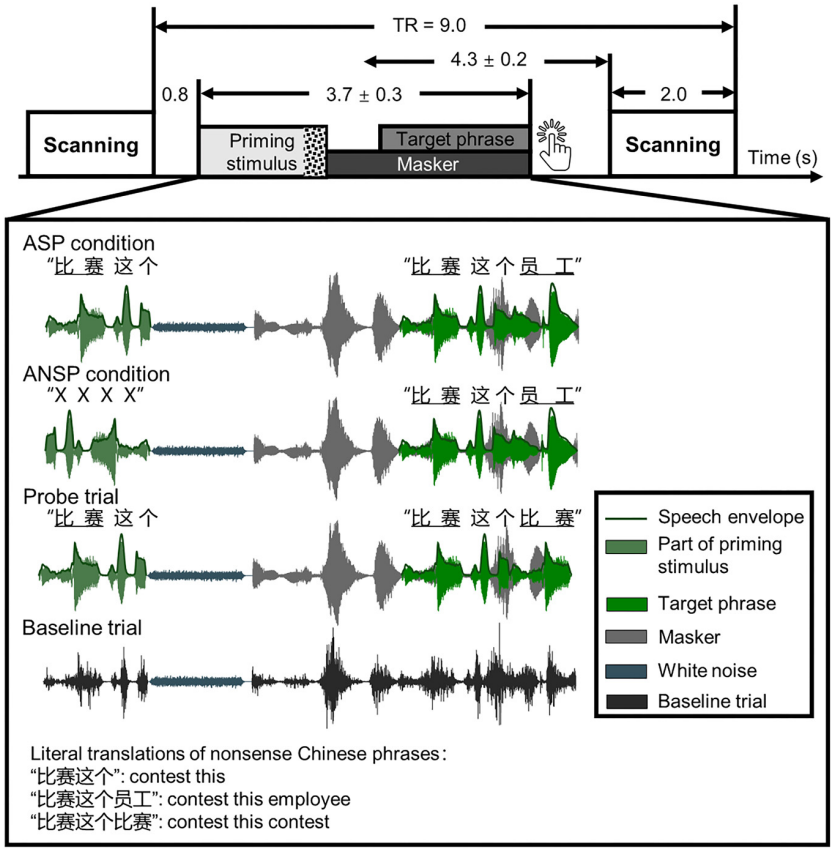

FIGURE 1 | Schematic of a single trial presenting sound stimuli in a functional run. Sparse temporal sampling was used. Four kinds of sound structures of the functional runs are illustrated separately. The temporal midpoint of the sound stimulus was 4,300 ms prior to the onset of scanning for a given trial. The unit of time in this figure is second(s). TR, time to repeat.

Random-Effects Analysis and *Post hoc* Tests

The first-level general linear model for each participant contained seven regressors in total: five for the listening conditions (PS, PC, nPS, nPC, and baseline), one for blank trials, and one for uninteresting/irrelevant response trials. Six realignment parameters of head movements were entered to account for residual movement-related effects. The blood oxygenation level–dependent response for each event was modeled using the canonical hemodynamic response function. Contrast images of "PS > baseline," "PC > baseline," "nPS > baseline," and "nPC > baseline" for each participant, were entered into a 2 (priming type: ASP, ANSP) × 2 (perceived laterality relationship: separation, co-location) repeated-measure analysis of variance (ANOVA). Only clusters passing the cluster-level family wise error correction for multiple comparisons (we set at FWE corrected cluster-level threshold of $p_{FWE} < 0.05$) were entered into the *post hoc* tests. To identify the causes of main effects, *post hoc* paired-samples *t* tests were conducted by inclusively masking specific *t*-contrast images with the corresponding *F* contrasts ($p < 0.001$ at the voxel level, uncorrected). We calculated the ASP effect (i.e., "ASP > ANSP" contrast) by the formula (PS + PC) > (nPS + nPC) and the spatial unmasking effect (i.e., "perceived separation > perceived co-location" contrast) by the formula (PS + nPS) > (PC + nPC). Moreover, referring to the method in Wild et al. (2012a), we calculated the logical intersections of clusters activated by both the ASP effect and the spatial unmasking effect. Overlapping clusters containing more than 10 voxels were reported and used as regions of interest for behavioral–neural correlation analyses.

Correlation Analysis

To identify the correlation between brain activations and behavioral improvements by the ASP effect and the spatial unmasking effect, we conducted Spearman correlation analyses by using the IBM SPSS.20 software. As we have introduced in the *Introduction*, the left TriIFG is important in semantic processing of speech under adverse listening conditions. We used the left TriIFG that was activated by both unmasking effects as the region of interest in the correlation analyses and extracted contrast values within this overlapping cluster by the MarsBar toolbox (version 0.44[2]). The correlations for four contrasts (i.e., "PS > nPS," "PC > nPC," "PS > PC," and "nPS > nPC") were examined by using corresponding contrast values and behavioral improvements at the −4 dB SMR.

[2]http://marsbar.sourceforge.net/

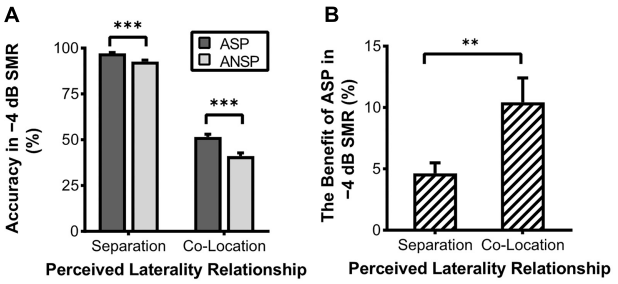

FIGURE 2 | Accuracy of recognition of the last keyword in target phrases under the −4 dB SMR in the behavioral testing. **(A)** Comparisons of recognition accuracy under different priming types and perceived laterality relationships. **(B)** The enhanced recognition accuracy obtained through ASP against ANSP under different relationships of perceived laterality. The error bars represent standard errors. ***$p < 0.001$, **$p < 0.01$.

Psychophysiological Interaction Analysis

We conducted generalized form of context-dependent psychophysiological interaction (PPI) analyses (McLaren et al., 2012) to identify brain regions showing significant functional connectivity with the brain substrates of interest (i.e., the left IFG), by using the gPPI Toolbox (version 7.12[3], based on SPM8). Each seed region was defined as a sphere with a 6-mm radius centered at the peak voxel. The first-level generalized PPI model contained all regressors in a first-level general linear model, additional PPI regressors (for the PS, PC, nPS, and nPC conditions, noise baseline trials, blank trials, and response trials), time courses of the seed region, and a constant (McLaren et al., 2012). Single-participant contrast images ("PS > nPS," "PC > nPC," "PS > PC," and "nPS > nPC" contrasts) of the first-level generalized PPI model were subjected to second-level one-sample *t* tests to identify brain regions showing increased coactivation with the seed region due to corresponding contrasts. The level of significance was set at $p < 0.001$ uncorrected with an extent threshold with minimum cluster size of 20 voxels (Wandschneider et al., 2014).

Behavioral Testing

Behavioral testing was conducted before the fMRI testing, and it featured three intraparticipant variables: (1) priming type (ASP, ANSP), (2) perceived laterality relationship (separation, co-location), and (3) SMR (−12, −8, −4, 0 dB). Variables for the priming type and perceived laterality relationship were the same as in the fMRI testing.

In each trial, a speech stimulus started with a priming stimulus, following which a target phrase mixed with a two-talker speech masker was presented. The onset of the target phrase was 1,000 ± 200 ms later than that of the masker, and they were terminated simultaneously. Participants started each trial by themselves, and they were instructed to verbally repeat the entire target phrase as much as possible when the sounds ended in each trial. The experimenter scored and calculated the number of correctly recognized syllables of the last keywords (participants were not aware that only the last keywords were scored). Each syllable of the last keyword was counted as one point. Before the formal testing, a training session was conducted to ensure that participants had understood the task of the behavioral testing.

Four combinations of priming types and perceived laterality relationships were assigned to four blocks, and their orders of presentation were counterbalanced across participants by using the Latin square order. The four SMRs were randomly ordered in each block. For each participant, 12 trials (12 target phrases) were conducted for each of 16 conditions. Sounds were binaurally presented to participants via the headphones (HD 650, Sennheiser electronic GmbH & Co., KG, Germany), driven by Presentation software (version 0.70).

RESULTS

Behavioral Improvements Due to ASP and Perceived Spatial Separation

In the behavioral testing, the 2 (priming type: ASP, ANSP) × 2 (perceived laterality relationship: separation, co-location) × 4 (SMR: −12, −8, −4, 0 dB) repeated-measured ANOVA showed significant main effects of priming type ($F_{1,\ 35} = 34.98$, $p < 0.001$, $\eta_p{}^2 = 0.50$), perceived laterality relationship ($F_{1,\ 35} = 2{,}539.91$, $p < 0.001$, $\eta_p{}^2 = 0.99$), and SMR ($F_{3,\ 105} = 1{,}127.50$, $p < 0.001$, $\eta_p{}^2 = 0.97$). Their interaction was also significant ($F_{3,\ 105} = 7.33$, $p < 0.001$, $\eta_p{}^2 = 0.17$). Simple-effect analyses (Bonferroni-corrected) showed that the ASP effect was not consistently significant between the two perceived laterality relationships (detailed results are provided in **Supplementary Table 1**). Only when the SMR was −4 dB did the ASP contrast to ANSP improve the recognition of the target under both the perceived separation (PS:0.97 ± 0.02, nPS:0.93 ± 0.05, $F_{1,\ 35} = 29.05$, $p < 0.001$) and the perceived co-location (PC:0.51 ± 0.09, nPC:0.41 ± 0.10, $F_{1,\ 35} = 27.44$, $p < 0.001$) conditions (see **Figure 2A**). In the −4 dB SMR, the benefits of ASP (recognition accuracy of ASP minus that of ANSP) between the two perceived laterality relationships were compared by paired-sample *t* tests. The results showed that the benefit of ASP under the perceived co-location condition (i.e., recognition accuracy of PC minus that of nPC; the benefit of ASP:0.10 ± 0.12) was greater than that under the perceived separation condition (i.e., recognition accuracy of PS minus nPS; the benefit of ASP:0.05 ± 0.05) ($t_{35} = 3$, $p = 0.005$, see **Figure 2B**). These results indicate that the unmasking effect of ASP decreases when the perceived spatial separation helps improve the recognition of the target speech.

Brain Regions Activated by ASP and Perceived Spatial Separation

A 2 (priming type: ASP, ANSP) × 2 (perceived laterality relationship: separation, co-location) repeated-measured ANOVA was conducted, and no suprathreshold cluster was activated by the interaction. We marked out clusters activated by the ASP effect (i.e., "ASP > ANSP") from these by the main effect of priming type, using a voxel-wise threshold of

[3]http://www.nitrc.org/projects/gppi/

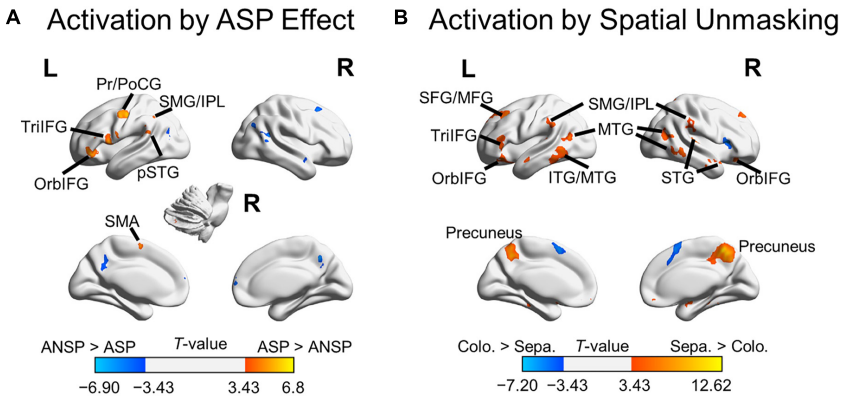

FIGURE 3 | Areas of the brain activated by the ASP effect and spatial unmasking effect. **(A)** Clusters activated by the "ASP > ANSP" contrast (i.e., the ASP effect, the warm color) and by the "ANSP > ASP" contrast (the cold color). **(B)** Clusters activated by the "perceived separation > perceived co-location" contrast (i.e., spatial unmasking effect, the warm color) and by the "perceived co-location > perceived separation" contrast (the cold color). ITG, inferior temporal gyrus; MTG, middle temporal gyrus; PrCG, precentral gyrus; PoCG, postcentral gyrus; SFG, superior frontal gyrus; SMA, supplementary motor area; STG, superior temporal gyrus; pSTG, posterior STG. L, left; R, right.

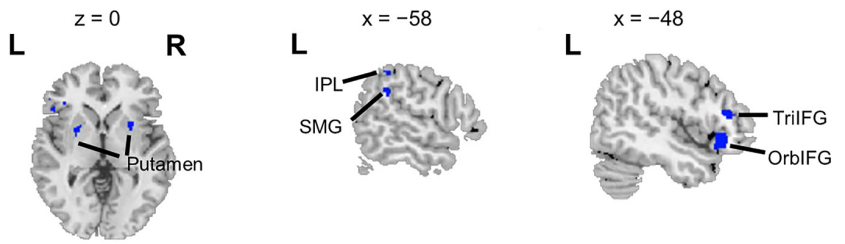

FIGURE 4 | Overlapping brain areas activated by both the ASP effect and the spatial unmasking effect. L, left; R, right. Negative values of the *x* axis (the coronal axis) denote the left hemisphere. The variable *z* denotes the vertical axis.

$p < 0.001$, uncorrected (clusters activated by the main effect of priming type are in both warm and cold colors in **Figure 3A**). The ASP effect activated the left motor cortex, left IFG, left posterior superior temporal gyrus, left inferior parietal region, bilateral putamen, and the right cerebellum (as shown in warm color clusters in **Figure 3A**). Meanwhile, we marked out clusters activated by the spatial unmasking effect (i.e., "perceived separation > perceived co-location") from these by the main effect of perceived laterality relationship, using a voxel-wise threshold of $p < 0.001$, uncorrected (clusters activated by the main effect of perceived laterality relationship are in both warm and cold colors in **Figure 3B**). The spatial unmasking effect activated the bilateral precuneus, bilateral IFG, bilateral middle temporal gyrus, bilateral superior frontal gyrus extending into middle frontal gyrus (MFG), bilateral inferior parietal region, bilateral superior temporal gyrus, and bilateral putamen (as shown in warm color clusters in **Figure 3B**). The detailed results are also shown in **Supplementary Table 2**.

Overlapping brain areas activated by both the ASP effect and the spatial unmasking effect were the bilateral putamen, left inferior parietal lobule (IPL), left supramarginal gyrus (SMG), left pars orbitalis of the IFG (OrbIFG), and left TriIFG (**Figure 4**), revealing shared neural bases of the unmasking effects of ASP and perceived separation.

Correlations Between Brain Activation and Behavioral Improvement Due to ASP and Perceived Spatial Separation

Within the left TriIFG, the overlapping brain area activated by both the ASP effect and the spatial unmasking effect, the Spearman correlation analyses showed that contrast values for the ASP effect under the perceived separation condition (i.e., "PS > nPS" contrast) were significantly correlated with behavioral improvements for the corresponding contrast ($r = -0.55$, $p < 0.01$, **Figure 5A**), and contrast values for the spatial unmasking effect under the ASP condition (i.e., "PS > PC" contrasts) were also significantly correlated with behavioral improvements for the corresponding contrast (PS > PC: $r = 0.43$, $p = 0.03$, **Figure 5B**). The sizes of these correlations were moderate. No significant correlation was observed for the ASP effect under the perceived co-location condition (i.e., "PC > nPC" contrast) or the spatial unmasking effect under the ANSP condition (i.e., "nPS > nPC" contrast).

Enhanced Functional Connectivity Due to ASP and Perceived Spatial Separation

The seed regions were located in the left IFG. As shown in **Figure 6A** and **Table 1**, the ASP effect under the perceived

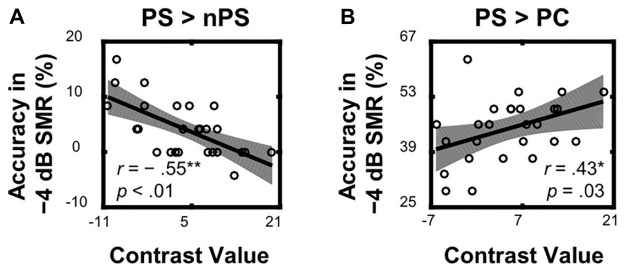

FIGURE 5 | Correlations between neural activities of the overlapping left TriIFG and target recognition accuracy under a −4 dB SMR in behavioral testing. **(A)** Contrast values of the ASP effect under the perceived separation condition ("PS > nPS" contrast) were correlated with the corresponding contrasts of recognition accuracy. **(B)** Contrast values of the spatial unmasking effect under the ASP condition ("PS > PC" contrast) were correlated with the corresponding contrasts of recognition accuracy. The shades represent 95% confidence intervals.

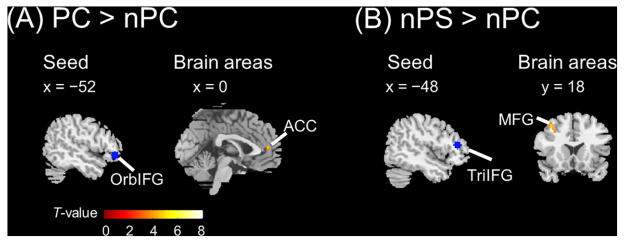

FIGURE 6 | Brain areas exhibiting significant functional connectivity with the left IFG by the ASP effect and spatial unmasking effect. **(A)** The ASP effect under the perceived co-location condition ("PC > nPC" contrast) enhanced the functional connectivity between the left OrbIFG and the bilateral ACC. **(B)** The spatial unmasking effect under the ANSP condition ("nPS > nPC" contrast) enhanced the functional connectivity between the left TriIFG and the left MFG. All peaks survived an uncorrected threshold of $p < 0.001$ at the voxel level ($t > 3.43$), and each cluster consisted of more than 20 contiguous voxels. The positive value of x (the coronal axis) denotes the right hemisphere. Negative values of the x axis (the coronal axis) denote the left hemisphere. Positive values of y (the sagittal axis) denote the front. The blue areas are seed regions.

co-location condition (i.e., "PC > nPC" contrast) enhanced the functional connectivity of the left OrbIFG [the seed locus at (−52, 38, −6)] with the bilateral anterior cingulate cortex (ACC). Meanwhile, the spatial unmasking effect under the ANSP condition (i.e., "nPS > nPC" contrast) enhanced the functional connectivity of the left TriIFG [the seed locus at (−48, 36, 14)] with the left MFG (**Figure 6B** and **Table 1**). Clusters that survived an uncorrected threshold of $p < 0.001$ at the voxel level with more than 20 contiguous voxels are reported.

DISCUSSION

In the present study, we introduced ASP and perceived spatial separation to investigate how they improve recognition of the target speech in a "cocktail party" environment. The behavioral results showed that the benefits of ASP and perceived separation for speech recognition cannot be added. Neuroimaging results showed that neural underpinnings underlying the unmasking effects of ASP and perceived separation partly overlapped in brain areas related to the speech motor system, especially in the left IFG. These findings suggest the intersection of neural bases underlying two unmasking effects.

The behavioral results were consistent with previous findings that recognizing the target speech masked by two-talker speech can be improved by the prior knowledge of an early segment of the target speech and the perceived spatial separation (Freyman et al., 1999, 2004; Li et al., 2004; Wu et al., 2005, 2012, 2017a; Yang et al., 2007; Ezzatian et al., 2011; Zheng et al., 2016). Moreover, even though ASP improved the speech recognition in two perceived laterality relationships under a moderate degree of masking (such as the −4 dB in this study), the benefit of ASP was more significant when the maskers were perceived to be co-located with the target than when they were perceived to be separated from the target (Freyman et al., 2004). That is, we observed the non-additive benefits of ASP and perceived separation to speech recognition. Considering the additive effect of two cues suggesting independent processing of them (Du et al., 2011; Bronkhorst, 2015; Lu et al., 2018), the behavioral results suggest intersecting processing of ASP and perceived spatial separation.

The above inference is supported by remarkable neuroimaging results, which showed overlaps in neural underpinnings underlying the unmasking effects of ASP and perceived spatial separation. These intersectant brain areas included bilateral putamen, left IPL, left SMG, left OrbIFG, and left TriIFG. The common feature of these brain substrates is their involvement in the speech motor representation (for putamen, Abutalebi et al., 2013; for SMG, Deschamps et al., 2014; for IPL, Hickok and Poeppel, 2000; Shum et al., 2011; for IFG, Klaus and Hartwigsen, 2019), indicating a notable and common role of the speech motor representation for different unmasking effects in a noisy environment. This finding is in accordance with the idea that the auditory-to-motor transformation is important for speech perception in challenging listening situations (Price, 2010; D'Ausilio et al., 2012; Wu et al., 2014) and suggests that the enhancement of the speech motor representation is possibly the common neural mechanism underlying both the unmasking effects of ASP and perceived separation in a multitalker condition. Specifically, the inferior parietal region, containing the IPL and the SMG, is an important component of the network for the sensorimotor integration of speech. The sensorimotor integration is likely to be an emulation process for speech perception in challenging scenarios (Nuttall et al., 2016). It is assumed that covert emulation is processed in parallel with external events through the generation of top-down predictions of ongoing listening events, and perceptual processing is modulated by the feedback generated by these predictions (Wilson and Knoblich, 2005; Hickok et al., 2011; Nuttall et al., 2016). Interestingly, previous studies have discovered mirror neurons in the IPL and the left IFG, even though these studies focused on how actions are processed (Fogassi et al., 2005; Chong et al., 2008; Kilner et al., 2009). In consideration of the roles of IPL and IFG in this study, it can be proposed that the assumed perceptual emulator beneficial for challenging speech perception

TABLE 1 | Areas of the brain exhibiting significant functional connectivity associated with the ASP effect under the perceived co-location condition ("PC > nPC" contrast) and with the spatial unmasking effect under the ANSP condition ("nPS > nPC" contrast).

Contrast	Seed	MNI coordinates (mm)			Statistics				Location
		x	*y*	*z*	*k*	*t*	*z* value	p_{unc}	
PC > nPC	L. OrbIFG	−2	48	10	26	5.24	4.29	9.02E−06	L. ACC
nPS > nPC	L. TriIFG	−38	18	48	66	4.98	4.14	1.77E−05	L. MFG
		−34	20	38		4.05	3.53	2.07E−04	L. MFG

The activation reported here survived an uncorrected threshold of $p < 0.001$ at the voxel level ($t > 3.43$), and the clusters consisted of more than 20 contiguous voxels. The MNI coordinates, k (the number of voxels), t value, z score, and uncorrected p values are provided. L, left.

(Wilson and Knoblich, 2005) may be also in these two areas. In short, the involvement of inferior parietal region (i.e., the IPL and the SMG) in both the unmasking effects of ASP and perceived separation suggests that the two unmasking cues might improve recognition of the target speech by promoting the sensorimotor integration of the target in a noisy environment.

This study also underlines the notable role of the left IFG in the unmasking effects of ASP and perceived spatial separation. The left IFG is critically involved in the speech motor representation, typically the speech production (Liakakis et al., 2011; Du et al., 2014; Klaus and Hartwigsen, 2019). Previous studies suggest that the left IPL contributes to the phonological storage mechanism, while the left frontal network supports subvocal rehearsal and articulatory representation (Hickok and Poeppel, 2000; Hickok et al., 2011). In the present study, the left IFG was not only activated by both unmasking effects, but was also significantly correlated with behavioral improvements caused by them. These results suggest that the improvement of speech recognition due to ASP and perceived separation may be substantially in charge of enhanced subvocal rehearsal and covert speech production. These processes might be especially beneficial to improve speech intelligibility under adverse listening conditions. Apart from this, both ASP and perceived separation enhanced the functional connectivity of the left IFG with brain areas related to the suppression of distractive speech signals (i.e., the ACC and the MFG, respectively; for the ACC, Botvinick et al., 2004; Orr and Weissman, 2009; Kim et al., 2013; for the MFG, Bledowski et al., 2004). Even though the seeds were not identical, both the left OrbIFG and the left TriIFG receive inputs from temporal cortex to support the semantic processing of speech (Friederici, 2012). The results indicate that both ASP and perceived separation can facilitate the collaboration of the semantic processing to the target speech and the suppression to non-target speech. Previous studies have reported that the left IFG plays important roles in the high-level processing of speech, including recovering words and meaning from an impoverished speech signal (Wild et al., 2012b), unifying speech information (Hagoort, 2005; Acheson and Hagoort, 2013), and semantic selection in semantic competition situations (Grindrod et al., 2008; Whitney et al., 2011, 2012). Accordingly, when dealing with complex scenes, the left IFG may be an integration node for high-level processes, such as selecting features of the target speech from competing information, binding features related to the target speech into a consolidated target object, and allocating selective attention to the target speech. It is supposed that when enhancing certain features related to the target speech (e.g., spatial location, the voice, and the content of speech sounds), the overall salience of the target may increase among non-targets. The specific attentional mechanisms and binding processing should be examined in the future work.

In addition to overlapping brain areas, ASP and perceived spatial separation activated brain areas specific to themselves. The results showed that the ASP effect activated motor cortices, including the left precentral gyrus, left postcentral gyrus, and left supplementary area. These areas are involved in the higher-order speech motor representation, such as articulatory planning and execution and predictive speech motor control (Pulvermüller et al., 2006; Connie et al., 2016; Hertrich et al., 2016). The spatial unmasking effect activated the MFG and the precuneus, which have also been reported in Zheng et al. (2016). Both the MFG and the precuneus are involved in the spatial processing, such as spatial attention control and spatial computations in a complex listening environment (for the MFG, Giesbrecht et al., 2003; for the precuneus, Shomstein and Yantis, 2006; Wu et al., 2007; Krumbholz et al., 2009; Zündorf et al., 2013). Furthermore, the spatial unmasking effect activated the inferior and middle temporal gyri responsible for sound-to-meaning representation (Buckner et al., 2000; Hickok and Poeppel, 2007; Simanova et al., 2014). Thus, ASP and perceived separation were also processed by their cue-specific brain substrates. Specifically, the ASP featured its neural underpinnings in the motor cortex, whereas the perceived separation involved dual pathways in the auditory system. Besides, their individualities may also reflect in the suppression of distractive speech signals. The ASP enhanced the functional connectivity of the ACC with the left IFG, but its main effect did not involve the ACC. While the perceived separation enhanced the functional connectivity of the MFG with the left IFG, its main effect also activated the MFG. Considering the ACC is involved in monitoring conflicts between targets and distractors (Botvinick et al., 2004), and the MFG is involved in the distractor processing (Bledowski et al., 2004), these results may indicate that the ASP enhances the suppression of distractors by allocating cognitive resources between targets and distractors. By contrast, along with this process, the perceived separation may also directly crack down on the distractor processing. Future work in this research area should explore the relationships between the commonality and the individuality of

these unmasking effects. They likely work together to improve speech recognition under a complex listening environment.

In conclusion, our study introduced ASP and perceived spatial separation in a multitalker acoustic environment to investigate common neural bases underlying their unmasking effects. We observed that the benefits of perceived separation on speech recognition could not add to that of ASP, and the speech motor system, especially the left IFG, was responsible for both the unmasking effects of ASP and perceived separation. These findings suggest that the unmasking effects of ASP and perceived separation are not independently processed, and there exist common neural bases supporting both the unmasking effects of ASP and perceived separation.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the data were identifiable and not anonymous for each participants. Requests to access the datasets should be directed to LLi, liangli@pku.edu.cn.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Committee for Protecting Human and Animal Subjects at the School of Psychological and Cognitive Sciences at Peking University. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JW: conceptualization, formal analysis, investigation, methodology, software, validation, visualization, writing—original draft, and writing—review and editing. JC: funding acquisition, writing—original draft, and writing—review and editing. XY: methodology. LLiu: formal analysis. CW: formal analysis, methodology, and software. LLu: methodology. LLi: conceptualization, funding acquisition, methodology, project administration, resources, validation, writing—review and editing, and supervision. YW: supervision. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant numbers 31771252 and 12074012).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.664985/full#supplementary-material

REFERENCES

Abrams, D. A., Ryali, S., Chen, T., Balaban, E., Levitin, D. J., and Menon, V. (2013). Multivariate activation and connectivity patterns discriminate speech intelligibility in Wernicke's, Broca's, and Geschwind's areas. *Cereb. Cortex* 23, 1703–1714. doi: 10.1093/cercor/bhs165

Abutalebi, J., Della Rosa, P. A., Gonzaga, A. K., Keim, R., Costa, A., and Perani, D. (2013). The role of the left putamen in multilingual language production. *Brain Lang.* 125, 307–315. doi: 10.1016/j.bandl.2012.03.009

Acheson, D. J., and Hagoort, P. (2013). Stimulating the brain's language network: syntactic ambiguity resolution after TMS to the inferior frontal gyrus and middle temporal gyrus. *J. Cogn. Neurosci.* 25, 1664–1677. doi: 10.1162/jocn_a_00430

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

Bledowski, C., Prvulovic, D., Goebel, R., Zanella, F. E., and Linden, D. E. (2004). Attentional systems in target and distractor processing: a combined ERP and fMRI study. *Neuroimage* 22, 530–540. doi: 10.1016/j.neuroimage.2003.12.034

Botvinick, M. M., Cohen, J. D., and Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends Cogn. Sci.* 8, 539–546. doi: 10.1016/j.tics.2004.10.003

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Atten. Percept. Psychophys.* 77, 1465–1487. doi: 10.3758/s13414-015-0882-9

Buckner, R. L., Koutstaal, W., Schacter, D. L., and Rosen, B. R. (2000). Functional MRI evidence for a role of frontal and inferior temporal cortex in amodal components of priming. *Brain* 123, 620–640. doi: 10.1093/brain/123.3.620

Carlile, S. (2015). Auditory perception: attentive solution to the cocktail party problem. *Curr. Biol.* 25, R757–R759. doi: 10.1016/j.cub.2015.07.064

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

Chong, T. T. J., Cunnington, R., Williams, M. A., Kanwisher, N., and Mattingley, J. B. (2008). fMRI adaptation reveals mirror neurons in human inferior parietal cortex. *Curr. Biol.* 18, 1576–1580. doi: 10.1016/j.cub.2008.08.068

Connie, C., Hamilton, L. S., Keith, J., and Chang, E. F. (2016). The auditory representation of speech sounds in human motor cortex. *Elife* 5, e12577. doi: 10.7554/eLife.12577

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

D'Ausilio, A., Bufalari, I., Salmas, P., and Fadiga, L. (2012). The role of the motor system in discriminating normal and degraded speech sounds. *Cortex* 48, 882–887. doi: 10.1016/j.cortex.2011.05.017

Deschamps, I., Baum, S. R., and Gracco, V. L. (2014). On the role of the supramarginal gyrus in phonological processing and verbal working memory: evidence from rTMS studies. *Neuropsychologia* 53, 39–46. doi: 10.1016/j.neuropsychologia.2013.10.015

Du, Y., Buchsbaum, B. R., Grady, C. L., and Alain, C. (2014). Noise differentially impacts phoneme representations in the auditory and speech motor systems. *Proc. Natl. Acad. Sci. U. S. A.* 111, 7126–7131. doi: 10.1073/pnas.1318738111

Du, Y., He, Y., Ross, B., Bardouille, T., Wu, X., Li, L., et al. (2011). Human auditory cortex activity shows additive effects of spectral and spatial cues during speech segregation. *Cereb. Cortex* 21, 698–707. doi: 10.1093/cercor/bhq136

Ezzatian, P., Li, L., Pichora-Fuller, K., and Schneider, B. A. (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

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., and Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science* 308, 662–667. doi: 10.1126/science.1106138

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., Helfer, K. S., Mccall, D. D., and Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *J. Acoust. Soc. Am.* 106, 3578–3588. doi: 10.1121/1.428211

Friederici, A. D. (2012). The cortical language circuit: from auditory perception to sentence comprehension. *Trends Cogn. Sci.* 16, 262–268. doi: 10.1016/j.tics.2012.04.001

Frühholz, S., and Grandjean, D. (2013). Processing of emotional vocalizations in bilateral inferior frontal cortex. *Neurosci. Biobehav. Rev.* 37, 2847–2855. doi: 10.1016/j.neubiorev.2013.10.007

Giesbrecht, B., Woldorff, M. G., Song, A. W., and Mangun, G. R. (2003). Neural mechanisms of top-down control during spatial and feature attention. *Neuroimage* 19, 496–512. doi: 10.1016/s1053-8119(03)00162-9

Grindrod, C. M., Bilenko, N. Y., Myers, E. B., and Blumstein, S. E. (2008). The role of the left inferior frontal gyrus in implicit semantic competition and selection: an event-related fMRI study. *Brain Res.* 1229, 167–178. doi: 10.1016/j.brainres.2008.07.017

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends Cogn. Sci.* 9, 416–423. doi: 10.1016/j.tics.2005.07.004

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., et al. (1999). Sparse" temporal sampling in auditory fMRI. *Hum. Brain Mapp.* 7, 213–223. doi: 10.1002/(SICI)1097-019319997:3<213::AID-HBM5<3.0.CO;2-N

Hertrich, I., Dietrich, S., and Ackermann, H. (2016). The role of the supplementary motor area for speech and language processing. *Neurosci. Biobehav. Rev.* 68, 602–610. doi: 10.1016/j.neubiorev.2016.06.030

Hickok, G., Houde, J., and Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron* 69, 407–422. doi: 10.1016/j.neuron.2011.01.019

Hickok, G., and Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends Cogn. Sci.* 4, 131–138. doi: 10.1016/S1364-6613(00)01463-7

Hickok, G., and Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92, 67–99. doi: 10.1016/j.cognition.2003.10.011

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402. doi: 10.1038/nrn2113

Kidd, G., Arbogast, T. L., Mason, C. R., and Gallun, F. J. (2005). The advantage of knowing where to listen. *J. Acoust. Soc. Am.* 117, 3804–3815. doi: 10.1121/1.2109187

Kilner, J. M., Neal, A., Weiskopf, N., Friston, K. J., and Frith, C. D. (2009). Evidence of mirror neurons in human inferior frontal gyrus. *J. Neurosci.* 29, 10153–10159. doi: 10.1523/JNEUROSCI.2668-09.2009

Kim, C., Chung, C., and Kim, J. (2013). Task-dependent response conflict monitoring and cognitive control in anterior cingulate and dorsolateral prefrontal cortices. *Brain Res.* 1537, 216–223. doi: 10.1016/j.brainres.2013.08.055

Klaus, J., and Hartwigsen, G. (2019). Dissociating semantic and phonological contributions of the left inferior frontal gyrus to language production. *Hum. Brain Mapp.* 40, 3279–3287. doi: 10.1002/hbm.24597

Krumbholz, K., Nobis, E. A., Weatheritt, R. J., and Fink, G. R. (2009). Executive control of spatial attention shifts in the auditory compared to the visual modality. *Hum. Brain Mapp.* 30, 1457–1469. doi: 10.1002/hbm.20615

Li, L., Daneman, M., Qi, J. G., and Schneider, B. A. (2004). Does the information content of an irrelevant source differentially affect spoken word recognition in younger and older adults? *J. Exp. Psychol. Hum. Percept. Perform.* 30, 1077–1091. doi: 10.1037/0096-1523.30.6.1077

Liakakis, G., Nickel, J., and Seitz, R. J. (2011). Diversity of the inferior frontal gyrus–a meta-analysis of neuroimaging studies. *Behav. Brain Res.* 225, 341–347. doi: 10.1016/j.bbr.2011.06.022

Lu, L., Bao, X., Chen, J., Qu, T., Wu, X., and Li, L. (2018). Emotionally conditioning the target-speech voice enhances recognition of the target speech under "cocktail-party" listening conditions. *Atten. Percept. Psychophys.* 80, 871–883. doi: 10.3758/s13414-018-1489-8

McDermott, J. H. (2009). The cocktail party problem. *Curr. Biol.* 19, R1024–R1027. doi: 10.1016/j.cub.2009.09.005

McLaren, D. G., Ries, M. L., Xu, G., and Johnson, S. C. (2012). A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. *Neuroimage* 61, 1277–1286. doi: 10.1016/j.neuroimage.2012.03.068

Nuttall, H. E., Kennedy-Higgins, D., Hogan, J., Devlin, J. T., and Adank, P. (2016). The effect of speech distortion on the excitability of articulatory motor cortex. *Neuroimage* 128, 218–226. doi: 10.1016/j.neuroimage.2015.12.038

Orr, J. M., and Weissman, D. H. (2009). Anterior cingulate cortex makes 2 contributions to minimizing distraction. *Cereb. Cortex* 19, 703–711. doi: 10.1093/cercor/bhn119

Price, C. J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. *Ann. N. Y. Acad. Sci.* 1191, 62–88. doi: 10.1111/j.1749-6632.2010.05444.x

Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F. M., Hauk, O., and Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proc. Natl. Acad. Sci. U. S. A.* 103, 7865–7870. doi: 10.1073/pnas.0509989103

Qu, T., Xiao, Z., Gong, M., Huang, Y., Li, X., and Wu, X. (2009). Distance-dependent head-related transfer functions measured with high spatial resolution using a spark gap. *IEEE Trans. Audio Speech Lang. Process.* 17, 1124–1132. doi: 10.1109/tasl.2009.2020532

Schneider, B. A., Li, L., and Daneman, M. (2007). How competing speech interferes with speech comprehension in everyday listening situations. *J. Am. Acad. Audiol.* 18, 559–572. doi: 10.3766/jaaa.18.7.4

Shomstein, S., and Yantis, S. (2006). Parietal cortex mediates voluntary control of spatial and nonspatial auditory attention. *J. Neurosci.* 26, 435–439. doi: 10.1523/JNEUROSCI.4408-05.2006

Shum, M., Shiller, D. M., Baum, S. R., and Gracco, V. L. (2011). Sensorimotor integration for speech motor learning involves the inferior parietal cortex. *Eur. J. Neurosci.* 34, 1817–1822. doi: 10.1111/j.1460-9568.2011.07889.x

Simanova, I., Hagoort, P., Oostenveld, R., and Van Gerven, M. A. (2014). Modality-independent decoding of semantic information from the human brain. *Cereb. Cortex* 24, 426–434. doi: 10.1093/cercor/bhs324

Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., and Small, S. L. (2007). Speech-associated gestures. Broca's area, and the human mirror system. *Brain Lang.* 101, 260–277. doi: 10.1016/j.bandl.2007.02.008

Wandschneider, B., Centeno, M., Vollmar, C., Symms, M., Thompson, P. J., Duncan, J. S., et al. (2014). Motor co-activation in siblings of patients with juvenile myoclonic epilepsy: an imaging endophenotype?. *Brain* 137, 2469–2479. doi: 10.1093/brain/awu175

Wang, Y., Zhang, J., Zou, J., Luo, H., and Ding, N. (2019). Prior knowledge guides speech segregation in human auditory cortex. *Cereb. Cortex* 29, 1561–1571. doi: 10.1093/cercor/bhy052

Whitney, C., Kirk, M., O'Sullivan, J., Lambon Ralph, M. A., and Jefferies, E. (2011). The neural organization of semantic control: TMS evidence for a distributed network in left inferior frontal and posterior middle temporal gyrus. *Cereb. Cortex* 21, 1066–1075. doi: 10.1093/cercor/bhq180

Whitney, C., Kirk, M., O'Sullivan, J., Lambon Ralph, M. A., and Jefferies, E. (2012). Executive semantic processing is underpinned by a large-scale neural network: revealing the contribution of left prefrontal, posterior temporal, and parietal cortex to controlled retrieval and selection using TMS. *J. Cogn. Neurosci.* 24, 133–147. doi: 10.1162/jocn_a_00123

Wild, C. J., Davis, M. H., and Johnsrude, I. S. (2012a). Human auditory cortex is sensitive to the perceived clarity of speech. *Neuroimage* 60, 1490–1502. doi: 10.1016/j.neuroimage.2012.01.035

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., and Johnsrude, I. S. (2012b). Effortful listening: the processing of degraded speech depends critically on attention. *J. Neurosci.* 32, 14010–14021. doi: 10.1523/jneurosci.1528-12.2012

Wilson, M., and Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychol. Bull.* 131, 460–473. doi: 10.1037/0033-2909.131.3.460

Wu, C., Cao, S., Zhou, F., Wang, C., Wu, X., and Li, L. (2012). Masking of speech in people with first-episode schizophrenia and people with chronic schizophrenia. *Schizophr. Res.* 134, 33–41. doi: 10.1016/j.schres.2011.09.019

Wu, C., Zheng, Y., Li, J., Wu, H., She, S., Liu, S., et al. (2017a). Brain substrates underlying auditory speech priming in healthy listeners and listeners with schizophrenia. *Psychol. Med.* 47, 837–852. doi: 10.1017/S0033291716002816

Wu, C., Zheng, Y., Li, J., Zhang, B., Li, R., Wu, H., et al. (2017b). Activation and functional connectivity of the left inferior temporal gyrus during visual speech priming in healthy listeners and listeners with schizophrenia. *Front. Neurosci.* 11:107. doi: 10.3389/fnins.2017.00107

Wu, C. T., Weissman, D. H., Roberts, K. C., and Woldorff, M. G. (2007). The neural circuitry underlying the executive control of auditory spatial attention. *Brain Res.* 1134, 187–198. doi: 10.1016/j.brainres.2006.11.088

Wu, M., Li, H., Gao, Y., Lei, M., Teng, X., Wu, X., et al. (2012a). Adding irrelevant information to the content prime reduces the prime-induced unmasking effect on speech recognition. *Hear. Res.* 283, 136–143. doi: 10.1016/j.heares.2011.11.001

Wu, M., Li, H., Hong, Z., Xian, X., Li, J., Wu, X., et al. (2012b). Effects of aging on the ability to benefit from prior knowledge of message content in masked speech recognition. *Speech Commun.* 54, 529–542. doi: 10.1016/j.specom.2011.11.003

Wu, X., Wang, C., Chen, J., Qu, H., and Li, W. (2005). The effect of perceived spatial separation on informational masking of Chinese speech. *Hear. Res.* 199, 1–10. doi: 10.1016/j.heares.2004.03.010

Wu, Z. M., Chen, M. L., Wu, X. H., and Li, L. (2014). Interaction between auditory and motor systems in speech perception. *Neurosci. Bull.* 30, 490–496. doi: 10.1007/s12264-013-1428-6

Xing, Y., Fu, Z., Wu, X., and Chen, J. (2016). "Evaluation of Apple iOS-based automated audiometry," In *Proceedings of the 22nd International Congress on Acoustics*, (Rome: ICA).

Yang, Z., Chen, J., Huang, Q., Wu, X., Wu, Y., Schneider, B. A., et al. (2007). The effect of voice cuing on releasing Chinese speech from informational masking. *Speech Commun.* 49, 892–904. doi: 10.1016/j.specom.2007.05.005

Zheng, Y., Wu, C., Li, J., Wu, H., She, S., Liu, S., et al. (2016). Brain substrates of perceived spatial separation between speech sources under simulated reverberant listening conditions in schizophrenia. *Psychol. Med.* 46, 477–491. doi: 10.1017/S0033291715001828

Zündorf, I. C., Lewald, J., and Karnath, H. O. (2013). Neural correlates of sound localization in complex acoustic environments. *PLoS One* 8:e64259. doi: 10.1371/journal.pone.0064259

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

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 01 July 2021
doi: 10.3389/fnins.2021.628866

Changes of the Brain Causal Connectivity Networks in Patients With Long-Term Bilateral Hearing Loss

Gang Zhang[1†], Long-Chun Xu[2†], Min-Feng Zhang[2†], Yue Zou[1], Le-Min He[3], Yun-Fu Cheng[3], Dong-Sheng Zhang[3], Wen-Bo Zhao[1], Xiao-Yan Wang[3], Peng-Cheng Wang[3] and Guang-Yu Zhang[3*]*

[1] Department of Otorhinolaryngology and Head-Neck Surgery, The Second Affiliated Hospital, Shandong First Medical University, Tai'an, China, [2] Department of Radiology, The Second Affiliated Hospital, Shandong First Medical University, Tai'an, China, [3] Department of Radiology, Shandong First Medical University & Shandong Academy of Medical Sciences, Tai'an, China

OPEN ACCESS

Edited by:
Howard Charles Nusbaum, University of Chicago, United States

Reviewed by:
Christian Füllgrabe, Loughborough University, United Kingdom
Andrej Kral, Hannover Medical School, Germany

***Correspondence:**
Yun-Fu Cheng
yfcheng@tsmc.edu.cn
Guang-Yu Zhang
gyuzhn@163.com

[†] These authors have contributed equally to this work

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

Received: *13 November 2020*
Accepted: *20 April 2021*
Published: *01 July 2021*

Citation:
Zhang G, Xu L-C, Zhang M-F, Zou Y, He L-M, Cheng Y-F, Zhang D-S, Zhao W-B, Wang X-Y, Wang P-C and Zhang G-Y (2021) Changes of the Brain Causal Connectivity Networks in Patients With Long-Term Bilateral Hearing Loss. Front. Neurosci. 15:628866. doi: 10.3389/fnins.2021.628866

It remains poorly understood how brain causal connectivity networks change following hearing loss and their effects on cognition. In the current study, we investigated this issue. Twelve patients with long-term bilateral sensorineural hearing loss [mean age, 55.7 ± 2.0; range, 39–63 years; threshold of hearing level (HL): left ear, 49.0 ± 4.1 dB HL, range, 31.25–76.25 dB HL; right ear, 55.1 ± 7.1 dB HL, range, 35–115 dB HL; the duration of hearing loss, 16.67 ± 4.5, range, 3–55 years] and 12 matched normally hearing controls (mean age, 52.3 ± 1.8; range, 42–63 years; threshold of hearing level: left ear, 17.6 ± 1.3 dB HL, range, 11.25–26.25 dB HL; right ear, 19.7 ± 1.3 dB HL, range, 8.75–26.25 dB HL) participated in this experiment. We constructed and analyzed the causal connectivity networks based on functional magnetic resonance imaging data of these participants. Two-sample *t*-tests revealed significant changes of causal connections and nodal degrees in the right secondary visual cortex, associative visual cortex, right dorsolateral prefrontal cortex, left subgenual cortex, and the left cingulate cortex, as well as the shortest causal connectivity paths from the right secondary visual cortex to Broca's area in hearing loss patients. Neuropsychological tests indicated that hearing loss patients presented significant cognitive decline. Pearson's correlation analysis indicated that changes of nodal degrees and the shortest causal connectivity paths were significantly related with poor cognitive performances. We also found a cross-modal reorganization between associative visual cortex and auditory cortex in patients with hearing loss. Additionally, we noted that visual and auditory signals had different effects on neural activities of Broca's area, respectively. These results suggest that changes in brain causal connectivity network are an important neuroimaging mark of cognitive decline. Our findings provide some implications for rehabilitation of hearing loss patients.

Keywords: auditory-visual reorganization, the shortest causal connectivity path, auditory- visual inhibition, the virtual digital brain, hearing loss, speech processing

INTRODUCTION

Behavioral studies have reported that individuals with early or congenitally hearing loss are highly reliant on their remaining intact sensory modalities to interact with their surrounding environment (Merabet and Pascual-Leone, 2010) and displayed superior visual behavioral performing skills compared with normal-hearing controls, such as enhanced peripheral attention to moving stimuli (Neville and Lawson, 1987; Bavelier et al., 2000; Proksch and Bavelier, 2002; Dye et al., 2009) and increased performance abilities in visual motion detection (Hauthal et al., 2013; Shiell et al., 2014, 2016). Subjects with early hearing loss also exhibited faster and more accurate visual working memory performance compared with normal-hearing controls (Ding et al., 2015). However, some disadvantages of early or congenitally hearing loss on visual functions have also been reported; for example, early auditory deprivation results in visual selection deficits (Smith et al., 1998; Horn et al., 2005; Quittner et al., 2010) but does not contribute to better or worse visual attention. Selected aspects of visual attention of early hearing loss individuals are able to be modified in various ways along the developmental trajectory (Dye and Bavelier, 2010).

An animal study indicated that those improvements in visual performances were caused by cross-modal functional reorganization of auditory cortex in individuals with early or congenitally hearing loss, and several experimental approaches had been used to localize individual visual functions in discrete portions of reorganized auditory cortex (Lomber et al., 2010). Projections from a number of fields in the visual cortex were observed in the dorsal zone of the auditory cortex (Barone et al., 2013) and the posterior auditory field (Butler et al., 2017). Furthermore, the structural connectivity changes between visual and auditory cortices contribute to reduced resting-state alpha band activity in the posterior auditory field (Yusuf et al., 2017) and reduced stimulus-related effective connectivity from higher-order auditory cortical areas involved in the cross-modal reorganization to primary areas (Yusuf et al., 2021) in congenitally hearing loss cats.

Studies of hearing loss in human beings have revealed neural activity changes associated with cortical plasticity. Karns et al. (2012) found that adults with congenitally hearing loss exhibited a greater visual stimulating response in the Heschl's gyrus compared with normal-hearing controls. Finney et al. (2001, 2003) also reported a visual-evoked activation of the auditory cortex in individuals with early hearing loss. Dewey and Hartley (2015) found that participants with profoundly hearing loss presented a larger group response elicited by visual motion within the right temporal lobe compared with normal-hearing controls.

Previous investigations have demonstrated cross-modal recruitment of auditory cortical areas in subjects with profoundly hearing loss viewing sign language (Nishimura et al., 1999; Petitto et al., 2000; MacSweeney et al., 2002). Visual cross-modal reorganization is dependent on hearing loss severity and is also observed in adults with mild-moderate hearing loss (Campbell and Sharma, 2014, 2020; Glick and Sharma, 2017). On one hand, this kind of cortical plasticity contributes to superior visual performance abilities and increased visual-evoked activation of the auditory cortex. On the other hand, investigators (Doucet et al., 2006; Sandmann et al., 2012; Lazard and Giraud, 2017) have also found that auditory–visual reorganization impacts rehabilitation of individuals with hearing loss after applying restorative strategies such as cochlear implantation (CI). In these studies, enhanced visual-evoked activation of the auditory cortex was always associated with poor speech comprehension in individuals with early or congenitally hearing loss following CI. As a result, it has been assumed that visual language is maladaptive for hearing restoration with a CI. However, a recent study has reported that visual stimulus provides adaptive benefits to speech processing following CI (Anderson et al., 2017). These inconsistent study results indicate that it is necessary to further investigate the cross-modal reorganization, especially brain causal connectivity network changes following auditory–visual reorganization and their effects on cognitive and speech processing abilities of hearing loss patients. Here, we study this issue. We hypothesize that cross-modal functional reorganization contributes to changes in the brain causal connectivity network, which may affect cognitive and speech processing abilities of patients with hearing loss. Auditory and visual language signals might have different effects on speech processing.

MATERIALS AND METHODS

Participants

Written informed consent was obtained from all participants according to a protocol approved by the Institutional Ethics Committee of Shandong First Medical University. In the current study, the severity levels of the hearing loss are labeled according to the Global Burden of Disease (GBD) 2013 and World report on hearing (i.e., mild: 20 to <35 dB, moderate: 35–49 dB, moderately severe: 50–64 dB, severe: 65–79 dB, profound: 80–94 dB, and complete $\geq$ 95 dB) (Xu et al., 2019b; World Health Organization, 2021). The less severe hearing loss is used as evaluating criteria in the bilateral. Twelve patients with long-term bilateral mild-to-severe sensorineural hearing loss and 12 age-, sex-, and education-matched normally hearing volunteers participated in this study. Hearing loss participants were asked about their hearing loss, including the etiology of hearing loss, age at onset and duration of hearing loss, duration of tinnitus, and hearing aid experience. Only one participant had a year of hearing aid use, and others had no hearing aid experience. All patients had post-lingual hearing loss, and 11 of them had self-reported tinnitus with a mean duration of 14.4 years $\pm$ 5.9 (standard error), ranging from 10 days to 55 years. Hearing loss was acquired due to infection in two patients and to noise and traumatism in one patient, and hearing loss in others had an unknown cause. All participants were right-handed and had normal or corrected-to-normal vision. None of them had a history of medical and stroke/cerebrovascular ischemia.

A clinical audiologist performed pure tone audiometry (PTA) using an Astera audiometer in the Department of Otorhinolaryngology and Head-neck Surgery of the Second Affiliated Hospital of Shandong First Medical University. Hearing level (HL) was defined as a speech-frequency pure tone

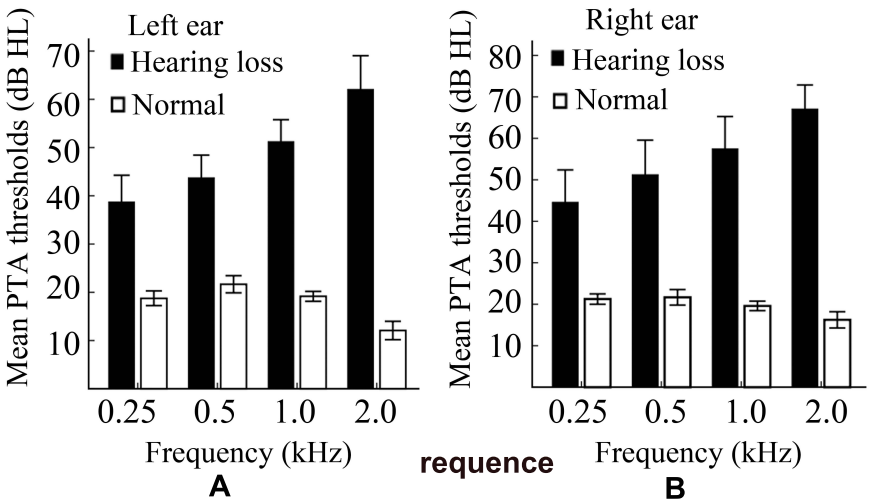

FIGURE 1 | A plot of the non-aided audiograms. **(A)** The non-aided audiogram of the left ear. **(B)** The non-aided audiogram of the right ear. PTA, pure tone audiometry; HL, hearing level.

average of air conduction thresholds at 0.25, 0.5, 1.0, and 2 kHz. All of the normally hearing participants had a HL of less than 27 dB, while all of the hearing loss patients had a HL of more than 30 dB. Non-aided audiograms based on pure-tone audiometry for these participants are shown in **Figure 1**. Details of the demographic characteristics of participants are shown in **Table 1**. Data of 21 subjects had been from a previously published study (Xu et al., 2017).

Participants were excluded if their head movements were greater than 2.5 mm or 2.5° during imaging. No participants were excluded in the present study.

Neuropsychological Tests

Mini-mental state examination (MMSE) is sensitive for measuring cognitive impairment (Tombaugh and Mcintyre, 1992). In this study, MMSE and Wechsler adult intelligence test (Wechsler, 1997) were performed to assess the cognitive abilities of the participants. Full details of all tests had been published in a previous study (Zhang et al., 2015).

It has been demonstrated that hearing impairment negatively impacts cognitive test performance (Füllgrabe, 2020a). To minimize the influence of hearing loss on the assessment of cognitive performances. All neuropsychological tests were administered in a quiet setting with minimal distractions and were provided in Mandarin. An experienced examiner had been accustomed to working with hearing loss patients and verbally gave all instructions in a face-to-face manner. However, as none of these precautions offset a possible bias introduced by comparing normal-hearing and hearing-impaired participants on an auditory cognitive task, the reported cognitive data are possibly biased and their association with changes in brain connectivity needs to be interpreted with caution (Füllgrabe, 2020b). Participants were tested on (a) MMSE, (b) digit symbol substitution test (DSST), (c) forward digit spans (FDS) and backward digit spans (BDS), (d) verbal fluency (VF), (e) trail making test part A (TMTA), (f) trail making test part B (TMTB), (g) Stroop color–word test A (SCWA), (h) Stroop color–word test B (SCWB), and (i) Stroop color–word test C (SCWC).

Magnetic Resonance Imaging (MRI) Data Acquisition

We acquired MRI data of 24 participants using a GE Discovery MR 750 3-T scanner. MRI data of 21 participants had been from a previously published study (Xu et al., 2017). Functional images were collected axially using an echo-planar imaging sequence (echo time: 30 ms; repetition time: 2000 ms; 41 slices, 64 × 64 matrix; voxel size: 3.4375 mm × 3.4375 mm × 3.2 mm). T1-weighted structural images were also acquired with parameters: echo time = 8.156 ms; repetition time = 3.18 ms; 176 slices with 1 mm × 1 mm × 1 mm voxels.

All of the participants underwent a functional magnetic resonance scan during a conscious resting state with their eyes closed. To reduce noise and disturbance, these participants wore earplugs and headphones and were instructed to keep still, as motionless as possible and not to think about anything in particular. Head movement was minimized by placing soft pads at the sides of the head. The light was switched off during the resting-state scanning.

Data Preprocessing

All of the MRI data were firstly preprocessed by using the software package spm8[1]. Slice timing, motion correction, space normalization, and spatial smoothing (8 mm × 8 mm × 8 mm) were first executed. Segmentation for structural images generated white matter, gray matter, and cerebrospinal fluid images. Full details of all preprocessing steps are provided in the previously published study (Xu et al., 2017). Then, we performed the following procedures utilizing the virtual digital brain software

[1]http://www.fil.ion.ucl.ac.uk/spm/software/spm8/

TABLE 1 | Demographic characteristics of participants.

Characteristics	Hearing loss (n = 12)	Normal (n = 12)	Statistic (df)	p-value
Age (yrs)	55.7 ± 2.0	52.3 ± 1.8	T = 1.256 (22)	0.222
Education (yrs)	10.1 ± 0.7	11.7 ± 0.9	T = −1.373 (22)	0.184
Male/Female	9/3	5/7	χ^2 = 2.743 (1)	0.098
Threshold (dB HL)	L: 49.0 ± 4.1	L: 17.6 ± 1.3	T = 7.229 (13.2)	<0.001*
	R: 55.1 ± 7.1	R: 19.7 ± 1.3	T = 4.925 (11.8) **	<0.001*
	L: 49.0 ± 4.1 R: 55.1 ± 7.1	–	T = -0.751 (22)#	0.461
	–	L: 17.6 ± 1.3 R: 19.7 ± 1.3	T = -1.094 (22)#	0.286
Hearing loss duration (yrs)	16.7 ± 4.5	na	–	–

Data are mean ± standard errors. yrs, years; L, left; R, right; df, degrees of freedom; na, not applicable; HL, hearing level.
**$p < 0.05$ is considered to indicate a significant difference.*
***Homogeneity test (Levene's test), $p < 0.05$, revised t-test.*
#Comparisons of hearing levels across both left and right ears (two-sample t-tests). Age of the hearing loss participant ranges from 39 to 63 years; onset of hearing loss ranges from 5 to 58 years of age; the duration of hearing loss ranges from 3 to 55 years; hearing level of left ear ranges from 31.25 to 76.25 dB HL; hearing level of right ear ranges from 35 to 115 dB HL; age of the normally hearing volunteer ranges from 42 to 63 years; hearing level of left ear ranges from 11.25 to 26.25 dB HL; hearing level of right ear ranges from 8.75 to 26.25 dB HL.

package VDB1.6[2] for the preprocessed data: (1) the removal of linear and quadratic trends; (2) regressing out covariates including realignment parameters, mean white matter, and cerebrospinal fluid signals; (3) band-pass temporal filtering (0.01–0.1 Hz); and (4) calculating the interregional causal connections of each participant. Finally, we obtained the nodal degrees and the shortest causal connectivity paths.

The Shortest Causal Connectivity Paths and Transfer Probability

In the current study, causal connectivity was realized by entropy connectivity method, as defined in a previous study (Zhang et al., 2016), and the causal connectivity is related to BOLD (blood oxygen level-dependent) signal amplitude. Suppose there is a strong correlation between amplitude changes of the BOLD signal across two brain regions; moreover, the current change of the BOLD signal amplitude in one brain region presents synchronous or asynchronous coupling with the future change of the other, then there exist a causal connectivity between the two brain regions (i.e., one is response system, and the other is drive system).

[2]https://www.nitrc.org/projects/vdb/

Let, $L_{k\to n}$ denote the length of the shortest causal connectivity path from brain region BA k to n, then $L_{k\to n}$ was defined as:

$$L_{k\to n} = \sum 1/S^4_{i\to j} \quad (1)$$

where, $S_{i\to j}$ denotes the strength of causal connectivity from brain region BA i to j. BA i and BA j are two adjacent brain regions in the shortest causal connectivity path from BA k to n (**Figure 2A**). The shortest causal connectivity path reflects the strongest causal interaction from one brain region to the other. Let $TP_{k\to n}$ denote the transfer probability of the shortest causal connectivity path from brain region BA k to n, then $TP_{k\to n}$ was defined as:

$$TP_{k\to n} = \prod P_{i\to j} \quad (2)$$

where, $P_{i\to j}$ denotes the probability of synchronous or asynchronous causal connectivity from brain region BA i to j (**Figure 2A**). The transfer probability reflects the influence extent of one brain region's neural activity on the other through the shortest causal connectivity path. $TP_{k\to n} > 0$ indicates a positive influence, and $TP_{k\to n} < 0$ indicates a negative influence. In general, $TP_{k\to n} > 0.05$ or $TP_{k\to n} < -0.05$ is regarded as a significant influence.

Positive Reproducible Test (PRT)

In statistical hypothesis testing, a type I error is the incorrect rejection of a true null hypothesis (i.e., a "false positive"), while a type II error is the failure to reject a false null hypothesis (i.e., a "false negative"). A false positive probability is regarded as the exact probability of committing a type I error or the exact level of significance (i.e., p-value). A false-negative probability is

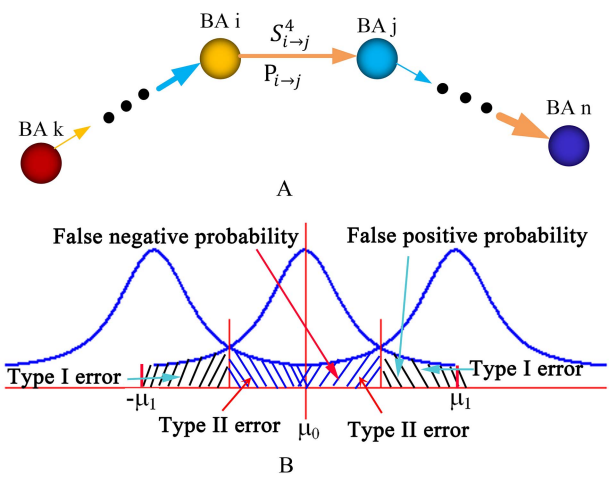

FIGURE 2 | Diagram of the shortest causal connectivity path, transfer probability, and type I and II errors. **(A)** The diagram of the shortest causal connectivity path and transfer probability. **(B)** The diagram of type I error and type II error. The curves in the diagram are statistical distributions. The parameters μ_0 and μ_1 are the means or expectations of the distributions. The value of μ_0 is equal to zero, and the value of μ_1 depends on the actual appilcation. BA, Brodmann's Area.

the exact probability of committing a type II error (**Figure 2B**). Types I and II error probabilities (i.e., false-positive and -negative probabilities) are a zero-sum game for any given sample size. Any method that protects more against one type of error is guaranteed to increase the probability of the other kind of error (Lieberman and Cunningham, 2009). Reduced type I error will lead to increased type II error. Thus, we suppose a positive reproducible test method to obtain a trade-off between the false-positive and -negative probabilities. This method is described as follows: (1) set p-value (i.e., false-positive probability) as a significant level in a statistical hypothesis testing for all samples and obtain n positive effects $A_1, A_2, \cdots, A_n$; (2) randomly select k samples and repeat the statistical hypothesis testing; (3) repeat step (2) M times; (4) if the positive effect A_i occurs m_i times, then the reproducible rate R_i of the positive effect A_i is defined as

$$R_i = \frac{m_i}{M} \tag{3}$$

where, i 1, 2, $\cdots$, n, higher R_i means fewer type I errors. This supposed method can obtain low false-negative probability and few type I errors through selecting high R_i in the statistical hypothesis testing.

Causal Connections and Nodal Degrees

The total number of synchronous or asynchronous input causal connectivity of a brain region is defined as synchronous or asynchronous input nodal degree of the brain region. Similarly, the total number of synchronous or asynchronous output causal connectivity of a brain region is defined as synchronous or asynchronous output nodal degree of the brain region.

Enhanced synchronous input causal connectivity and increased synchronous input nodal degree of a brain region mean that some original functions associated with the brain region will be weakened due to increased resource occupation. Similarly, enhanced asynchronous input causal connectivity and increased asynchronous input nodal degree of a brain region indicate that neural activity of the brain region will be depressed and related functions will be weakened.

Weakened synchronous input causal connectivity and reduced synchronous input nodal degree of a brain region imply that some functions associated with the brain region will be enhanced due to decreased resource occupation. Similarly, weakened asynchronous input causal connectivity and reduced asynchronous input nodal degree of a brain region indicate that the neural activity of this brain region will be enhanced due to reduced neural inhibition.

Enhanced synchronous output causal connectivity and increased synchronous output nodal degree of a brain region imply that the information processing ability of the brain region will be enhanced due to occupying more resources. Similarly, enhanced asynchronous output causal connectivity and increased asynchronous output nodal degree of a brain region indicate that the brain region enhances its information processing abilities through depressing neural activities of other brain regions.

Weakened synchronous output causal connectivity and reduced synchronous output nodal degree of a brain region imply that the information processing ability associated with the brain region will decline due to interrupted causal connectivity. Similarly, weakened asynchronous output causal connectivity and reduced asynchronous output nodal degree of a brain region imply the decline of some functions associated with the brain region.

All performances were executed using the VDB1.6 software package.

Statistical Analyses

We performed two-sample t-tests to examine group differences in age, educational levels, hearing levels, and neuropsychological test scores. A Chi-square test was conducted to examine group differences in sex. Statistical analyses were conducted using SPSS software (version 19.0), and a $p < 0.05$ was considered to indicate a significant difference in any single analysis. A $p < 0.05$ (FDR corrected) is considered as a significant difference across the hearing loss and normally hearing control group for the two-sample t-tests of neuropsychological performance scores.

RESULTS

Neuropsychological Test Results

The results of neuropsychological tests are summarized in **Table 2**. Compared with normally hearing controls, patients with hearing loss presented significantly decreased test scores ($p < 0.05$, FDR corrected) in MMSE, VF, and BDS, but significantly increased test scores ($p < 0.05$, FDR corrected) in TMTA, SCWTA, SCWTB, and SCWTC. A lower score means a worse performance for MMSE, VF, and BDS. A higher score means a worse performance for TMTA, SCWTA, SCWTB, and SCWTC. No significant difference was observed in other tests across the two groups.

TABLE 2 | Summary of neuropsychological tests.

Tests	Hearing loss (n = 12)	Normal (n = 12)	Statistic (df)	p-value
MMSE	27.3 ± 0.54	29 ± 0.30	T = −2.776 (17.3)**	0.013*
DSST	35.4 ± 4.1	45.1 ± 3.8	T = −1.735 (22)	0.370
VF	25.7 ± 1.8	32.8 ± 1.8	T = −2.807 (22)	0.01*
FDS	6.9 ± 0.29	7.6 ± 0.26	T = −1.72 (22)	0.099
BDS	4.2 ± 0.17	5.2 ± 0.34	T = −2.613 (22)	0.016*
TMTA	58.6 ± 5.62	42.7 ± 2.47	T = 2.589 (15.1)**	0.020*
TMTB	160.6 ± 16.5	123.0 ± 11.3	T = 1.876 (22)	0.074
SCWA	30.5 ± 2.1	24.3 ± 1.3	T = 2.519 (18.4)**	0.021*
SCWB	47.7 ± 3.8	36.3 ± 1.9	T = 2.677 (22)	0.014*
SCWC	86.3 ± 6.14	65.5 ± 4.7	T = 2.683 (22)	0.014*

Data are mean ± standard errors. df, degrees of freedom.
**$p < 0.05$ (FDR corrected) is considered as a significant difference.*
***Homogeneity test (Levene's test), $p < 0.05$, revised t-test. MMSE, mini-mental state examination; DSST, digit symbol substitution test; VF, verbal fluency; FDS, forward digit spans; BDS, backward digit spans; TMTA, trail making test part A; TMTB, trail making test part B; SCWTA, Stroop color–word test A; SCWTB, Stroop color–word test B; SCWTC, Stroop color–word test C.*

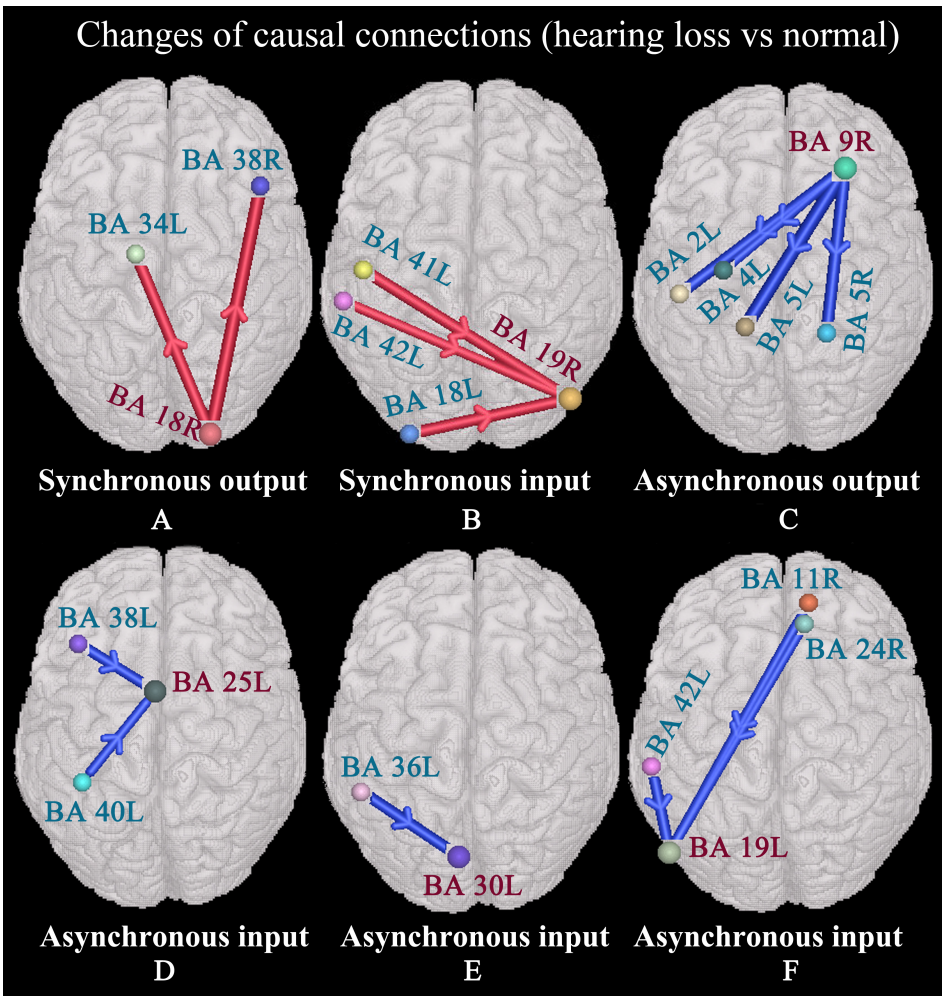

FIGURE 3 | Changes of causal connections in patients with hearing loss ($p < 0.05$, two-sided, PRT corrected; reproducible rate: 0.5; number of subjects: 10; PRT threshold: 4000). The red bar denotes enhanced causal connectivity, and the blue bar denotes weakened causal connectivity. The width of the bar indicates the strength of causal connectivity. The direction of the arrow indicates the direction of causal connectivity. See **Table 3** for the BA index. **(A)** Enhanced synchronous output causal connections for BA 18R. **(B)** Enhanced synchronous input causal connections for BA 19R. **(C)** Weakened asynchronous output causal connections for BA 9R. **(D)** Weakened asynchronous input causal connections for BA 25L. **(E)** Weakened asynchronous input causal connections for BA 30L. **(F)** Weakened asynchronous input causal connections for BA 19L.

Changes of Causal Connections

We studied changes of causal connections and nodal degrees in patients with hearing loss, as well as the relationships between nodal degrees and cognitive performance scores. The strength of interregional causal connectivity was corrected using the positive reproducible test (PRT) method. Only those significant changes and correlations are described as follows.

We found that the right secondary visual cortex (BA 18R) presented enhanced synchronous output causal connections with the left anterior entorhinal cortex (BA 34L) and right temporopolar area (BA 38R) (**Figure 3A**); the right associative visual cortex (BA 19R) exhibited enhanced synchronous input causal connectivity with the left primary auditory cortex (BAs 41L and 42L) and secondary visual cortex (BA 18L) (**Figure 3B**); the right dorsolateral prefrontal cortex (BA 9R) exhibited weakened asynchronous output causal connections with the left primary somatosensory cortex (BA 2L), left primary motor cortex (BA 4L), and the somatosensory association cortex (BAs 5L and 5R) (**Figure 3C**); the left subgenual cortex (BA 25L) presented weakened asynchronous input causal connectivity with the left temporopolar area (BA 38L) and supramarginal gyrus (BA 40L) (**Figure 3D**); the left cingulate cortex (BA 30L) presented weakened asynchronous input causal connections with the left parahippocampal cortex (BA 36L) (**Figure 3E**); the left associative visual cortex (BA 19L) presented weakened asynchronous input causal connectivity with the left primary auditory cortex (BA 42L), right orbitofrontal cortex (BA 11R), and the ventral anterior cingulate cortex (BA 24R) (**Figure 3F**).

TABLE 3 | Indexes and corresponding brain regions.

Indexes	Brain regions	Indexes	Brain regions
BA 2L	Left primary somatosensory cortex	BA 11R	Right orbitofrontal cortex
BA 4L	Left primary motor cortex	BA 13L	Left insular cortex
BA 5L	Left somatosensory association cortex	BA 13R	Right insular cortex
BA 5R	Right somatosensory association cortex	BA 17R	Right primary visual cortex
BA 8R	Right dorsal frontal cortex	BA 18R	Right secondary visual cortex
BA 9R	Right dorsolateral prefrontal cortex	BA 19R	Right associative visual cortex
BA 9L	Left dorsolateral prefrontal cortex	BA 19L	Left associative visual cortex
BA 24R	Right ventral anterior cingulate cortex	BA 25L	Left subgenual cortex
BA 32R	Right dorsal anterior cingulate cortex	BA 30L	Left cingulate cortex
BA 34L	Left anterior entorhinal cortex	BA 36L	Left parahippocampal cortex
BA 36R	Right parahippocampal cortex	BA 38R	Right temporopolar area
BA 38L	Left temporopolar area	BA 40L	Left supramarginal gyrus
BA 41L	left primary auditory cortex	BA 41R	Right primary auditory cortex
BA 42L	left primary auditory cortex	BA 43L	Left subcentral area
BA 46R	Right dorsolateral prefrontal cortex	BA 44R	Right IFC pars opercularis
BA 46L	Left dorsolateral prefrontal cortex	BA 45L	Left IFC pars triangularis

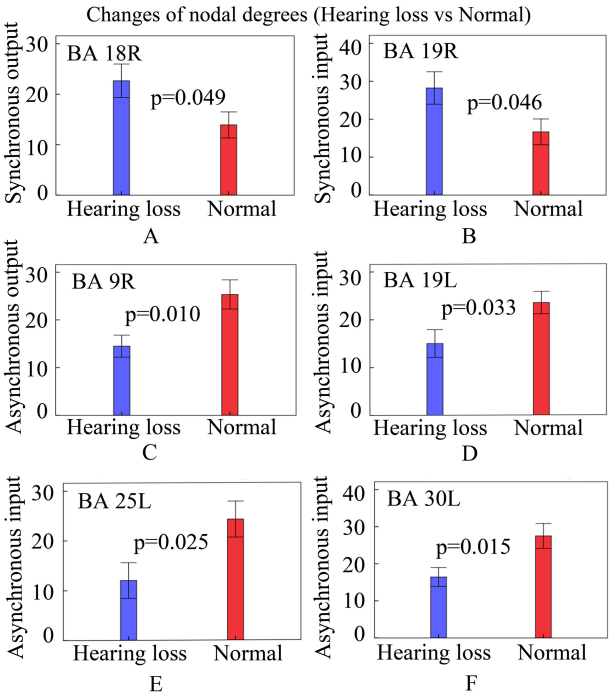

FIGURE 4 | Changes of nodal degrees in patients with hearing loss. Bars: mean ± standard errors. See **Table 3** for the BA index. **(A)** The bar graph of synchronous output nodal degrees for BA 18R. **(B)** The bar graph of synchronous input nodal degrees for BA 19R. **(C)** The bar graph of asynchronous output nodal degrees for BA 9R. **(D)** The bar graph of asynchronous input nodal degrees for BA 19L. **(E)** The bar graph of asynchronous input nodal degrees for BA 25L. **(F)** The bar graph of asynchronous input nodal degrees for BA 30L.

Changes of Nodal Degrees

In order to observe whether causal connectivity changes had an impact on nodal degrees, we studied the nodal degree utilizing two-sample *t*-tests and found significant changes ($p < 0.05$, two-sided) in hearing loss patients. The results are shown in **Figure 4**.

Compared with normally hearing subjects, patients with hearing loss presented increased synchronous output nodal degrees in BA 18R (**Figure 4A**); increased synchronous input nodal degrees in BA 19R (**Figure 4B**); reduced asynchronous output nodal degrees in BA 9R (**Figure 4C**); and reduced asynchronous input nodal degrees in BAs 19L, 25L, and 30L (**Figures 4D–F**).

Changes of the Shortest Causal Connectivity Paths

The shortest causal connectivity path from one brain region to the other reflects the strongest causal interaction between these two brain regions. It is unclear whether the change of nodal degree also leads to the change of the shortest causal connectivity path; we studied this issue. Our study focused on the shortest causal connectivity paths between the visual cortex, auditory cortex, and those brain areas associated with speech processing. The results are shown in **Figure 5**.

Compared with normal-hearing subjects, patients with hearing loss presented a shorter causal connectivity path ($p < 0.0001$, two-sided, two-sample *t*-test, uncorrected) and a bigger transfer probability (hearing loss group = −0.41, normal-hearing control = −0.24) from BA 18R to Broca's area (BA 45L) (**Figures 5A,B**). In addition, we also studied the shortest causal connectivity paths and transfer probabilities (hearing loss group = −0.25, normal-hearing control = −0.61) from BA 18R to 41L (**Figures 5C,D**); BA 19L to 45L with transfer probabilities (hearing loss group = −0.41, normal-hearing control = −0.37) (**Figures 5E,F**); BA 19L to 41L with transfer probabilities (hearing loss group = +0.25, normal-hearing control = −0.10) (**Figures 5G,H**); and BA 41R to 45L with transfer probabilities (hearing loss group = +0.16, normal-hearing control = +0.24) (**Figures 4I,J**). The transfer probability > 0.05 or < −0.05 indicates that the change of neural signal in one brain region has significant effects on the other through the shortest causal connectivity path from one brain region to the other.

Correlation Analysis

In order to investigate whether nodal degree changes can contribute to cognitive decline, we studied the relationship of cognitive performance scores with nodal degrees utilizing Pearson's correlation analysis method. The results are described as follows. VF test scores exhibited a significant negative

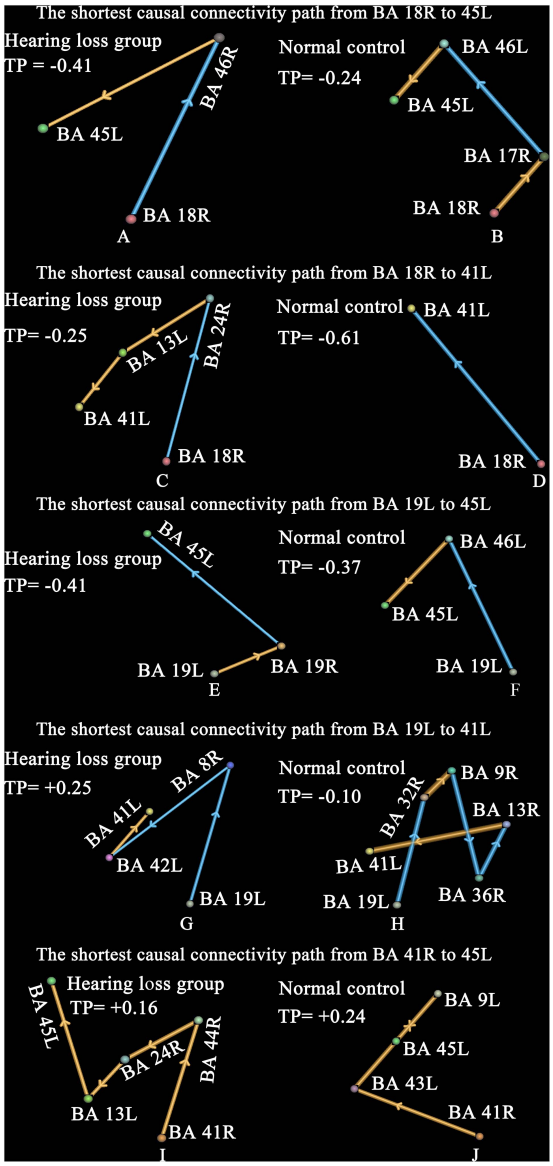

FIGURE 5 | The shortest causal connectivity paths and transfer probabilities. Strengths of interregional causal connections were corrected by using positive reproducible test (PRT) method: $p < 0.05$, PRT corrected; reproducible rate 0.5; number of subjects 10; PRT threshold 1000. The golden bar denotes synchronous causal connectivity, and the light blue bar denotes asynchronous causal connectivity. The width of the bar indicates the strength of causal connectivity. The direction of the bar indicates the direction of causal connectivity. The transfer probability indicates the influence of one brain region on the other through the shortest causal connectivity path. Transfer probability > 0.05 denotes a significant positive effect, and transfer probability < −0.05 denotes a significant negative effect. See **Table 3** for the BA index. **(A,B)** Denote the shortest causal connectivity path from BA 18R to 45L in the hearing loss and the normal control,respectively. **(C,D)** Denote the shortest causal connectivity path from BA 18R to 41L in the hearing loss and the normal control,respectively. **(E,F)** Denote the shortest causal connectivity path from BA 19L to 45L in the hearing loss and the normal control, respectively. **(G,H)** Denote the shortest causal connectivity path from BA 19L to 41L in the hearing loss and the normal control, respectively. **(I,J)** Denote the shortest causal connectivity path from BA 41R to 45L in the hearing loss and the normal control, respectively.

correlation with synchronous output nodal degrees of BA 18R (**Figure 6A**); TMTA test scores exhibited a significant positive correlation with synchronous input nodal degrees of BA 19R (**Figure 6B**); SCWTA test scores presented a significant negative correlation with asynchronous input nodal degrees of BA 19L (**Figure 6C**); SCWTC test scores presented a significant negative correlation with asynchronous output nodal degrees of BA 9R (**Figure 6D**); MMSE test scores presented significant positive correlations with asynchronous input nodal degrees of both BA 25L and BA 30L (**Figures 6E,F**).

In addition, we also analyzed the correlations of cognitive performance scores with lengths of the shortest causal connectivity paths from BA 18R to Broca's area (BA 45L). We found that both MMSE and VF test scores were positively correlated with lengths of the shortest causal connectivity paths from BA 18R to BA 45L (**Figures 7A,B**). On the contrary, SCWTB and SCWTC test scores presented negative correlations with the path lengths (**Figures 7C,D**).

In this study, we have mixed controls and hearing loss subjects into one set and computed correlations (**Figures 6**, **7**). This may lead to a curious correlation given the fact that the hearing loss subjects may simply show a different outcome than controls. To verify this issue, we recomputed those correlations by separating controls and hearing loss subjects and found that only several correlations are significant: synchronous output nodal degrees of BA 18R with VF test scores in the control group (**Figure 8A**); MMSE test scores with asynchronous input nodal degrees of BA 30L in the hearing loss group (**Figure 8B**); SCWTC test scores with asynchronous output nodal degrees of BA 9R in the hearing loss group (**Figure 8C**); and lengths of the shortest causal connectivity paths from BA 18R to BA 45L with MMSE test scores in the hearing loss group (**Figure 8D**).

To further explore the influence of hearing loss on causal connectivity network, we also studied the relationships between nodal degrees and durations of hearing loss. Pearson's correlation analysis revealed that the asynchronous output nodal degree of BA 23L strongly negatively related with the duration of hearing loss (**Figure 9A**); similarly, the synchronous output nodal degree of BA 23R strongly negatively related with the duration of hearing loss as well (**Figure 9B**).

DISCUSSION

The present study investigated adult-onset hearing loss and its effects on the connectivity in the brain using resting-state functional magnetic resonance imaging (fMRI). The main findings of the study are described as follows. Patients with hearing loss presented increased synchronous input causal connectivity for the right associative visual cortex with the auditory cortex and exhibited a shorter causal connectivity path from the right secondary visual cortex to Broca's area. These findings were consistent with cross-modal reorganization. Moreover, we also observed nodal degree changes in the right secondary visual cortex, associative visual cortex, right dorsolateral prefrontal cortex, left subgenual cortex, and the left cingulate cortex. These changes were significantly

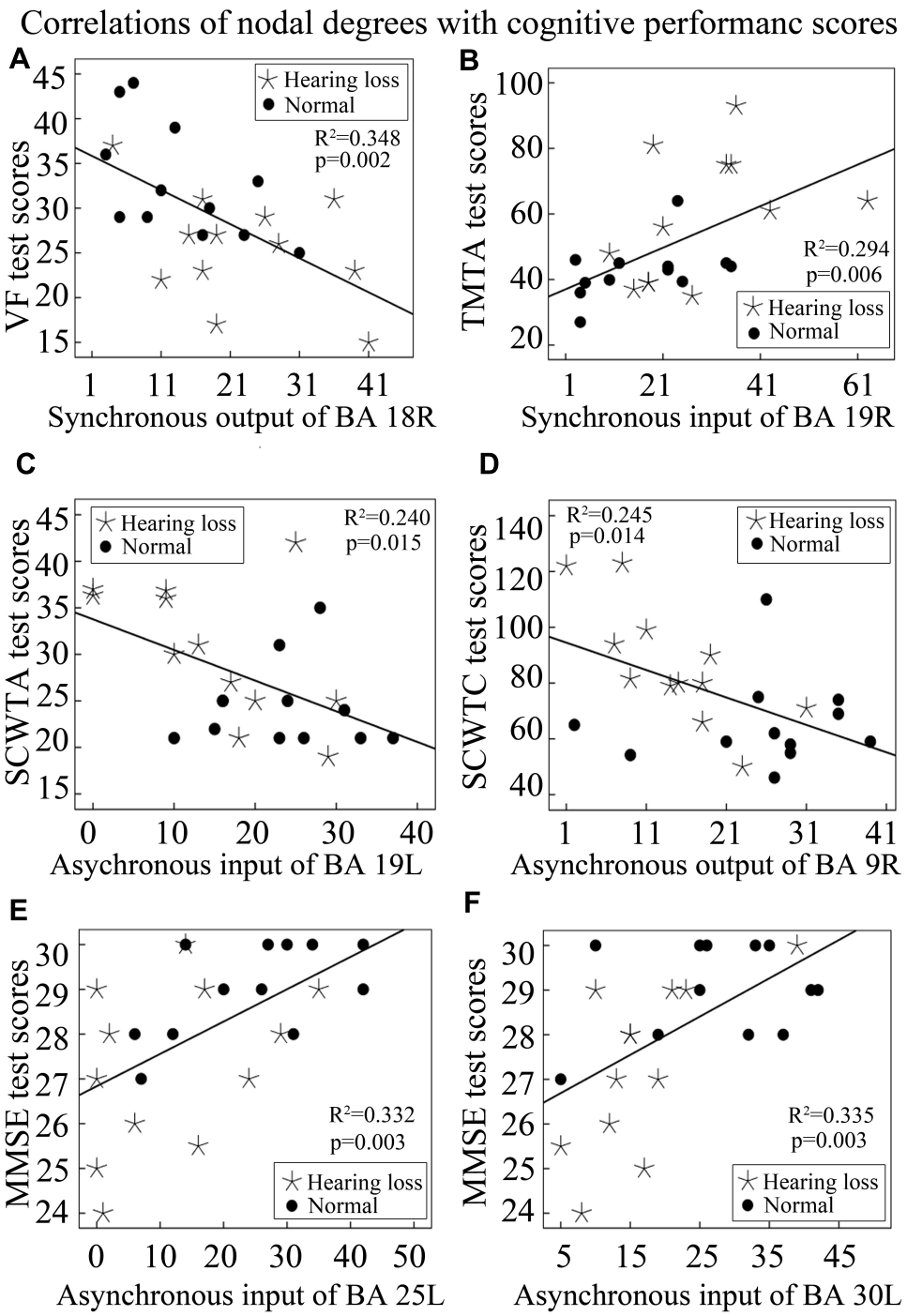

FIGURE 6 | Correlations between nodal degrees and cognitive performance scores. See **Table 3** for the BA index. **(A)** The correlation of synchronous output nodal degrees for BA 18R with VF test scores. **(B)** The correlation of synchronous input nodal degrees for BA 19R with TMTA test scores. **(C)** The correlation of asynchronous input nodal degrees for BA 19L with SCWTA test scores. **(D)** The correlation of asynchronous output nodal degrees for BA 9R with SCWTC test scores. **(E)** The correlation of asynchronous input nodal degrees for BA 25L with MMSE test scores. **(F)** The correlation of asynchronous input nodal degrees for BA 30L with MMSE test scores.

related with poor cognitive performances. In addition, we also noted that changes of neural signal in the visual cortex had negative effects on neural activity of Broca's area but, in the auditory cortex, had positive effects on neural activity of Broca's area through the shortest causal connectivity paths.

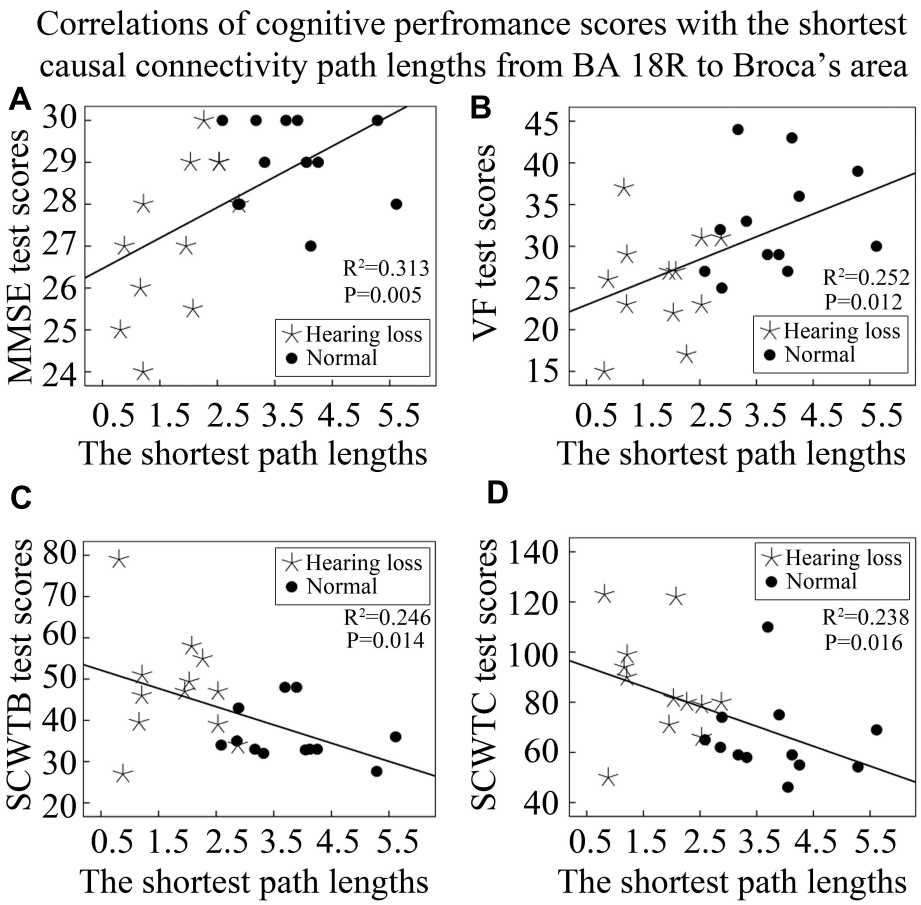

FIGURE 7 | Correlations between cognitive performance scores and lengths of the shortest causal connectivity paths from BA 18R to Broca's area (BA 45L). See **Table 3** for the BA index. **(A)** The correlation of MMSE test scores with the shortest path lengths from BA 18R to Broca's area. **(B)** The correlation of VF test scores with the shortest path lengths from BA 18R to Broca's area. **(C)** The correlation of SCWTB test scores with the shortest path lengths from BA 18R to Broca's area. **(D)** The correlation of SCWTC test scores with the shortest path lengths from BA 18R to Broca's area.

Auditory–Visual Inhibition and Cross-Modal Reorganization

Molloy et al. (2015) found that a visual search task of high perceptual load led to reduced auditory evoked activity in auditory cortical areas compared with that of low perceptual load. Moreover, the reduction in neural responses of the auditory cortex was associated with reduced awareness of the sound. Study on animals found that activation of the auditory cortex by a noise burst suppressed neural activities of the visual cortex *via* cortico-cortical inhibitory circuits (Iurilli et al., 2012). These findings support a neural account of shared audiovisual resources, which lead to auditory–visual inhibition. Our results are consistent with previous studies. Furthermore, this current study also found that patients with hearing loss presented reduced auditory–visual inhibition through the shortest causal connectivity path from the right secondary visual cortex to the left primary auditory cortex (BA 41L) and weakened asynchronous input causal connectivity from the left primary auditory cortex (BA 42L) to the left associative visual cortex. Reduced auditory–visual inhibition might contribute to the auditory cortex recruited by visual modality. In addition, hearing loss patients also exhibited enhanced synchronous input causal connectivity from the left primary auditory cortex to the right associative visual cortex, as well as enhanced causal connectivity through a shortest causal connectivity path from the left associative visual cortex to the left primary auditory cortex. These results suggest that there exists auditory–visual reorganization in patients with long-term bilateral mild-to-severe hearing loss.

Influence of Visual and Auditory Signals on Neural Activities of Broca's Area

Previous investigations found that enhanced visual-evoked activation in auditory brain regions of individuals with early or profoundly hearing loss was associated with poor speech comprehension (Doucet et al., 2006; Sandmann et al., 2012; Lazard and Giraud, 2017), and greater activation of auditory cortical areas in adult CI users during lip-reading predicted poorer speech understanding abilities (Strelnikov et al., 2013). Activation of auditory cortex by visual language might limit its capacity for auditory signal processing (Schormans et al., 2017).

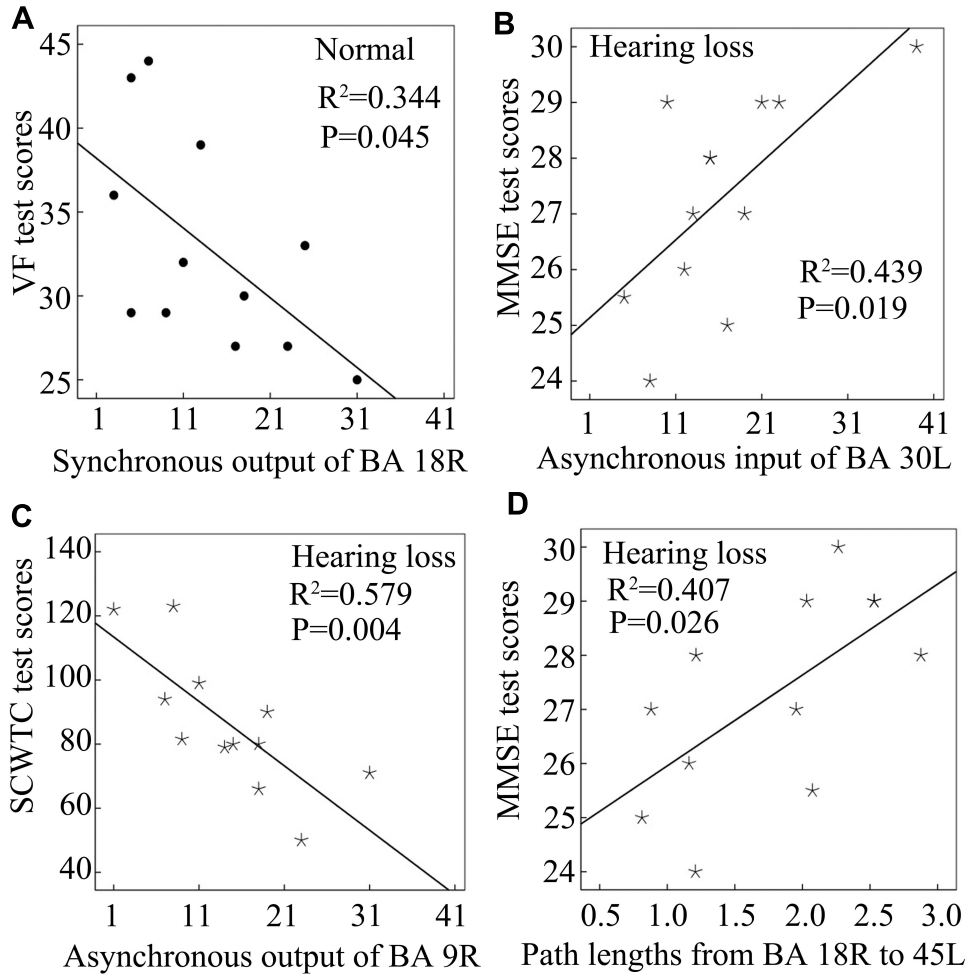

FIGURE 8 | Correlations of the shortest causal connectivity path lengths and nodal degrees with cognitive performance scores in the control or the hearing loss group. See **Table 3** for the BA index. **(A)** The correlation of synchronous output nodal degrees for BA 18R with VF test scores in the normal control group. **(B)** The correlation of asynchronous input nodal degrees for BA 30L with MMSE test scores in the hearing loss group. **(C)** The correlation of asynchronous output nodal degrees for BA 9R with SCWTC test scores in the hearing loss group. **(D)** The correlation of the shortest path lengths from BA 18R to 45L with MMSE test scores in the hearing loss group.

Therefore, auditory–visual reorganization is maladaptive for hearing restoration with a CI (Sandmann et al., 2012). It is well known that Broca's area is associated with speech processing (Grewe et al., 2005; DeWitt and Rauschecker, 2012). This current study found that the signal from the visual cortex had a negative effect on neural activity of Broca's area through the shortest causal interactive path with big transfer probability. In particular, the path from the visual cortex to Broca's area is shorter in patients with hearing loss than in normally hearing controls; one possible reason is that hearing loss leads to auditory–visual reorganization, which contributes to larger synchronous output nodal degree of the visual cortex in hearing loss patients. As a result, more brain regions are linked with the visual cortex and lead to the shorter causal connectivity path, and these changes contribute to stronger interregional causal interaction and inhibition of visual signal for speech processing. Furthermore, shorter causal interactive paths from the visual cortex to Broca's area and larger synchronous output nodal degrees of the visual cortex are associated with poorer speech processing performances manifested as lower VF test scores and higher SCWTB and SCWTC test scores in hearing loss patients. Measures of VF are mainly used to investigate language deficits (Tombaugh et al., 1999). Low VF test scores suggest poor speech processing. SCWTB and SCWTC performances are widely used to measure the ability of inhibiting cognitive interference (Scarpina and Tagini, 2017). High SCWTB and SCWTC test scores reflect a strong inhibition of the visual signal for speech processing. In addition, BA 18R in patients

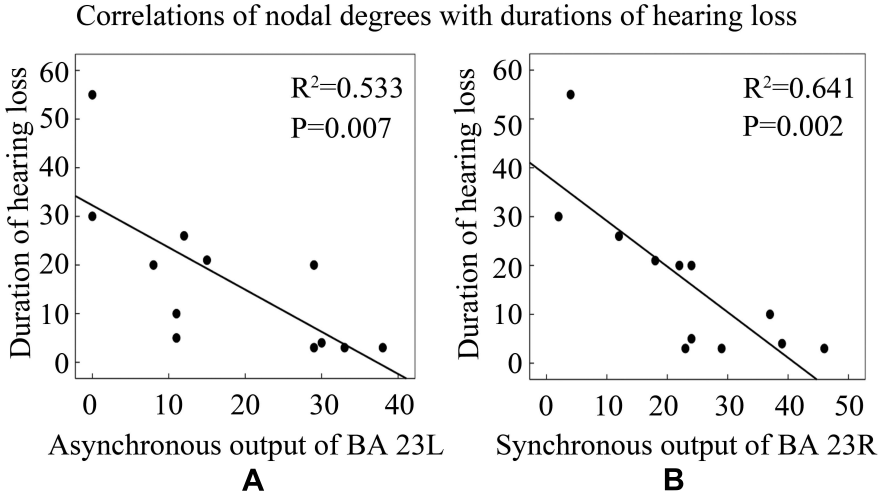

FIGURE 9 | Correlations between nodal degrees and durations of hearing loss. See **Table 3** for the BA index. **(A)** The correlation of asynchronous output nodal degrees for BA 23L with durations of hearing loss (years). **(B)** The correlation of synchronous output nodal degrees for BA 23R with durations of hearing loss (years).

with hearing loss presented enhanced synchronous output casual connectivity with the left anterior entorhinal cortex (BA 34L) and right temporopolar area (BA 38R). BA 34L is associated with cognitive processing (Kovari et al., 2003), and BA 38R is involved in language processing and visual object naming (Poch et al., 2016). These results imply that visual language may be maladaptive for speech processing.

This current study found that the neural signal from the right primary auditory cortex (BA 41R) enhanced neural activity of Broca's area through the shortest causal connectivity path with big transfer probability. Auditory language is adaptive for speech processing; thus current findings support the opinion that intensive and rigorous oral language training can improve language performance of patients with hearing loss following restorative strategies (Kos et al., 2009).

Cognitive Decline Associated With Cerebral Cortex Plasticity

Behavior studies found that hearing loss was associated with incident all-cause dementia (Lin et al., 2011b) and poor cognitive performances such as low scores on DDST and MMSE tests (Lin, 2011; Xu et al., 2017), as well as memory and executive function tests (Lin et al., 2011a). A lot of adults over the age of 60 have both hearing loss and cognitive decline simultaneously (Lin et al., 2011c, 2013). Accumulating evidence from the clinical investigations indicates that there exists a link between hearing loss and cognitive decline (Lindenberger and Baltes, 1994; Lin et al., 2013; Heywood et al., 2017; Huh, 2018; Jayakody et al., 2018; Loughrey et al., 2018). Our current results are consistent with previous investigations. Compared with normally hearing subjects, patients with long-term bilateral hearing loss exhibited significantly poor cognitive performances in MMSE, VF, BDS, TMTA, SCWTA, SCWTB, and SCWTC tests.

Hearing loss contributes to changes in the brain network associated with cognitive processing. Resting-state fMRI studies reported increased functional connectivity and abnormal neural activities in the default mode network (DMN) of individuals with hearing loss compared with normally hearing subjects (Husain et al., 2014; Xu et al., 2017). In addition, previous studies also reported decreased resting-state functional connectivity from the insula to the DMN (Xu et al., 2019a,d), as well as reduced thalamic connectivity with the DMN (Xu et al., 2019c). These investigations indicate that hearing loss leads to neural activity changes in the DMN, which is responsible for cognitive processing (Wheeler et al., 2006; Leech et al., 2011). However, it is not entirely clear whether changes in brain networks contribute to cognitive decline in hearing loss patients. This current study found that the right dorsolateral prefrontal cortex (BA 9R), left subgenual cortex (BA 25L), left cingulate cortex (BA 30L), and the associative visual cortex presented significantly weakened asynchronous causal connections with those brain regions associated with cognitive, language, auditory, and sensor motor processing. Moreover, BAs 9R, 25L, 30L, and 19L presented significantly reduced nodal degrees, which were associated with poor cognitive performances. Previous studies indicate that BA 9R plays a crucial role in visual processing, working memory, and the decisional processes (Manoach et al., 1999; Pierrot-Deseilligny et al., 2003); BA 25L is involved in cognitive behavior activity associated with emotional process (Drevets et al., 1997); BA 30L is associated with implementing retrieval strategies for episodic memory (Nestor et al., 2003); and the associative visual cortex is associated with visual processing, associative memory formation, and cognitive function (McKee et al., 2006; Gazzaley et al., 2007; Rosen et al., 2018). These results suggest that changes in the causal connectivity network might be an important reason that contributes to cognitive decline in hearing loss patients.

In a word, our findings suggest that changes in brain causal connectivity networks are an important mark of cognitive decline in patients with hearing loss. Auditory and visual language

signals might have different effects on speech processing. Our research provides some implications for rehabilitation of patients with hearing loss.

Limitations

Our study has a lower limit of sample size; therefore, significant effects observed may also be the consequence of small power. Hearing loss during development and hearing loss in adulthood may have different effects. The aim of the present study was to investigate adult-onset long-term bilateral mild-to-severe sensorineural hearing loss. However, the use of auditorily presented cognitive tests might have biased the results. In addition, we only investigated resting-state causal connectivity, and these results might be different with stimulus-related connectivity (Okano et al., 2020; Yusuf et al., 2021).

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Institutional Ethics Committee of Shandong First Medical University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

G-YZ and L-CX designed the general approach and wrote the manuscript. W-BZ, GZ, and YZ collected the clinical data. M-FZ and L-CX collected the MRI data. G-YZ, Y-FC, D-SZ, and L-MH processed the MRI data. P-CW and X-YW advised on the experimental design. GZ, P-CW, and X-YW revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the Shandong Provincial Natural Science Foundation, China (grant number: ZR2018MH033, ZR2015HL095, and ZR2014HM072); the High-level project nurturing program of Taishan Medical University (2018GCC13); the Shandong Province Higher Educational Science and Technology Program (No. J17KA262); and the Academic promotion program of Shandong First Medical University (2019LJ004).

REFERENCES

Anderson, C. A. I, Wiggins, M., Kitterick, P. T., and Hartley, D. E. H. (2017). Adaptive benefit of cross-modal plasticity following cochlear implantation in deaf adults. *Proc. Natl. Acad. Sci. U.S.A.* 114, 10256–10261. doi: 10.1073/pnas.1704785114

Barone, P., Lacassagne, L., and Kral, A. (2013). Reorganization of the connectivity of cortical field DZ in congenitally deaf cat. *PLoS One* 8:e60093. doi: 10.1371/journal.pone.0060093

Bavelier, D., Tomann, A., Hutton, C., Mitchell, T., Corina, D., Liu, G., et al. (2000). Visual attention to the periphery is enhanced in congenitally deaf individuals. *J. Neurosci.* 20:RC93.

Butler, B. E., Chabot, N., Kral, A., and Lomber, S. G. (2017). Origins of thalamic and cortical projections to the posterior auditory field in congenitally deaf cats. *Hear. Res.* 343, 118–127. doi: 10.1016/j.heares.2016.06.003

Campbell, J., and Sharma, A. (2014). Cross-modal re-organization in adults with early stage hearing loss. *PLoS One* 9:e90594. doi: 10.1371/journal.pone.0090594

Campbell, J., and Sharma, A. (2020). Frontal cortical modulation of temporal visual cross-modal re-organization in adults with hearing loss. *Brain Sci.* 10, 1–16.

Dewey, R. S., and Hartley, D. E. (2015). Cortical cross-modal plasticity following deafness measured using functional near-infrared spectroscopy. *Hear. Res.* 325, 55–63. doi: 10.1016/j.heares.2015.03.007

DeWitt, I., and Rauschecker, J. P. (2012). Phoneme and word recognition in the auditory ventral stream. *Proc. Natl. Acad. Sci. U.S.A.* 109, E505–E514.

Ding, H., Qin, W., Liang, M., Ming, D., Wan, B., Li, Q., et al. (2015). Cross-modal activation of auditory regions during visuo-spatial working memory in early deafness. *Brain* 138, 2750–2765. doi: 10.1093/brain/awv165

Doucet, M. E., Bergeron, F., Lassonde, M., Ferron, P., and Lepore, F. (2006). Cross-modal reorganization and speech perception in cochlear implant users. *Brain* 129, 3376–3383. doi: 10.1093/brain/awl264

Drevets, W. C., Price, J. L., Simpson, J. R. Jr., Todd, R. D., Reich, T., Vannier, M., et al. (1997). Subgenual prefrontal cortex abnormalities in mood disorders. *Nature* 386, 824–827. doi: 10.1038/386824a0

Dye, M. W., and Bavelier, D. (2010). Attentional enhancements and deficits in deaf populations: an integrative review. *Restor. Neurol. Neurosci.* 28, 181–192. doi: 10.3233/rnn-2010-0501

Dye, M. W., Hauser, P. C., and Bavelier, D. (2009). Is visual selective attention in deaf individuals enhanced or deficient? The case of the useful field of view. *PLoS One* 4:e5640. doi: 10.1371/journal.pone.0005640

Finney, E. M., Clementz, B. A., Hickok, G., and Dobkins, K. R. (2003). Visual stimuli activate auditory cortex in deaf subjects: evidence from MEG. *Neuroreport* 14, 1425–1427. doi: 10.1097/00001756-200308060-00004

Finney, E. M., Fine, I., and Dobkins, K. R. (2001). Visual stimuli activate auditory cortex in the deaf. *Nat. Neurosci.* 4, 1171–1173. doi: 10.1038/nn763

Füllgrabe, C. (2020a). On the possible overestimation of cognitive decline: the impact of age-related hearing loss on cognitive-test performance. *Front. Neurosci.* 14:454. doi: 10.3389/fnins.2020.00454

Füllgrabe, C. (2020b). When hearing loss masquerades as cognitive decline. *J. Neurol. Neurosurg. Psychiatry* 91:1248. doi: 10.1136/jnnp-2020-324707

Gazzaley, A., Rissman, J., Cooney, J., Rutman, A., Seibert, T., Clapp, W., et al. (2007). Functional interactions between prefrontal and visual association cortex contribute to top-down modulation of visual processing. *Cereb. Cortex* 17(Suppl. 1), i125–i135.

Glick, H., and Sharma, A. (2017). Cross-modal plasticity in developmental and age-related hearing loss: clinical implications. *Hear. Res.* 343, 191–201. doi: 10.1016/j.heares.2016.08.012

Grewe, T., Bornkessel, I., Zysset, S., Wiese, R., von Cramon, D. Y., and Schlesewsky, M. (2005). The emergence of the unmarked: a new perspective on the language-specific function of Broca's area. *Hum. Brain Mapp.* 26, 178–190. doi: 10.1002/hbm.20154

Hauthal, N., Sandmann, P., Debener, S., and Thorne, J. D. (2013). Visual movement perception in deaf and hearing individuals. *Adv. Cogn. Psychol.* 9, 53–61. doi: 10.5709/acp-0131-z

Heywood, R., Gao, Q., Nyunt, M. S. Z., Feng, L., Chong, M. S., Lim, W. S., et al. (2017). Hearing loss and risk of mild cognitive impairment and dementia:

findings from the singapore longitudinal ageing study. *Dement. Geriatr. Cogn. Disord.* 43, 259–268. doi: 10.1159/000464281

Horn, D. L., Davis, R. A. O., Pisoni, D. B., and Miyamoto, R. T. (2005). Development of visual attention skills in prelingually deaf children who use cochlear implants. *Ear Hear.* 26, 389–408. doi: 10.1097/00003446-200508000-00003

Huh, M. (2018). The relationships between cognitive function and hearing loss among the elderly. *J. Phys. Ther. Sci.* 30, 174–176. doi: 10.1589/jpts.30.174

Husain, F. T., Carpenter-Thompson, J. R., and Schmidt, S. A. (2014). The effect of mild-to-moderate hearing loss on auditory and emotion processing networks. *Front. Syst. Neurosci.* 8:10. doi: 10.3389/fnsys.2014.00010

Iurilli, G., Ghezzi, D., Olcese, U., Lassi, G., Nazzaro, C., Tonini, R., et al. (2012). Sound-driven synaptic inhibition in primary visual cortex. *Neuron* 73, 814–828. doi: 10.1016/j.neuron.2011.12.026

Jayakody, D. M. P., Friedland, P. L., Eikelboom, R. H., Martins, R. N., and Sohrabi, H. R. (2018). A novel study on association between untreated hearing loss and cognitive functions of older adults: baseline non-verbal cognitive assessment results. *Clin. Otolaryngol.* 43, 182–191. doi: 10.1111/coa.12937

Karns, C. M., Dow, M. W., and Neville, H. J. (2012). Altered cross-modal processing in the primary auditory cortex of congenitally deaf adults: a visual-somatosensory fMRI study with a double-flash illusion. *J. Neurosci.* 32, 9626–9638. doi: 10.1523/jneurosci.6488-11.2012

Kos, M. I., Deriaz, M., Guyot, J. P., and Pelizzone, M. (2009). What can be expected from a late cochlear implantation? *Int. J. Pediatr. Otorhinolaryngol.* 73, 189–193. doi: 10.1016/j.ijporl.2008.10.009

Kovari, E., Gold, G., Herrmann, F. R., Canuto, A., Hof, P. R., Bouras, C., et al. (2003). Lewy body densities in the entorhinal and anterior cingulate cortex predict cognitive deficits in Parkinson's disease. *Acta Neuropathol.* 106, 83–88. doi: 10.1007/s00401-003-0705-2

Lazard, D. S., and Giraud, A. L. (2017). Faster phonological processing and right occipito-temporal coupling in deaf adults signal poor cochlear implant outcome. *Nat. Commun.* 8:14872.

Leech, R., Kamourieh, S., Beckmann, C. F., and Sharp, D. J. (2011). Fractionating the default mode network: distinct contributions of the ventral and dorsal posterior cingulate cortex to cognitive control. *J. Neurosci.* 31, 3217–3224. doi: 10.1523/jneurosci.5626-10.2011

Lieberman, M. D., and Cunningham, W. A. (2009). Type I and Type II error concerns in fMRI research: re-balancing the scale. *Soc. Cogn. Affect. Neurosci.* 4, 423–428. doi: 10.1093/scan/nsp052

Lin, F. R. (2011). Hearing loss and cognition among older adults in the United States. *J. Gerontol. A Biol. Sci. Med. Sci.* 66, 1131–1136. doi: 10.1093/gerona/glr115

Lin, F. R., Ferrucci, L., Metter, E. J., An, Y., Zonderman, A. B., and Resnick, S. M. (2011a). Hearing loss and cognition in the baltimore longitudinal study of aging. *Neuropsychology* 25, 763–770. doi: 10.1037/a0024238

Lin, F. R., Metter, E. J., O'Brien, R. J., Resnick, S. M., Zonderman, A. B., and Ferrucci, L. (2011b). Hearing loss and incident dementia. *Arch. Neurol.* 68, 214–220.

Lin, F. R., Thorpe, R., Gordon-Salant, S., and Ferrucci, L. (2011c). Hearing loss prevalence and risk factors among older adults in the United States. *J. Gerontol. A Biol. Sci. Med. Sci.* 66, 582–590. doi: 10.1093/gerona/glr002

Lin, F. R., Yaffe, K., Xia, J., Xue, Q. L., Harris, T. B., Purchase-Helzner, E., et al. (2013). Hearing loss and cognitive decline in older adults. *JAMA Intern. Med.* 173, 293–299.

Lindenberger, U., and Baltes, P. B. (1994). Sensory functioning and intelligence in old age: a strong connection. *Psychol. Aging* 9, 339–355. doi: 10.1037/0882-7974.9.3.339

Lomber, S. G., Meredith, M. A., and Kral, A. (2010). Cross-modal plasticity in specific auditory cortices underlies visual compensations in the deaf. *Nat. Neurosci.* 13, 1421–1427. doi: 10.1038/nn.2653

Loughrey, D. G., Kelly, M. E., Kelley, G. A., Brennan, S., and Lawlor, B. A. (2018). Association of age-related hearing loss with cognitive function, cognitive impairment, and dementia: a systematic review and meta-analysis. *JAMA Otolaryngol. Head Neck Surg.* 144, 115–126. doi: 10.1001/jamaoto.2017.2513

MacSweeney, M., Woll, B., Campbell, R., McGuire, P. K., David, A. S., Williams, S. C., et al. (2002). Neural systems underlying British Sign Language and audio-visual English processing in native users. *Brain* 125, 1583–1593. doi: 10.1093/brain/awf153

Manoach, D. S., Press, D. Z., Thangaraj, V., Searl, M. M., Goff, D. C., Halpern, E., et al. (1999). Schizophrenic subjects activate dorsolateral prefrontal cortex during a working memory task, as measured by fMRI. *Biol. Psychiatry* 45, 1128–1137. doi: 10.1016/s0006-3223(98)00318-7

McKee, A. C., Au, R., Cabral, H. J., Kowall, N. W., Seshadri, S., Kubilus, C. A., et al. (2006). Visual association pathology in preclinical Alzheimer disease. *J. Neuropathol. Exp. Neurol.* 65, 621–630. doi: 10.1097/00005072-200606000-00010

Merabet, L. B., and Pascual-Leone, A. (2010). Neural reorganization following sensory loss: the opportunity of change. *Nat. Rev. Neurosci.* 11, 44–52. doi: 10.1038/nrn2758

Molloy, K., Griffiths, T. D., Chait, M., and Lavie, N. (2015). Inattentional deafness: visual load leads to time-specific suppression of auditory evoked responses. *J. Neurosci.* 35, 16046–16054. doi: 10.1523/jneurosci.2931-15.2015

Nestor, P. J., Fryer, T. D., Ikeda, M., and Hodges, J. R. (2003). Retrosplenial cortex (BA 29/30) hypometabolism in mild cognitive impairment (prodromal Alzheimer's disease). *Eur. J. Neurosci.* 18, 2663–2667. doi: 10.1046/j.1460-9568.2003.02999.x

Neville, H. J., and Lawson, D. (1987). Attention to central and peripheral visual space in a movement detection task: an event-related potential and behavioral study. II. Congenitally deaf adults. *Brain Res.* 405, 268–283. doi: 10.1016/0006-8993(87)90296-4

Nishimura, H., Hashikawa, K., Doi, K., Iwaki, T., Watanabe, Y., Kusuoka, H., et al. (1999). Sign language 'heard' in the auditory cortex. *Nature* 397:116. doi: 10.1038/16376

Okano, T., Yamamoto, Y., Kuzuya, A., Egawa, N., Kawakami, K., Furuta, I., et al. (2020). Development of the reading cognitive test Kyoto (ReaCT Kyoto) for early detection of cognitive decline in patients with hearing loss. *J. Alzheimers Dis.* 73, 981–990. doi: 10.3233/jad-190982

Petitto, L. A., Zatorre, R. J., Gauna, K., Nikelski, E. J., Dostie, D., and Evans, A. C. (2000). Speech-like cerebral activity in profoundly deaf people processing signed languages: implications for the neural basis of human language. *Proc. Natl. Acad. Sci. U.S.A.* 97, 13961–13966. doi: 10.1073/pnas.97.25.13961

Pierrot-Deseilligny, C., Muri, R. M., Ploner, C. J., Gaymard, B., Demeret, S., and Rivaud-Pechoux, S. (2003). Decisional role of the dorsolateral prefrontal cortex in ocular motor behaviour. *Brain* 126, 1460–1473. doi: 10.1093/brain/awg148

Poch, C., Toledano, R., Jimenez-Huete, A., Garcia-Morales, I., Gil-Nagel, A., and Campo, P. (2016). Differences in visual naming performance between patients with temporal lobe epilepsy associated with temporopolar lesions versus hippocampal sclerosis. *Neuropsychology* 30, 841–852. doi: 10.1037/neu0000269

Proksch, J., and Bavelier, D. (2002). Changes in the spatial distribution of visual attention after early deafness. *J. Cogn. Neurosci.* 14, 687–701. doi: 10.1162/08989290260138591

Quittner, A. L., Smith, L. B., Osberger, M. J., and Mitchell. (2010). The impact of audition on the development of visual attention. *Psychol. Sci.* 5, 347–353. doi: 10.1111/j.1467-9280.1994.tb00284.x

Rosen, M. L., Sheridan, M. A., Sambrook, K. A., Peverill, M. R., Meltzoff, A. N., and McLaughlin, K. A. (2018). The role of visual association cortex in associative memory formation across development. *J. Cogn. Neurosci.* 30, 365–380. doi: 10.1162/jocn_a_01202

Sandmann, P., Dillier, N., Eichele, T., Meyer, M., Kegel, A., Pascual-Marqui, R. D., et al. (2012). Visual activation of auditory cortex reflects maladaptive plasticity in cochlear implant users. *Brain* 135, 555–568. doi: 10.1093/brain/awr329

Scarpina, F., and Tagini, S. (2017). The stroop color and word test. *Front. Psychol.* 8:557. doi: 10.3389/fpsyg.2017.00557

Schormans, A. L., Typlt, M., and Allman, B. L. (2017). Crossmodal plasticity in auditory, visual and multisensory cortical areas following noise-induced hearing loss in adulthood. *Hear. Res.* 343, 92–107. doi: 10.1016/j.heares.2016.06.017

Shiell, M. M., Champoux, F., and Zatorre, R. J. (2014). Enhancement of visual motion detection thresholds in early deaf people. *PLoS One* 9:e90498. doi: 10.1371/journal.pone.0090498

Shiell, M. M., Champoux, F., and Zatorre, R. J. (2016). The right hemisphere planum temporale supports enhanced visual motion detection ability in deaf people: evidence from cortical thickness. *Neural Plast.* 2016:7217630.

Smith, L. B., Quittner, A. L., Osberger, M. J., and Miyamoto, R. (1998). Audition and visual attention: the developmental trajectory in deaf and hearing populations. *Dev. Psychol.* 34, 840–850. doi: 10.1037/0012-1649.34.5.840

Strelnikov, K., Rouger, J., Demonet, J. F., Lagleyre, S., Fraysse, B., Deguine, O., et al. (2013). Visual activity predicts auditory recovery from deafness after adult cochlear implantation. *Brain* 136, 3682–3695. doi: 10.1093/brain/awt274

Tombaugh, T. N., and Mcintyre, N. J. (1992). The mini-mental state examination: a comprehensive review. *J. Am. Geriatr. Soc.* 40, 922–935. doi: 10.1111/j.1532-5415.1992.tb01992.x

Tombaugh, T. N., Kozak, J., and Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Arch. Clin. Neuropsychol.* 14, 167–177. doi: 10.1016/s0887-6177(97)00095-4

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale*, Vol. 3. New York, NY: Psychological Corporation.

Wheeler, M. E., Shulman, G. L., Buckner, R. L., Miezin, F. M., Velanova, K., and Petersen, S. E. (2006). Evidence for separate perceptual reactivation and search processes during remembering. *Cereb. Cortex* 16, 949–959. doi: 10.1093/cercor/bhj037

World Health Organization (2021). *World Report on Hearing*. Geneva: World Health Organization.

Xu, L. C., Zhang, G., Zou, Y., Zhang, M. F., Zhang, D. S., Ma, H., et al. (2017). Abnormal neural activities of directional brain networks in patients with long-term bilateral hearing loss. *Oncotarget* 8, 84168–84179. doi: 10.18632/oncotarget.20361

Xu, X. M., Jiao, Y., Tang, T. Y., Lu, C. Q., Zhang, J., Salvi, R., et al. (2019a). Altered spatial and temporal brain connectivity in the salience network of sensorineural hearing loss and tinnitus. *Front. Neurosci.* 13:246. doi: 10.3389/fnins.2019.00246

Xu, X. M., Jiao, Y., Tang, T. Y., Zhang, J., Lu, C. Q., Luan, Y., et al. (2019b). Dissociation between cerebellar and cerebral neural activities in humans with long-term bilateral sensorineural hearing loss. *Neural Plast.* 2019: 8354849.

Xu, X. M., Jiao, Y., Tang, T. Y., Zhang, J., Lu, C. Q., Salvi, R., et al. (2019c). Sensorineural hearing loss and cognitive impairments: contributions of thalamus using multiparametric MRI. *J. Magn. Reson. Imaging* 50, 787–797. doi: 10.1002/jmri.26665

Xu, X. M., Jiao, Y., Tang, T. Y., Zhang, J., Salvi, R., and Teng, G. J. (2019d). Inefficient involvement of insula in sensorineural hearing loss. *Front. Neurosci.* 13:133. doi: 10.3389/fnins.2019.00133

Yusuf, P. A., Hubka, P., Tillein, J., and Kral, A. (2017). Induced cortical responses require developmental sensory experience. *Brain* 140, 3153–3165. doi: 10.1093/brain/awx286

Yusuf, P. A., Hubka, P., Tillein, J., Vinck, M., and Kral, A. (2021). Deafness weakens interareal couplings in the auditory cortex. *Front. Neurosci.* 14:625721. doi: 10.3389/fnins.2020.625721

Zhang, G. Y., Yang, M., Liu, B., Huang, Z. C., Chen, H., Zhang, P. P., et al. (2015). Changes in the default mode networks of individuals with long-term unilateral sensorineural hearing loss. *Neuroscience* 285, 333–342. doi: 10.1016/j.neuroscience.2014.11.034

Zhang, G. Y., Yang, M., Liu, B., Huang, Z. C., Li, J., Chen, J. Y., et al. (2016). Changes of the directional brain networks related with brain plasticity in patients with long-term unilateral sensorineural hearing loss. *Neuroscience* 313, 149–161. doi: 10.1016/j.neuroscience.2015.11.042

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

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 15 July 2021
doi: 10.3389/fnins.2021.670192

The Relative Weight of Temporal Envelope Cues in Different Frequency Regions for Mandarin Disyllabic Word Recognition

Zhong Zheng[1,2], Keyi Li[3], Yang Guo[4], Xinrong Wang[1,2], Lili Xiao[1,2], Chengqi Liu[1,2], Shouhuan He[5], Gang Feng[6] and Yanmei Feng[1,2]**

[1] Department of Otolaryngology-Head and Neck Surgery, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China, [2] Shanghai Key Laboratory of Sleep Disordered Breathing, Shanghai, China, [3] Sydney Institute of Language and Commerce, Shanghai University, Shanghai, China, [4] Ear, Nose, and Throat Institute and Otorhinolaryngology Department, Eye and ENT Hospital of Fudan University, Shanghai, China, [5] Department of Otolaryngology, Qingpu Branch of Zhongshan Hospital Affiliated to Fudan University, Shanghai, China, [6] The First Affiliated Hospital of Jinzhou Medical University, Jinzhou, China

OPEN ACCESS

Edited by:
Stephen Charles Van Hedger, Huron University College, Canada

Reviewed by:
Christian Füllgrabe, Loughborough University, United Kingdom
Nan Yan, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences (CAS), China

***Correspondence:**
Gang Feng
fghyc@163.com
Yanmei Feng
ymfeng@sjtu.edu.cn

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

Received: 20 February 2021
Accepted: 14 June 2021
Published: 15 July 2021

Citation:
Zheng Z, Li K, Guo Y, Wang X, Xiao L, Liu C, He S, Feng G and Feng Y (2021) The Relative Weight of Temporal Envelope Cues in Different Frequency Regions for Mandarin Disyllabic Word Recognition. Front. Neurosci. 15:670192. doi: 10.3389/fnins.2021.670192

Objectives: Acoustic temporal envelope (E) cues containing speech information are distributed across all frequency spectra. To provide a theoretical basis for the signal coding of hearing devices, we examined the relative weight of E cues in different frequency regions for Mandarin disyllabic word recognition in quiet.

Design: E cues were extracted from 30 continuous frequency bands within the range of 80 to 7,562 Hz using Hilbert decomposition and assigned to five frequency regions from low to high. Disyllabic word recognition of 20 normal-hearing participants were obtained using the E cues available in two, three, or four frequency regions. The relative weights of the five frequency regions were calculated using least-squares approach.

Results: Participants correctly identified 3.13–38.13%, 27.50–83.13%, or 75.00–93.13% of words when presented with two, three, or four frequency regions, respectively. Increasing the number of frequency region combinations improved recognition scores and decreased the magnitude of the differences in scores between combinations. This suggested a synergistic effect among E cues from different frequency regions. The mean weights of E cues of frequency regions 1–5 were 0.31, 0.19, 0.26, 0.22, and 0.02, respectively.

Conclusion: For Mandarin disyllabic words, E cues of frequency regions 1 (80–502 Hz) and 3 (1,022–1,913 Hz) contributed more to word recognition than other regions, while frequency region 5 (3,856–7,562) contributed little.

Keywords: relative weight, envelope cues, frequency region, Mandarin Chinese, disyllabic word

INTRODUCTION

The World Health Organization estimates that > 5% of the world's population (approximately 466 million people) suffer from disabling hearing loss (WHO, 2020). Approximately one-third of people over the age of 65 years suffer from different degrees of sensorineural hearing loss (SNHL), one of the most common forms of hearing loss (WHO, 2020). Cochlear implants (CIs) remain the only

effective device for restoring speech communication ability in patients with severe to profound SNHL (Tavakoli et al., 2015). In people with normal hearing, the cochlea converts speech signals into bioelectrical signals, which are transmitted through the auditory nerve to the brain so that listeners are able to sense various sounds. The basilar membrane in the cochlea can be regarded as a series of overlapping bandpass filters, each of which has its own unique characteristic frequency. When the basilar membrane is vibrated by sound, the sound signal of a characteristic frequency causes amplitude to peak at the corresponding basilar membrane partition (Smith et al., 2002). As electronic devices, CIs stimulate the auditory nerve *via* amplitude-modulated pulses that carry important temporal information of speech signals.

Speech is a complex acoustic signal and can be viewed in terms of two domains: temporal information and spectral information. Temporal information refers to information in speech signals with time-varying wave rates, which can be divided into the temporal envelope (E) below 50 Hz, periodic fluctuations in the range of 50–500 Hz, and temporal fine structure in the range of 500–10,000 Hz (Rosen, 1992). E cues contain temporal modulation information, which is most important for speech perception in quiet conditions, whereas the temporal fine structure can provide information in noisy environments and for tonal and pitch recognition (Smith et al., 2002; Xu and Pfingst, 2003; Moore, 2008; Ardoint and Lorenzi, 2010; Wang et al., 2016). Vocoder studies have shown that E modulation rates of 4–16 Hz are most important for speech intelligibility in quiet (Drullman et al., 1994a,b; Shannon et al., 1995). However, when speech has to be perceived against interfering speech, both slower and faster modulation rates, which are associated with prosodic (Füllgrabe et al., 2009) and fundamental frequency (Xu et al., 2002; Stone et al., 2008), respectively, become important for identification.

Recently, many researchers have investigated the relative importance of E cues from different frequency regions. Shannon et al. (2002) studied the influence of temporal E cues from different frequency regions on English recognition by removing specific spectral information. They found that the removal medium and high frequencies had greater impacts than low frequencies. This is supported by Apoux and Bacon (2004) who also reported that temporal E cues from different frequency regions are important in quiet conditions, while the high-frequency region (>2,500 Hz) is more important in noisy environments. In addition, using a classic high-pass and low-pass filtering experiment paradigm (French and Steinberg, 2005), Ardoint et al. (Ardoint and Lorenzi, 2010) demonstrated that E cues for frequency bands approximately 1,000–2,000 Hz are most important in French vowel–consonant–vowel speech recognition. The different frequency ranges and their respective importance in these investigations of non-tonal languages motivated us to examine the relative weight of E cues in different frequency regions for Mandarin Chinese speech recognition.

Mandarin Chinese is a tonal language and has the most first-language speakers of any language in the world. It includes 23 consonants, 38 vowels, and 5 tones. The 38 vowels consist of 9 monophthongs, 13 diphthongs and triphthongs, and 16 nasal finals. The tones include tone 1 (high-level), tone 2 (mid-rising), tone 3 (low-dipping), and tone 4 (high-falling) (Xu and Zhou, 2011). In addition, there is a fifth tone, usually called a neutral tone or tone 0, that occurs in unstressed syllables in multisyllabic words or connected speech (Yang et al., 2017). There are many polysyllabic words in Mandarin, most of which are disyllabic, and different tones can represent many different meanings (Nissen et al., 2005). Thus, the recognition of disyllabic words plays an important role in Mandarin speech recognition. Despite the large number of people who speak Mandarin as a mother tongue, there has been little emphasis on the relative weight of temporal and spectral information in different frequency regions for Mandarin Chinese. The present study intends to fill this knowledge gap. Our results may benefit CI wearers whose native language is Mandarin Chinese and ultimately improve their speech recognition performance and quality of life (McRackan et al., 2018).

Speech recognition is an interactive process between the speech characteristics of auditory signals and the long-term language knowledge of the listener, enabled by the decoding of speech through integrative bottom-up and top-down processes (Tuennerhoff and Noppeney, 2016). Speech recognition includes phonemes, syntax, and semantic recognition (Etchepareborda, 2003; Desroches et al., 2009). Phonemes are the smallest unit of sound that distinguish one word from another word in a language. Syntax refers to the meaning and interpretation of words, signs, and sentence structure, and depends on elements such as language environment and contextual information. Based on the semantic information of language, we can roughly judge the content range. Bottom-up mechanisms include phoneme recognition (Tuennerhoff and Noppeney, 2016), which disyllabic word recognition primarily relies on. Top-down mechanisms include syntactic and semantic information (Etchepareborda, 2003; Desroches et al., 2009). In an fMRI study, Tuennerhoff and Noppeney (2016) found that activations of unintelligible fine-structure speech were limited to the primary auditory cortices, but when top-down mechanisms made speech intelligible, the activation spread to posterior middle temporal regions, allowing for lexical access and speech recognition. Sentences can provide listeners with an envelope template where lexical and phonological constraints can help segment the acoustic signal into larger comprehensible temporal units, similar to the spatial "pop-out" phenomenon in visual object recognition (Dolan et al., 1997). Both top-down and bottom-up mechanisms are essential in the recognition of sentences.

While Mandarin Chinese word recognition relies more on tone recognition, sentence recognition can be inferred from context, which is consistent with the top-down mechanisms of speech recognition. Our team previously found that acoustic temporal E cues in frequency regions 80–502 Hz and 1,022–1,913 Hz contributed significantly to Mandarin sentence recognition (Guo et al., 2017). The present study builds on these findings and further investigates the relative weights of E cues for Mandarin disyllabic word recognition.

MATERIALS AND METHODS

Participants

We recruited a total of 20 participants (10 males and 10 females), who were graduates of Shanghai Jiao Tong University with normal audiometric thresholds ($\leq$ 20 dB HL) bilaterally at octave frequencies of 0.25–8 kHz. Their ages ranged from 22 to 28 (average, 24) years. All participants were native Mandarin speakers with no reported history of ear disease or hearing loss. Pure-tone audiometric thresholds were measured with a GSI-61 audiometer (Grason-Stadler, Madison, WI, United States) using standard audiometric procedures. No participant had any preceding exposure to the speech materials used in the present study. All participants provided signed consent forms before the experiment and were compensated on an hourly basis for their participation. The protocol was approved by the ethics committee of Shanghai Jiao Tong University Affiliated Sixth People's Hospital (ChiCTR-ROC-17013460), and the experiment was performed in accordance with the Declaration of Helsinki.

Signal Processing

Mandarin disyllabic speech test materials issued by the Beijing Institute of Otolaryngology were used for disyllable word recognition in quiet conditions. The test materials included 10 lists, each of which contained 50 disyllabic words covering 96.65% of the words used in daily life (Wang et al., 2007). The disyllable words were filtered into 30 contiguous frequency bands using zero-phase, third-order Butterworth filters (18 dB/octave slopes), ranging from 80 to 7,562 Hz (Li et al., 2016). Each band had an equivalent rectangular bandwidth for normal-hearing participants, simulating the frequency selectivity of the normal auditory system (Glasberg and Moore, 1990).

The E cues were extracted from each band using Hilbert decomposition followed by low-pass filtering at 64 Hz with a third-order Butterworth filter. E cues were then used to modulate the amplitude of a white-noise carrier. The modulated noise was filtered using the same bandpass filters and summed across frequency bands to form the frequency regions of acoustic E cues. We focused on the parameters used in clinics to assess hearing levels; i.e., low-frequency (< 500 Hz), medium-low-frequency (500–1,000 Hz), medium-frequency (1,000–2,000 Hz), medium-high-frequency (2,000–4,000 Hz), and high-frequency (4,000–8,000 Hz) regions. We chose cutoff frequencies closest to the audiometric frequencies 500, 1,000, 2,000, 4,000, and 8,000 Hz. Thus, frequency bands 1–8, 9–13, 14–18, 19–24, and 25–30 were combined to form frequency regions 1–5 (**Table 1**).

TABLE 1 | Cutoff frequencies of the 30 frequency bands.

Frequency regions	Band	Lower frequency (Hz)	Upper frequency (Hz)
1	1	80	115
	2	115	154
	3	154	198
	4	198	246
	5	246	300
	6	300	360
	7	360	427
	8	427	502
2	9	502	585
	10	585	677
	11	677	780
	12	780	894
	13	894	1,022
3	14	1,022	1,164
	15	1,164	1,322
	16	1,322	1,499
	17	1,499	1,695
	18	1,695	1,913
4	19	1,913	2,157
	20	2,157	2,428
	21	2,428	2,729
	22	2,729	3,066
	23	3,066	3,440
	24	3,440	3,856
5	25	3,856	4,321
	26	4,321	4,837
	27	4,837	5,413
	28	5,413	6,054
	29	6,054	6,767
	30	6,767	7,562

To investigate the role of different frequency regions in Mandarin disyllabic word recognition, participants were presented with acoustic E cues from two frequency regions (10 conditions, namely, Region 1 + 2, Region 1 + 3, Region 1 + 4, Region 1 + 5, Region 2 + 3, Region 2 + 4, Region 2 + 5, Region 3 + 4, Region3 + 5, and Region 4 + 5), three frequency regions (10 conditions, namely, Region 1 + 2 + 3, Region 1 + 2 + 4, Region 1 + 2 + 5, Region 1 + 3 + 4, Region 1 + 3 + 5, Region 1 + 4 + 5, Region 2 + 3 + 4, Region 2 + 3 + 5, Region 2 + 4 + 5, and Region 3 + 4 + 5), and four frequency regions (5 conditions, namely, Region 1 + 2 + 3 + 4, Region 1 + 3 + 4 + 5, Region 1 + 2 + 4 + 5, Region 1 + 2 + 3 + 5, and Region 2 + 3 + 4 + 5). To prevent the possible use of transitional band information (Warren et al., 2004; Li et al., 2015), the frequency regions including disyllabic word E cues and complementary frequency regions (i.e., noise-masking), were combined to present a speech-to-noise ratio of 16 dB. For example, the "Region 1 + 2" condition indicates that the participant was presented with a stimulus consisting of E cues for disyllabic words in frequency regions 1 and 2 with noise in frequency regions 3, 4, and 5. Similarly, "Region 1 + 2 + 3 + 4" indicates that the stimulus comprised acoustic temporal E cues in frequency regions 1, 2, 3, and 4 with noise in frequency region 5.

Testing Procedure

None of the participants had previously participated in perceptual experiments testing acoustic temporal E cues. The participants were tested individually in a double-walled, soundproof room. All stimuli were delivered *via* HD 205 II circumaural headphones (Sennheiser, Wedemark, Germany) at approximately 65 dB SPL (range, 60–75 dB SPL).

About 30 min of practice was provided before the formal test. The practice vocabulary comprised 50 disyllabic words (i.e., one list of the test material). Words were first presented under the "full region" condition and then experimental stimuli were presented randomly. Feedback was given during practice. To familiarize participants with the stimuli, they were allowed to listen to words repeatedly for any number of times before moving on to the next word until their performance stabilized.

In the formal test, participants were allowed to listen to words as many times as they wanted. Most participants listened to each word two or three times before moving on. All conditions and corresponding material lists were presented in random order for each participant to avoid any order effect. Participants were encouraged to repeat words as accurately as possible and to guess if necessary. No feedback was provided during the test. Each keyword in the disyllabic word lists was rated as correct or incorrect, and the results were recorded as the percentage of correct words under different conditions. The participants were allowed to take breaks whenever necessary. Each participant required approximately 1.5–2 h to complete the set of tests.

Statistical Analysis

Statistical analyses were conducted using the Statistical Package for Social Sciences version 22.0 (IBM Co., Armonk, NY, United States). Because of our small sample size, we used the Kruskal–Wallis test to analyze results from different test conditions for disyllabic word recognition. Pairwise comparison using results of different test conditions for disyllabic word recognition was performed with *post hoc* analysis (the Bonferroni test). The relative weights of the five frequency regions were calculated using the least-squares approach (Kasturi et al., 2002). The Mann–Whitney *U* test was used to compare the relative weights of the five frequency regions for Mandarin disyllabic word and sentence recognition.

RESULTS

Scores for Mandarin Disyllabic Word Recognition Across Conditions Using E Cues

Participants correctly identified 3.13–38.13% of words when presented with E cues in two frequency regions (**Figure 1**); scores were highest for Region 1 + 2 (38.13% correct) and lowest for Region 2 + 4 (3.13% correct). Thus, these conditions were unfavorable for participants to understand the meaning of disyllabic words. In addition, the percentage of correct responses differed significantly among frequency region combinations (H = 167.288, $p < 0.05$). Regions 1 + 2 and Region 1 + 3 had significantly higher scores than other conditions with two frequency regions (adjusted $p < 0.05$).

Participants correctly identified 27.50%–83.13% of words when presented E cues in three frequency regions (**Figure 2**). Region 1 + 2 + 3 and Regions 2 + 3 + 4 had high scores (78.75% and 83.13% correct, respectively), while Region 2 + 3 + 5 scored relatively low (27.50% correct). In addition, the percentage of correct responses differed significantly among frequency region combinations (H = 168.938, $p < 0.05$). Region 1 + 2 + 3 and Region 2 + 3 + 4 had significantly higher scores than other conditions with three frequency regions (adjusted $p < 0.05$).

Participants correctly identified 75.00–93.13% of words when presented with E cues in four frequency regions (**Figure 3**). Region 1 + 2 + 3 + 4 had the highest score among all conditions (93.13% correct), while Region 1 + 2 + 4 + 5 had the lowest score among combinations of four frequency regions (75.00% correct). In addition, the percentage of correct responses differed

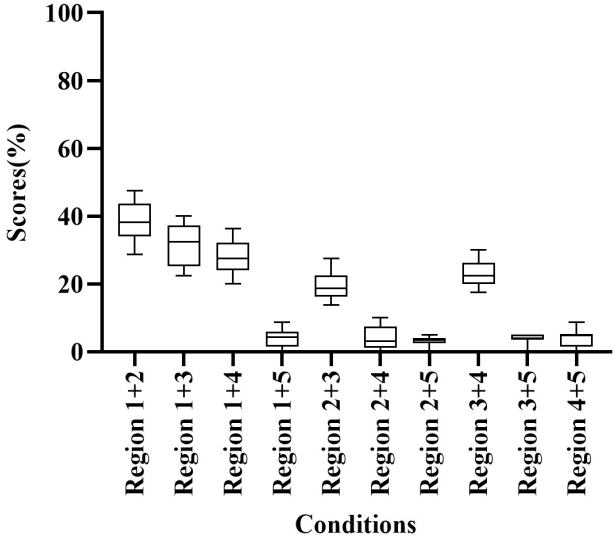

FIGURE 1 | Percent-correct scores for disyllabic word recognition using acoustic temporal envelope in two-frequency-region conditions.

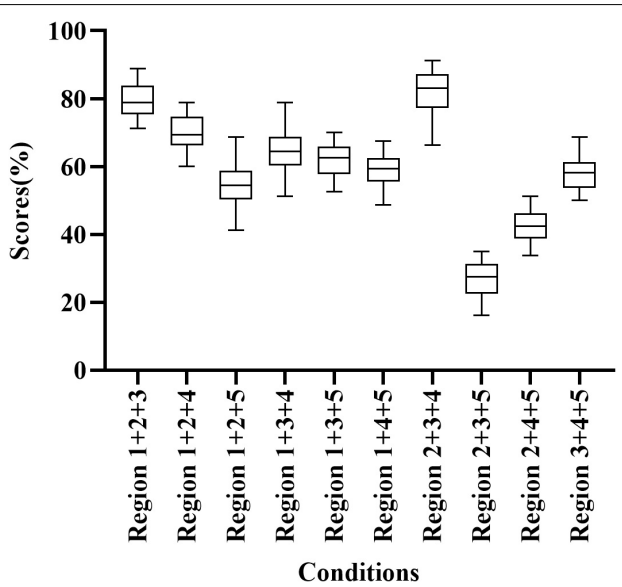

FIGURE 2 | Percent-correct scores for disyllabic word recognition using acoustic temporal envelope in three-frequency-region conditions.

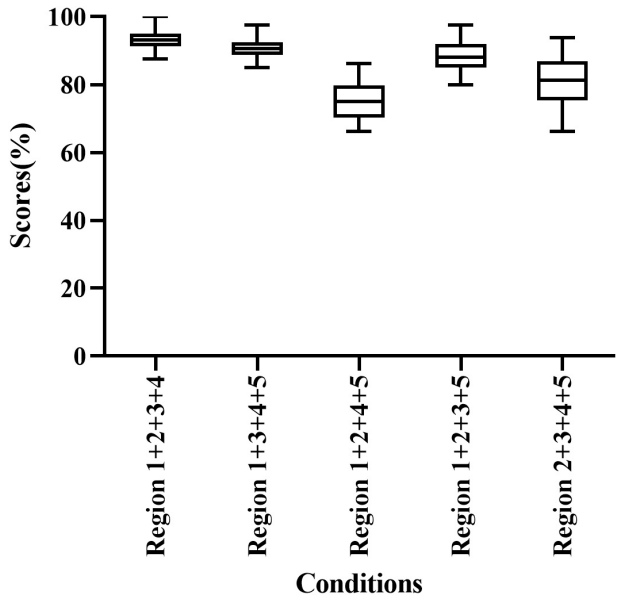

FIGURE 3 | Percent-correct scores for disyllabic word recognition using acoustic temporal envelope in four-frequency-region conditions.

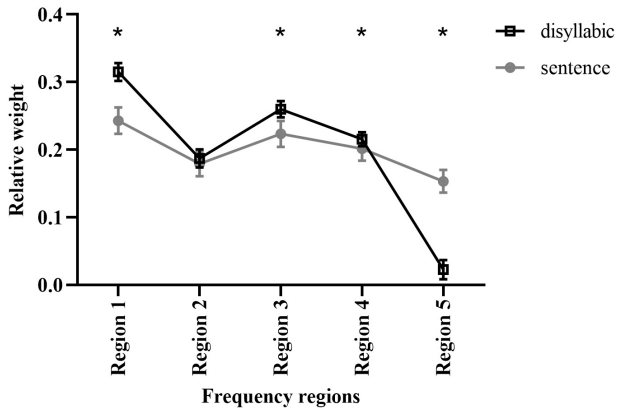

FIGURE 4 | The relative weights of different frequency regions for Mandarin disyllabic word and sentence recognition using acoustic temporal envelope. The data for Mandarin sentence recognition was adopted from a previous study (Guo et al., 2017). The error bars represent standard errors.
* Statistically significant ($p < 0.05$).

significantly among frequency region combinations (H = 60.762, $p < 0.05$). As the number of frequency regions increased, speech recognition scores increased and the magnitude of the difference between the groups decreased. Region 1 + 2 + 3 + 4 and Region 1 + 3 + 4 + 5 had significantly higher scores than other conditions with four frequency regions (adjusted $p < 0.05$).

Relative Weights of the Five Frequency Regions in Mandarin Disyllabic Word and Sentence Recognition

The relative weights of the five frequency regions for Mandarin disyllabic word recognition using acoustic temporal E cues were calculated using the least-squares approach (Kasturi et al., 2002). The strength of each frequency region was defined as a binary value (0 or 1) depending on whether the frequency region was present. The weight of each frequency region was then calculated by predicting the participant's response as a linear combination of each frequency region's intensity. The original weights for the five frequency regions of each participant were transformed to relative weights by summing their values, and each frequency region weight was represented as the original weight divided by the sum of all weights of the five frequency regions. Thus, the sum of the relative weights of the five frequency regions was equal to 1.0. The mean relative weights of frequency regions 1–5 were 0.31, 0.19, 0.26, 0.22, and 0.02, respectively (**Figure 4**). The relative weights differed significantly among frequency regions (H = 94.221, $p < 0.05$).

Our previous reports showed that the relative weights of the E cues from frequency regions 1–5 for Mandarin sentence recognition were 0.25, 0.18, 0.22, 0.20, and 0.15, respectively (Guo et al., 2017). Thus, the observed trends between sentence and disyllabic word recognition were similar. The E cues of frequency regions 1 and 3 contributed more to Mandarin disyllabic word and sentence recognition than those of the other frequency regions. Mandarin disyllabic word and sentence recognition had significantly different relative weights for frequency regions 1, 3, 4, and 5 ($p < 0.05$; **Figure 4**) but not for frequency region 2.

DISCUSSION

Vocoder technology is often used to simulate the signal processing of CIs when studying the function of E cues in speech recognition. The vocoder separates the speech signal into different frequency bands through bandpass filters. The E cues in different frequency bands are extracted and used to modulate white noise or sine waves. Eventually, the summed E cues of each frequency band are presented to the participants for perceptual experiments (Kim et al., 2015; Rosen and Hui, 2015). Previous research has shown that as the number of frequency bands segmented by a vocoder increases, the speech recognition rate of the subjects gradually increases (Shannon et al., 2004; Xu et al., 2005; Xu and Zheng, 2007). However, it is impossible to increase the number of intracochlear electrodes infinitely due to a number of constraints, including interference between adjacent electrodes (Shannon et al., 2004). Therefore, when the number of electrodes is fixed, it is necessary to study the relative importance of E cues in different frequency regions for speech recognition, especially given the trade-off between the number of spectral channels and the amount of temporal information (Xu and Pfingst, 2008).

It is worth noting that the bandpass filters used in the present study had a fairly shallow slope (18 dB/octave). The listening ability beyond the cutoff frequency that filtered the stimulus is called "off-frequency listening." However, our normal-hearing participants may not have been able to efficiently use the speech cues in the off-frequency bands because, in

vocoder processing, the fine structure of each frequency band is replaced by white noise.

The present study explored the relative importance of E cues across different frequency regions for Mandarin disyllabic word recognition. We found that E cues in different frequency regions contributed differentially to recognition scores. For Mandarin Chinese disyllabic words, frequency region 1 had the highest weight (0.31; **Figure 4**). Thus, the low-frequency region is most important for Mandarin recognition, which is consistent with a previous report on Mandarin Chinese sentence recognition (Guo et al., 2017). However, this is not consistent with the results of Ardoint et al., who found that E cues in higher regions (1,800–7,300 Hz) contributed more strongly to French consonant recognition (Ardoint et al., 2011). Explanations for this difference include differences in speech materials, since previous studies have shown that the type of speech material may have a strong impact on the use of acoustic temporal fine structure information (Lunner et al., 2012); the cutoff frequencies defined by different frequency regions used in different experiments; the methods of processing stimuli and evaluating weights; and, most importantly, differences inherent to the languages themselves. As a tone language, Mandarin Chinese has disyllabic words with different tones that can contain different lexical meanings (Fu et al., 1998; Wei et al., 2004). Fundamental frequency (F0) and its harmonics are the primary cues for lexical tone recognition, and tone contours play an important role in tonal language speech intelligibility (Feng et al., 2012). Kuo et al. (2008) found that when F0 information is provided, participants consistently have tone recognition rates of $> 90\%$; however, without F0 information, E cue information contributes to tone recognition to a lesser extent. Considering that F0 information is mainly in the low-frequency region, which plays an important role in tone recognition (Wong et al., 2007), the low-frequency region (i.e., region 1) likely has a strong influence on the recognition of Mandarin Chinese.

The E cues information in the intermediate frequencies (i.e., region 3, 1,022–1,913 Hz) is important for speech recognition, which is consistent with previous results. Kasturi et al. (2002) found that when E cues in a band centered at 1,685 Hz were removed, English vowel and consonant recognition declined. In addition, Ardoint et al. (Ardoint and Lorenzi, 2010) found that E cues conveyed important distinct phonetic cues in frequency regions between 1,000 and 2,000 Hz. These results indicate that E cues in the frequency band 1,000–2,000 Hz are important for speech recognition regardless of language.

We found that relative weights differed significantly between Mandarin disyllabic word and sentence recognition in frequency regions 1, 3, 4, and 5 ($p < 0.05$) but not in frequency region 2. This may have been due to differences in speech materials as described earlier and the fact that tone plays a more important role in understanding disyllabic words (Feng et al., 2012; Wong and Strange, 2017). Auditory speech input is rapidly and automatically bound by the participant into a speech representation in a short-term memory buffer. If this information matches the speech representation in long-term memory, relatively automatic and effortless lexical acquisition occurs. Mismatches are effortlessly controlled using higher-level linguistic knowledge such as semantics or syntactic contexts. Therefore, when bottom-up processes fail, controlled processing and sentence contextual information are used (Rönnberg et al., 2013). It should be noted that relationship between working memory (cognitive processing) and speech-in-noise intelligibility is less evident for younger participants than older hearing-impaired participants (Füllgrabe and Rosen, 2016). In this study, a common feature of most sentence-recognition models is that long-term language knowledge helps the participant choose the appropriate phonological and lexical candidates, thereby allowing the participant to make the correct selection step by step based on the acoustic speech characteristics of the speech signal (McClelland and Elman, 1986; Norris et al., 2016). The results of this study showed that Mandarin disyllabic recognition relies more on tone recognition, which is consistent with a bottom-up mechanism, whereas sentence recognition is inferred from context, which is consistent with the top-down mechanism of speech recognition.

We found that Region 1 + 2 has the highest score among two-frequency regions. In addition, Region 2 + 3, Region 2 + 3 + 4, and Region 1 + 2 + 4 also yielded high scores; however, Region 2 had the second-lowest relative weight. Warren et al. (1995) reported that regions centered at 370 and 6,000 Hz strongly synergize but provide little information when presented separately. Healy et al. (Healy and Warren, 2003) also found that unintelligible individual speech regions became comprehensible when combined. We assessed participant performance under all possible combinations of frequency regions and found that frequency region 2 had synergistic effects when combined with adjacent regions (i.e., frequency regions 1 or 3), which led to increased word recognition scores.

This study had some limitations. First, participants were well educated, and the influence of education level has not been evaluated in previous studies. In addition, since cognitive abilities are associated with changes in speech processing with age (even in the absence of audiometric hearing loss) (Füllgrabe et al., 2014), our findings may not be directly applicable to CI performance in different age groups. Future studies are needed to evaluate the factors of age, education level, and cognitive ability.

Overall, E cues contained in low-frequency spectral regions are more important in quiet environments for Mandarin disyllabic word recognition than for non-tonal languages such as English. These differences may be determined by the tone characteristics of Mandarin Chinese, but the influence of signal extraction parameters, test materials, and test environments cannot be excluded.

CONCLUSION

1. We found that E cues in frequency regions 1 (80–502 Hz) and 3 (1,022–1,913 Hz) significantly contributed to Mandarin disyllabic word recognition in quiet.

2. In contrast to English speech recognition, the low-frequency region contributed strongly to Mandarin Chinese disyllabic word recognition due to the tonal nature of the language.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the studies involving human participants were reviewed and approved by Ethics Committee of the Sixth People's Hospital affiliated to the Shanghai Jiao Tong University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ZZ: conceptualization and writing—original draft. KL and YG: methodology and data curation. XW, LX, and CL: investigation. SH: supervision. GF: validation and project administration. YF: conceptualization, resources, writing—review and editing, and funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the National Natural Science Foundation of China (Grant No. 81771015), the Shanghai Municipal Commission of Science and Technology (Grant No. 18DZ2260200), and the International Cooperation and Exchange of the National Natural Science Foundation of China (Grant No. 81720108010).

ACKNOWLEDGMENTS

We would like to thank all the patients and their families for supporting our work.

REFERENCES

Apoux, F., and Bacon, S. P. (2004). Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise. *J. Acoust. Soc. Am.* 116, 1671–1680. doi: 10.1121/1.1781329

Ardoint, M., Agus, T., Sheft, S., and Lorenzi, C. (2011). Importance of temporal-envelope speech cues in different spectral regions. *J. Acoust. Soc. Am.* 130, El115–El121.

Ardoint, M., and Lorenzi, C. (2010). Effects of lowpass and highpass filtering on the intelligibility of speech based on temporal fine structure or envelope cues. *Hear Res.* 260, 89–95. doi: 10.1016/j.heares.2009.12.002

Desroches, A. S., Newman, R. L., and Joanisse, M. F. (2009). Investigating the time course of spoken word recognition: electrophysiological evidence for the influences of phonological similarity. *J. Cogn. Neurosci.* 21, 1893–1906. doi: 10.1162/jocn.2008.21142

Dolan, R. J., Fink, G. R., Rolls, E., Booth, M., Holmes, A., Frackowiak, R. S., et al. (1997). How the brain learns to see objects and faces in an impoverished context. *Nature* 389, 596–599. doi: 10.1038/39309

Drullman, R., Festen, J. M., and Plomp, R. (1994a). Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.* 95(5 Pt 1), 2670–2680. doi: 10.1121/1.409836

Drullman, R., Festen, J. M., and Plomp, R. (1994b). Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.* 95, 1053–1064. doi: 10.1121/1.408467

Etchepareborda, M. C. (2003). Intervention in dyslexic disorders: phonological awareness training. *Rev. Neurol.* 36, S13–S19.

Feng, Y. M., Xu, L., Zhou, N., Yang, G., and Yin, S. K. (2012). Sine-wave speech recognition in a tonal language. *J. Acoust. Soc. Am.* 131, EL133–EL138.

French, N. R., and Steinberg, J. C. (2005). Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am.* 19, 90–119. doi: 10.1121/1.1916407

Fu, Q. J., Zeng, F. G., Shannon, R. V., and Soli, S. D. (1998). Importance of tonal envelope cues in Chinese speech recognition. *J. Acoust. Soc. Am.* 104, 505–510. doi: 10.1121/1.423251

Füllgrabe, C., and Rosen, S. (2016). On the (un)importance of working memory in speech-in-noise processing for listeners with normal hearing thresholds. *Front. Psychol.* 7:1268.

Füllgrabe, C., Moore, B. C., and Stone, M. A. (2014). Age-group differences in speech identification despite matched audiometrically normal hearing: contributions from auditory temporal processing and cognition. *Front. Aging Neurosci.* 6:347.

Füllgrabe, C., Stone, M. A., and Moore, B. C. (2009). Contribution of very low amplitude-modulation rates to intelligibility in a competing-speech task (L). *J. Acoust. Soc. Am.* 125, 1277–1280. doi: 10.1121/1.3075591

Glasberg, B. R., and Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear Res.* 47, 103–138. doi: 10.1016/0378-5955(90)90170-t

Guo, Y., Sun, Y., Feng, Y., Zhang, Y., and Yin, S. (2017). The relative weight of temporal envelope cues in different frequency regions for Mandarin sentence recognition. *Neural Plast.* 2017:7416727.

Healy, E. W., and Warren, R. M. (2003). The role of contrasting temporal amplitude patterns in the perception of speech. *J. Acoust. Soc. Am.* 113, 1676–1688. doi: 10.1121/1.1553464

Kasturi, K., Loizou, P. C., Dorman, M., and Spahr, T. (2002). The intelligibility of speech with "holes" in the spectrum. *J. Acoust. Soc. Am.* 112(3 Pt 1), 1102–1111.

Kim, B. J., Chang, S. A., Yang, J., Oh, S. H., and Xu, L. (2015). Relative contributions of spectral and temporal cues to Korean phoneme recognition. *PLoS One* 10:e0131807. doi: 10.1371/journal.pone.0131807

Kuo, Y. C., Rosen, S., and Faulkner, A. (2008). Acoustic cues to tonal contrasts in Mandarin: implications for cochlear implants. *J. Acoust. Soc. Am.* 123, 2815–2824. doi: 10.1121/1.2896755

Li, B., Hou, L., Xu, L., Wang, H., Yang, G., Yin, S., et al. (2015). Effects of steep high-frequency hearing loss on speech recognition using temporal fine structure in low-frequency region. *Hear Res.* 326, 66–74. doi: 10.1016/j.heares.2015.04.004

Li, B., Wang, H., Yang, G., Hou, L., Su, K., Feng, Y., et al. (2016). The importance of acoustic temporal fine structure cues in different spectral regions for Mandarin sentence recognition. *Ear Hear* 37, e52–e56.

Lunner, T., Hietkamp, R. K., Andersen, M. R., Hopkins, K., and Moore, B. C. (2012). Effect of speech material on the benefit of temporal fine structure information in speech for young normal-hearing and older hearing-impaired participants. *Ear Hear* 33, 377–388. doi: 10.1097/aud.0b013e3182387a8c

McClelland, J. L., and Elman, J. L. (1986). The TRACE model of speech perception. *Cog. Psychol.* 18, 1–86. doi: 10.1016/0010-0285(86)90015-0

McRackan, T. R., Bauschard, M., Hatch, J. L., Franko-Tobin, E., Droghini, H. R., Nguyen, S. A., et al. (2018). Meta-analysis of quality-of-life improvement after cochlear implantation and associations with speech recognition abilities. *Laryngoscope* 128, 982–990. doi: 10.1002/lary.26738

Moore, B. C. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *J. Assoc. Res. Otolaryngol.* 9, 399–406. doi: 10.1007/s10162-008-0143-x

Nissen, S. L., Harris, R. W., Jennings, L. J., Eggett, D. L., and Buck, H. (2005). Psychometrically equivalent Mandarin bisyllabic speech discrimination

materials spoken by male and female talkers. *Int. J. Audiol.* 44, 379–390. doi: 10.1080/14992020500147615

Norris, D., McQueen, J. M., and Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Lang. Cogn. Neurosci.* 31, 4–18. doi: 10.1080/23273798.2015.1081703

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., et al. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Front. Syst. Neurosci.* 7:31.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 336, 367–373. doi: 10.1098/rstb.1992.0070

Rosen, S., and Hui, S. N. (2015). Sine-wave and noise-vocoded sine-wave speech in a tone language: acoustic details matter. *J. Acoust. Soc. Am.* 138, 3698–3702. doi: 10.1121/1.4937605

Shannon, R. V., Fu, Q. J., and Galvin, J. (2004). The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta Otolaryngol. Suppl.* 552, 50–54. doi: 10.1080/03655230410017562

Shannon, R. V., Galvin, J. J., and Baskent, D. (2002). Holes in hearing. *J. Assoc. Res. Otolaryngol.* 3, 185–199. doi: 10.1007/s101620020021

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science* 270, 303–304. doi: 10.1126/science.270.5234.303

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature* 416, 87–90. doi: 10.1038/416087a

Stone, M. A., Füllgrabe, C., and Moore, B. C. (2008). Benefit of high-rate envelope cues in vocoder processing: effect of number of channels and spectral region. *J. Acoust. Soc. Am.* 124, 2272–2282. doi: 10.1121/1.2968678

Tavakoli, M., Jalilevand, N., Kamali, M., Modarresi, Y., and Zarandy, M. M. (2015). Language sampling for children with and without cochlear implant: MLU, NDW, and NTW. *Int. J. Pediatr. Otorhinolaryngol.* 79, 2191–2195. doi: 10.1016/j.ijporl.2015.10.001

Tuennerhoff, J., and Noppeney, U. (2016). When sentences live up to your expectations. *NeuroImage* 124, 641–653. doi: 10.1016/j.neuroimage.2015.09.004

Wang, S., Dong, R., Liu, D., Zhang, L., and Xu, L. (2016). The Relative contributions of temporal envelope and fine structure to Mandarin lexical tone perception in auditory neuropathy spectrum disorder. *Adv. Exp. Med. Biol.* 894, 241–248. doi: 10.1007/978-3-319-25474-6_25

Wang, S., Mannell, R., Newall, P., Zhang, H., and Han, D. (2007). Development and evaluation of Mandarin disyllabic materials for speech audiometry in China. *Int. J. Audiol.* 46, 719–731. doi: 10.1080/14992020701558511

Warren, R. M., Bashford, J. A., and Lenz, P. W. (2004). Intelligibility of bandpass filtered speech: steepness of slopes required to eliminate transition band contributions. *J. Acoust. Soc. Am.* 115, 1292–1295. doi: 10.1121/1.1646404

Warren, R. M., Riener, K. R., Bashford, J. A. Jr., and Brubaker, B. S. (1995). Spectral redundancy: intelligibility of sentences heard through narrow spectral slits. *Percept. Psychophys.* 57, 175–182. doi: 10.3758/bf03206503

Wei, C. G., Cao, K., and Zeng, F. G. (2004). Mandarin tone recognition in cochlear-implant subjects. *Hear Res.* 197, 87–95. doi: 10.1016/j.heares.2004.06.002

WHO (2020). *Deafness and Hearing Loss*. Available Online at: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss (accessed April 1, 2021).

Wong, L. L., Ho, A. H., Chua, E. W., and Soli, S. D. (2007). Development of the Cantonese speech intelligibility index. *J. Acoust. Soc. Am.* 121, 2350–2361.

Wong, P., and Strange, W. (2017). Phonetic complexity affects children's Mandarin tone production accuracy in disyllabic words: a perceptual study. *PLoS One* 12:e0182337. doi: 10.1371/journal.pone.0182337

Xu, L., and Pfingst, B. E. (2003). Relative importance of temporal envelope and fine structure in lexical-tone perception. *J. Acoust. Soc. Am.* 114(6 Pt 1), 3024–3027. doi: 10.1121/1.1623786

Xu, L., and Pfingst, B. E. (2008). Spectral and temporal cues for speech recognition: implications for auditory prostheses. *Hear Res.* 242, 132–140. doi: 10.1016/j.heares.2007.12.010

Xu, L., and Zheng, Y. (2007). Spectral and temporal cues for phoneme recognition in noise. *J. Acoust. Soc. Am.* 122:1758. doi: 10.1121/1.2767000

Xu, L., and Zhou, N. (2011). "Tonal languages and cochlear implants," in *Auditory Prostheses: New Horizons*, eds F. Zeng, A. Popper, and R. Fay (New York, NY: Springer), 341–364. doi: 10.1007/978-1-4419-9434-9_14

Xu, L., Thompson, C. S., and Pfingst, B. E. (2005). Relative contributions of spectral and temporal cues for phoneme recognition. *J. Acoust. Soc. Am.* 117, 3255–3267. doi: 10.1121/1.1886405

Xu, L., Tsai, Y., and Pfingst, B. E. (2002). Features of stimulation affecting tonal-speech perception: implications for cochlear prostheses. *J. Acoust. Soc. Am.* 112, 247–258. doi: 10.1121/1.1487843

Yang, J., Zhang, Y., Li, A., and Xu, L. (2017). "On the duration of Mandarin tones," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech 2017)*, (Stockholm).

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

frontiers in Psychology

ORIGINAL RESEARCH
published: 09 September 2021
doi: 10.3389/fpsyg.2021.714485

Semantic Predictability Facilitates Comprehension of Degraded Speech in a Graded Manner

Pratik Bhandari[1,2]*, *Vera Demberg*[2,3] and *Jutta Kray*[1]

[1]Department of Psychology, Saarland University, Saarbrücken, Germany, [2]Department of Language Science and Technology, Saarland University, Saarbrücken, Germany, [3]Department of Computer Science, Saarland University, Saarbrücken, Germany

Previous studies have shown that at moderate levels of spectral degradation, semantic predictability facilitates language comprehension. It is argued that when speech is degraded, listeners have *narrowed expectations* about the sentence endings; i.e., semantic prediction may be limited to only most highly predictable sentence completions. The main objectives of this study were to (i) examine whether listeners form narrowed expectations or whether they form predictions across a wide range of probable sentence endings, (ii) assess whether the facilitatory effect of semantic predictability is modulated by perceptual adaptation to degraded speech, and (iii) use and establish a sensitive metric for the measurement of language comprehension. For this, we created 360 German Subject-Verb-Object sentences that varied in semantic predictability of a sentence-final target word in a graded manner (high, medium, and low) and levels of spectral degradation (1, 4, 6, and 8 channels noise-vocoding). These sentences were presented auditorily to two groups: One group ($n = 48$) performed a listening task in an unpredictable channel context in which the degraded speech levels were randomized, while the other group ($n = 50$) performed the task in a predictable channel context in which the degraded speech levels were blocked. The results showed that at 4 channels noise-vocoding, response accuracy was higher in high-predictability sentences than in the medium-predictability sentences, which in turn was higher than in the low-predictability sentences. This suggests that, in contrast to the *narrowed expectations* view, comprehension of moderately degraded speech, ranging from low- to high- including medium-predictability sentences, is facilitated in a graded manner; listeners probabilistically preactivate upcoming words from a wide range of semantic space, not limiting only to highly probable sentence endings. Additionally, in both channel contexts, we did not observe learning effects; i.e., response accuracy did not increase over the course of experiment, and response accuracy was higher in the predictable than in the unpredictable channel context. We speculate from these observations that when there is no trial-by-trial variation of the levels of speech degradation, listeners adapt to speech quality at a long timescale; however, when there is a trial-by-trial variation of the high-level semantic feature (e.g., sentence predictability), listeners do not adapt to low-level perceptual property (e.g., speech quality) at a short timescale.

Keywords: speech perception, language comprehension, bottom-up processing, top-down prediction, semantic prediction, probabilistic prediction, perceptual adaptation, noise-vocoded speech

OPEN ACCESS

Edited by:
Howard Charles Nusbaum, University of Chicago, United States

Reviewed by:
Jed A. Meltzer, Baycrest Hospital, Canada
Kazuo Ueda, Kyushu University, Japan

***Correspondence:**
Pratik Bhandari
pratikb@coli.uni-saarland.de

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

Received: *25 May 2021*
Accepted: *06 August 2021*
Published: *09 September 2021*

Citation:
Bhandari P, Demberg V and Kray J (2021) Semantic Predictability Facilitates Comprehension of Degraded Speech in a Graded Manner. Front. Psychol. 12:714485. doi: 10.3389/fpsyg.2021.714485

INTRODUCTION

Understanding speech is highly automatized and seemingly easy when conditions are optimal. However, in our day-to-day communication, conditions are often far from being optimal. Intelligibility and comprehension of speech can be compromised at the source (speaker), at the receiver (listener), and at the transmission of the speech signal (environmental factor; Shannon, 1948). Ambient noise is an environmental factor that distorts the speech signal and renders it difficult to understand. For example, the noise coming from people talking in the background might make it difficult for you to understand what your friend is saying, while you are chatting in a café. Similarly, a conversation with a friend over the phone can be corrupted by a poor transmission of the speech signal which in turn hampers language comprehension. Interestingly, although the speech signal is sometimes bad or the environment is noisy, listeners do not always fail to understand what a friend is saying in the café or over the phone. Instead, listeners are successful in understanding distorted speech by utilizing context information which contains information in a given situation about a topic of conversation, semantic and syntactic information of a sentence structure, world knowledge, visual information, etc. (Kaiser and Trueswell, 2004; Knoeferle et al., 2005; Altmann and Kamide, 2007; Xiang and Kuperberg, 2015; for reviews, Stilp, 2020). The goals of the present study were threefold: First, to examine the interplay between perceptual and cognitive processing during language comprehension to answer the question whether the predictability of the sentence context facilitates language comprehension in a graded manner primarily when the speech signal is distorted, second, to assess whether perceptual adaptation influences the interplay between perceptual and language processing, and third, to establish a sensitive metric that takes into account the use of context in language comprehension.

Language Comprehension and Sentence Context (Predictability)

Research from various domains of cognitive (neuro)science, like emotion, vision, odor, and proprioception, has shown that predicting upcoming events influences human perception and cognition (Stadler et al., 2012; Clark, 2013; Seth, 2013; Marques et al., 2018). There is also a large body of evidence from psycholinguistics and cognitive neuroscience of language, suggesting that human language comprehension is predictive in nature (Lupyan and Clark, 2015; Pickering and Gambi, 2018; see also Huettig and Mani, 2016). Empirical evidence from a number of studies suggests that readers or listeners predict upcoming words in a sentence when the words are predictable from the preceding context (for reviews, Staub, 2015; Kuperberg and Jaeger, 2016). For instance, words that are highly predictable from the preceding context are read faster and are skipped compared to the less predictable ones (Ehrlich and Rayner, 1981; Frisson et al., 2005; Staub, 2011). Applying the visual world paradigm, several studies have demonstrated that participants show anticipatory eye movements toward the picture of the word predictable from the sentence context (Altmann and Kamide, 1999; Kamide et al., 2003; Jachmann et al., 2019). The sentence-final word in a highly predictable sentence context (e.g., "*She dribbles a ball.*") elicits a smaller N400, a negative going EEG component that peaks around 400 ms post-stimulus and is considered as a neural marker of context-based expectation, than that in a less predictable sentence context (e.g., "*She buys a ball.*"; Kutas and Hillyard, 1984; Federmeier et al., 2007; for reviews, see Kutas and Federmeier, 2011; see also Brouwer and Crocker, 2017). Similarly, event-related words (e.g., "luggage") elicited reduced N400 compared to event-unrelated words (e.g., "vegetables") which were not predictable from the context (e.g., in a "travel" scenario; Metusalem et al., 2012). In sum, as the sentence context builds up, listeners make predictions about upcoming words in the sentence, and these in turn facilitate language comprehension. Here, we will investigate whether individuals make use of the predictability of the sentence context when perceptual processing is hampered due to a bad quality of the speech signal.

Language Comprehension Under Reduced Quality of the Speech Signal

The detrimental effect of distortion of speech signal in language comprehension and speech intelligibility has been investigated for several types of artificial distortions, like multi-talker babble noise, reverberation, time compression, and noise-vocoding. For instance, it has been shown that speech intelligibility and comprehension decreases (a) with a decrease in signal-to-noise ratio under multi-talker babble noise conditions (e.g., Fontan et al., 2015), (b) faster rate of speech (e.g., Wingfield et al., 2006), and (c) longer reverberation time (e.g., Xia et al., 2018).

Similarly, noise-vocoding also impedes speech intelligibility. Noise-vocoding is an effective method to parametrically vary and control the intelligibility of speech in a graded manner. Noise-vocoding distorts speech by dividing a speech signal into specific frequency bands corresponding to the number of vocoder channels. The frequency bands are analogous to the electrodes of cochlear implant (Shannon et al., 1995, 1998). The amplitude envelope within each band is extracted and is used to modulate noise of the same bandwidth. This renders vocoded speech harder to understand by replacing the fine structure of the speech signal with noise while preserving the temporal characteristics and periodicity of perceptual cues. With the increase in number of channels, more frequency-specific information becomes available, spectral resolution of the speech signal increases, and hence, speech becomes more intelligible; for example, speech processed through 8 channels noise-vocoding is more intelligible than the one processed through 4 channels noise-vocoding (Loizou et al., 1999; Shannon et al., 2004). The level of degradation, i.e., the number of channels used for noise-vocoding, required for the same level of task accuracy can vary from 3 to 30 or more depending on the method implemented for noise-vocoding (e.g., Ueda et al., 2018), participant variables (age and language experience), test materials (words, sentences, and accented speech), and listening conditions (speech in quiet and speech with background

noise; Shannon et al., 2004). Here, we will systematically vary the level of speech degradation by noise-vocoding of unaccented speech to determine whether listeners benefit from the sentence context for language comprehension when the signal quality is not too bad or too good, hence at moderate levels of speech degradation.

Predictive Processing and Language Comprehension Under Degraded Speech

Top-down predictive and bottom-up perceptual processes interact dynamically in language comprehension. In a noisy environment, when the bottom-up perceptual input is less reliable, it has been shown that participants rely more on top-down predictions by narrowing down the predictions to smaller sets of semantic categories or words (e.g., Strauß et al., 2013; see also Corps and Rabagliati, 2020). Obleser and colleagues (Obleser et al., 2007; Obleser and Kotz, 2010, 2011), for instance, used sentences of two levels of semantic predictability (high and low) and systematically degraded the speech signal by passing it through various numbers of noise-vocoding channels ranging from 1 to 32 in a series of behavioral and neuroimaging studies. They found that semantic predictability facilitated language comprehension at a moderate level of speech degradation. That is, participants relied more on the sentence context when the speech signal was degraded but intelligible enough than when it was not degraded or was highly degraded. At such moderate levels of speech degradation, accuracy of word recognition was found to be higher for highly predictable target words than for less expected target words (Obleser and Kotz, 2010). For the extremes, i.e., when the speech signal was highly degraded or when it was clearly intelligible, the word recognition accuracy was similar across both levels of sentence predictability, meaning that predictability did not facilitate language comprehension. The conclusion of these findings is that at moderate levels of degradation, participants rely more on the top-down prediction generated by the sentence context and less on the bottom-up processing of unclear, less reliable, and degraded speech signal (Obleser, 2014). Reliance on prediction results in higher word recognition accuracy for target words with high-cloze probability than for the target words with low-cloze probability. In the case of a heavily degraded speech signal, participants may not be able to understand the sentence context and, therefore, be unable to form predictions of the target word, or their cognitive resources may already be occupied by decoding the signal, leaving little room for making predictions. Thus, there is no differential effect of levels of sentence predictability. On the other extreme, when the speech is clear and intelligible (at the behavioral level, i.e., when the participants respond what the target word of the sentence is), participants recognize the intelligible target word across all levels of sentence predictability. Hence, no differential effect of levels of predictability of target word can be expected.

These findings of Obleser and colleagues (Obleser and Kotz, 2011) were replicated and extended by Strauß et al. (2013; see Obleser, 2014). In a modified experimental design, they varied the target word predictability by manipulating its expectancy (i.e., how expected the target word is given the verb) and typicality (i.e., co-occurrence of target word and the preceding verb). They reported that at a moderate level of spectral degradation, N400 responses at strong-context, low-typical words and weak-context, low-typical words were largest. N400 responses at the latter two were not statistically different from each other. However, the N400 response was smallest at highly predictable (strong-context and high-typical) words. The authors interpreted these findings as a facilitatory effect of sentence predictability which might be limited to only highly predictable sentence endings at a moderate level of spectral degradation. In their "expectancy searchlight model," they suggested that listeners form "narrowed expectations" from a restricted semantic space when the sentence endings are highly predictable. When the sentence endings are less predictable, listeners cannot preactivate those less predictable sentence endings in an adverse listening condition. This is contrary to the view that readers and listeners form a probabilistic prediction of upcoming word in a sentence. For example, Nieuwland et al. (2018) showed in a large-scale replication study of DeLong et al. (2005) that the N400 amplitude at the sentence-final noun is directly proportional to its cloze probability across a range of high- and low-cloze words (see also, Kochari and Flecken, 2019; Nicenboim et al., 2020). Heilbron et al. (2020) also showed that a probabilistic prediction model outperforms a constrained guessing model, suggesting that linguistic prediction is not limited to highly predictable sentence endings, but it operates broadly in a wide range of probable sentence endings. However, a difference is that these studies were conducted in conditions without noise.

The probabilistic nature of prediction in comprehension of degraded speech has focused on a comparison of listeners' response to high-cloze target words and low-cloze target words (Obleser et al., 2007; Obleser and Kotz, 2011; see also, Strauß et al., 2013; Amichetti et al., 2018). In the present study, we included sentences with medium-cloze target words (see **Supplementary Material**). If the listeners form a narrowed prediction only for high-cloze target words, then the facilitatory effect of semantic prediction will be observed only at these highly predictable sentence endings. Listeners' response to medium-cloze target words and low-cloze target words would be expected to be quite similar as these two will fall out of the range of narrowed prediction. However, if the listeners' predictions are not restricted to highly predictable target words, then they form predictions across a wide range of semantic context proportional to the probability of occurrence of the target word. In addition to highly predictable sentence endings, listeners will also form predictions for less predictable sentence endings. Such predictions, however, will depend on the probability of occurrence of the target words. In other words, listeners form predictions also for less expected sentence endings; and the semantic space of prediction depends on the probability of occurrence of those sentence endings. The addition of sentences with medium-cloze target words in the present study thus allows us to differentiate whether listeners form all-or-none prediction restricted to high-cloze target words, or a probabilistic prediction for words across a wide range of cloze probability.

Adaptation to Degraded Speech

Listeners quickly adapt to novel speech with artificial acoustic distortions (e.g., Dupoux and Green, 1997). Repeated exposure to degraded speech leads to improved comprehension over time (for reviews, Samuel and Kraljic, 2009; Guediche et al., 2014). When the noise condition is constant throughout the experiment, listeners adapt to it and the performance (e.g., word recognition) improves with as little as 20 min of exposure (Rosen et al., 1999). For example, Davis et al. (2005, Experiment 1) presented listeners with sentences with 6 channels of noise-vocoding and found an increase in the proportion of correctly reported words over the course of experiment. Similarly, Erb et al. (2013) presented participants with sentences passed through 4 channels of noise-vocoding and reached a similar conclusion. In these experiments, only a single spectrally degraded speech signal (passed through 6 or 4 channels) was presented in one block. Therefore, it was predictable from the point of view of the participant which level of spectral degradation will appear in any trial within the block. Additionally, target word predictability was not varied.

When multiple types or levels of degraded speech signals are presented in a (pseudo-)randomized order within a block, then a listener is uncertain about any upcoming trials' signal quality; if such multiple levels of degradation are due to the presentation of multiple channels of noise-vocoded speech, then the global channel context is unpredictable or uncertain. This can influence perceptual adaptation. For instance, Mattys et al. (2012) note the possibility for a total absence of perceptual adaptation, when the characteristics of auditory signal change throughout an experiment. We also know from Sommers et al. (1994) that trial-by-trial variability in the characteristics of distorted speech impairs word recognition (see also, Dahan and Magnuson, 2006). We thus speculated that if the noise-vocoded speech varies from one trial to the next, then the adaptation to noise in this scenario might be different from the earlier case in which spectral degradation is constant throughout the experiment. Perceptual adaptation, however, is not limited to trial-by-trial variability of stimulus property. Listeners can adapt to auditory signal at different time courses and time scales (Atienza et al., 2002; see also, Whitmire and Stanley, 2016). In addition to differences in intrinsic trial-by-trial variability and resulting short timescale trial-by-trial adaptation in two channel contexts, the global differences in the presentation of vocoded speech can result in a difference in the general adaptation at a longer timescale between predictable and unpredictable channel contexts.

There is a limited number of studies that has looked at how next-trial noise-uncertainty and global context of speech property influence adaptation. For example, words were presented at +3 dB SNR and +10 dB SNR in a word-recognition task in a pseudorandom order (Vaden et al., 2013). The authors wanted to minimize the certainty about the noise conditions in the block. The same group of authors (Vaden et al., 2015, 2016; Eckert et al., 2016) proposed that an adaptive control system (cingulo-opercular circuit) might be involved to optimize task performance when listeners are uncertain about the upcoming trial. However, we cannot make a firm conclusion about perceptual adaptation *per se* from their studies as they do not report the change in performance over the course of experiment. Similarly, Obleser and colleagues (Obleser et al., 2007; Obleser and Kotz, 2011; Hartwigsen et al., 2015) also presented listeners with noise-vocoded sentences (ranging from 2 to 32 channels of noise-vocoding) in a pseudo-randomized order but did not report the presence or absence of perceptual adaptation to noise-vocoded speech. In the above-mentioned studies, the authors did not compare participants' task performance in blocked design against the presentation in a pseudorandomized block of different noise conditions to make an inference about general adaptation to degraded speech at a longer timescale. To examine the influence of uncertainty about next-trial speech features and the global context of speech features on perceptual adaptation, we will therefore compare language comprehension with a trial-by-trial variation of sentence predictability and speech degradation either in blocks in which the noise-vocoded channels are blocked, or in a randomized order.

Measurement of Language Comprehension

Another issue we would like to discuss is how to best measure language comprehension. The measurement of comprehension performance is inconsistent across studies. For instance, Erb et al. (2013) and Hakonen et al. (2017) measured participants' performance as proportion of correctly reported words per sentence ("report scores"; Peelle et al., 2013). On the other hand, Sheldon et al. (2008) asked participants to only report the final word of the sentence and then calculated the proportion of correctly reported words. One disadvantage of these approaches is that they do not consider whether participants correctly identified the sentence context or not. A crude word recognition score and the proportion of correct responses do not reflect an accurate picture of facilitation (or lack thereof) of language comprehension. Therefore, in the present study, we consider only those responses in which participants correctly identify the sentence context.

Goals of This Study

In sum, the goals of the study were threefold. Our first goal was to replicate and extend the behavioral results of Obleser and colleagues (Obleser et al., 2007; Obleser and Kotz, 2011), namely that the effect of semantic predictability will be observed only at a moderate level of speech degradation – as participants can realize the sentence context at the moderate level, their prediction will be narrowed down and the reliance on the bottom-up processing of the sentence final word will be over-ridden by top-down prediction. In contrast, when the sentence is clearly intelligible, even if the prediction is narrowed down and regardless of whether the clearly intelligible final word confirms or disconfirms those predictions, they respond based on what they hear, i.e., they rely mostly on acoustic-phonetic rather than lexical cues. Our study will provide new insights to the field by examining whether listeners indeed form *narrowed expectations* such that the facilitatory effect of predictability will be observed only for high-cloze target words and not at

medium- or low-cloze probability. To examine this, we created 360 German sentences at different levels of target word probability – low-cloze probability = 0.022 ± 0.027, medium-cloze probability = 0.1 ± 0.55, and high-cloze probability = 0.56 ± 1.0 – and varied the levels of spectral degradation by noise-vocoding through 1, 4, 6 and 8 channels.

Our second goal was to investigate the role of an uncertainty about the next-trial speech features on perceptual adaptation by varying the global channel context on the comprehension of degraded speech. To study this, we presented sentences of different levels of predictability blocked by each channel conditions (*predictable channel context*) and pseudo-randomized across all channels (*unpredictable channel context*). Based on previous findings, we expected that in the unpredictable channel context (i.e., when sentences are presented in a random order of spectral degradation) participants' word recognition performance will be worse than in the predictable channel context (i.e., when the sentences are blocked by noise-vocoding; Sommers et al., 1994; Garrido et al., 2011; Vaden et al., 2013). Moreover, to further examine perceptual adaptation, we also considered the effect of trial number in the analyses of data.

Our third goal was to establish a sensitive metric of measurement of language comprehension which considers the use of context by listeners. We note the caveat in the measurement of language comprehension in above-mentioned studies (e.g., Sheldon et al., 2008; Erb et al., 2013; Hakonen et al., 2017) and extend it further with a metric that we consider is a better measure of language comprehension (Amichetti et al., 2018) in the write-down paradigm (Samar and Metz, 1988).

MATERIALS AND METHODS

Participants

We recruited two groups of participants *via* Prolific Academic and assigned them to one of the two groups: "unpredictable channel context" (n = 48; $\bar{x} \pm SD = 24.44 \pm 3.5$ years; age range = 18–31 years; 16 females) and "predictable channel context" (n = 50; $\bar{x} \pm SD = 23.66 \pm 3.2$ years; age range = 18–30 years; 14 females). All participants were native speakers of German residing in Germany. Exclusion criteria for participating in this study were self-reported hearing disorder, speech-language disorder, or any neurological disorder. All participants received monetary compensation for their participation. The study was approved by the Deutsche Gesellschaft für Sprachwissenschaft (DGfS) Ethics Committee, and the participants provided consent in accordance with the Declaration of Helsinki.

Stimuli

The stimuli were digital recordings of 360 German sentences spoken by a female native speaker of German in a normal rate of speech. All sentences were in present tense consisting of pronoun, verb, determiner, and object (noun) in the Subject-Verb-Object form. We used 120 unique nouns to create three sentences that differed in cloze probability of target words. This resulted into sentences with low-, medium-, and high-cloze target word (for examples, see **Figure 1**).

We collected cloze probability ratings for each of these sentences in separate groups of younger participants (n = 60) of the same age range (18–30 years) prior to this study, while not all participants received the full set of 360 sentences. Mean cloze probabilities were 0.022 (SD = 0.027; range = 0.00–0.09) for sentences with low-cloze target word (low-predictability sentences), 0.274 (SD = 0.134; range = 0.1–0.55) for sentences with medium-cloze target word (medium-predictability sentences), and 0.752 (SD = 0.123; range = 0.56–1.00) for sentences with high-cloze target word (high-predictability sentences). The distribution of cloze probability across low-, medium-, and high-predictability sentences is shown in **Figure 1**.

The sentences were recorded and digitized at 44.1 kHz with 32-bit linear encoding. The spectral degradation conditions of 1, 4, 6, and 8 channels were achieved for each of the 360 recorded sentences using a customized noise-vocoding script originally written by Darwin (2005) in Praat. The speech signal was divided into 1, 4, 6, and 8 frequency bands between 70 and 9,000 Hz. The frequency boundaries were determined by cochlear-frequency position functions, and the boundary frequencies were approximately logarithmically spaced (Greenwood, 1990; Erb, 2014). The amplitude envelope of each band was extracted and applied to band-pass filtered white noise in the same frequency ranges; the upper and lower bounds for band extraction are specified in **Table 1**. Each of the modulated noises was then combined to produce distorted sentence. Scaling was performed to equate the root-mean-square value of the original undistorted sentence and the final distorted sentences. This resulted into four channel conditions: 1 channel, 4 channels, 6 channels, and 8 channels. The 1-channel noise-vocoding provides a baseline condition as speech encoded with only one frequency band is least to non-intelligible. However, speech vocoded through four or more channels has been shown to be well intelligible – Ueda and Nakajima (2017) derived from the factor analysis of spectral fluctuations in eight languages, including German, that four channels are sufficient, and they also identified the optimal boundary frequencies for 4 channels vocoding. These are similar, although not identical, to the cochlear-frequency position function-based boundary frequencies chosen in the current study.

In the *unpredictable channel* context, each participant was presented with 120 unique sentences: 40 high-predictability, 40 medium-predictability, and 40 low-predictability sentences. Channel condition was also balanced across each sentence type; i.e., in each of high-, medium-, and low-predictability sentences, 10 sentences passed through each noise-vocoding channels – 1, 4, 6, and 8 – were presented. This resulted into 12 experimental lists. The sentences in each list were pseudo-randomized, that is, not more than three sentences of same channel condition, or same predictability condition appeared consecutively. This randomization ascertained uncertainty of next-trial speech quality/degradation in the global context of the experiment.

The same set of stimuli and experimental lists were used in the predictable channel context. Each participant was presented with 120 unique sentences blocked by channel conditions. There

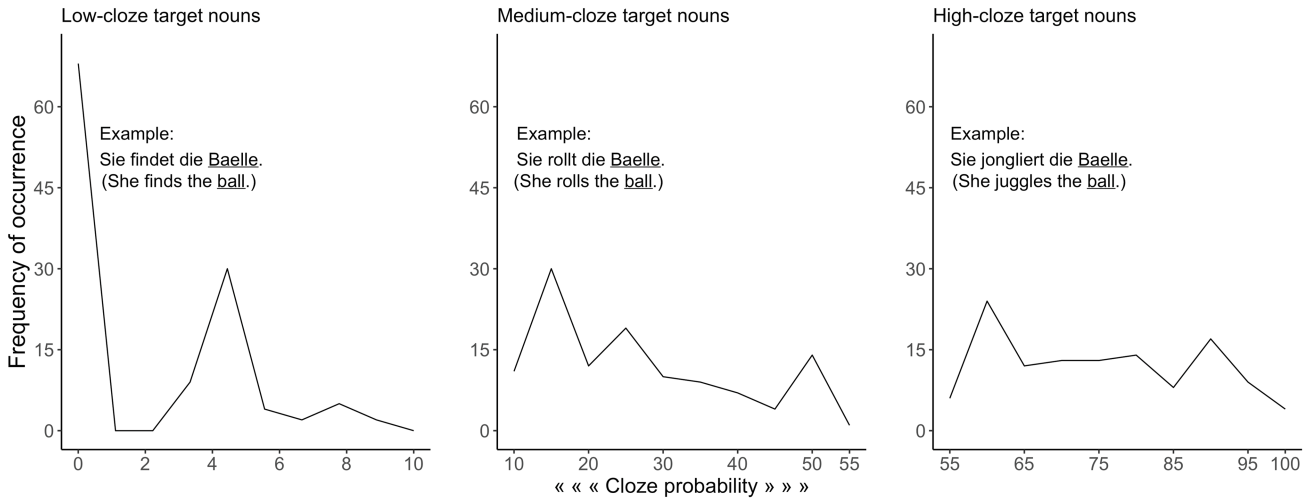

FIGURE 1 | The distribution of cloze probability values for low-, medium-, and high-cloze target nouns. Example sentences with their English translations are shown on each plot.

TABLE 1 | Boundary frequencies (in Hz) for 1, 4, 6, and 8 channels noise-vocoding conditions.

Number of channels	Boundary frequencies								
1	70	9,000							
4	70	423	1,304	3,504	9,000				
6	70	268	633	1,304	2,539	4,813	9,000		
8	70	207	423	764	1,304	2,156	3,504	5,634	9,000

were four blocks of stimuli. Thirty sentences were presented in each of the four blocks. In the first block, all sentences were of 8 channels, followed by blocks of 6 channels, 4 channels, and 1 channel speech, consecutively (Sheldon et al., 2008). Within each block, 10 high-predictability, 10 medium-predictability, and 10 low-predictability sentences were presented. All the sentences were pseudo-randomized so that not more than three sentences of the same predictability condition appeared consecutively in each block.

Procedure

Participants were asked to use headphones or earphones. A prompt to adjust loudness was displayed at the beginning of the experiment: A noise-vocoded sound not used in the main experiment was presented, and participants were asked to adjust the loudness at their level of comfort. One spoken sentence was presented in each trial. Eight practice trials were presented before presenting 120 experimental trials. They were asked to enter what they had heard (i.e., to type in the entire sentence) *via* keyboard. Guessing was encouraged. At the end of each trial, they were asked to judge their confidence in their response on a scale of 1 to 4, 1 being "just guessing" to 4 being "highly confident." The response was not timed. The experiment was about 40 min long.

Analyses

In the sentences used in our experiment, verbs evoke predictability of the sentence-final noun. Therefore, the effect of predictability (evoked by the verb) on language comprehension can be rightfully measured if we consider only those trials in which participants identify the verbs correctly. Verb-correct trials were considered as the sentence in which participants correctly understood the context (independent of whether they correctly understood the final target noun). Morphological inflections and typos were considered as correct. We first filtered out those trials in which context was not identified correctly, i.e., trials with incorrect verbs.[1] Therefore, we excluded 2,469 out of 5,760 trials in unpredictable channel context and 2,374 out of 6,000 trials in predictable channel context from the analyses. The condition with 1-channel noise-vocoding was dropped from the analyses as there were no correct responses in any of the trials in this condition. The number of trials excluded per condition in each group is shown in **Table 2**.

Data preprocessing and analyses were performed in R-Studio (Version 3.6.1; R Core Team, 2019). Accuracy was analyzed

[1]We also performed a complementary analysis with only noun-correct trials, which was suggested by a reviewer, to test whether there was a backward effect of guessing the verb after first recognizing the noun in a sentence. The results of the analysis are presented in the Supplementary Material.

TABLE 2 | Number of trials excluded per condition.

Channel context	Channel condition	Predictability			Total trials presented
		Low	Medium	High	
Predictable (blocked design)	4 channels	208	181	215	1,500
	6 channels	62	60	49	1,500
	8 channels	32	29	38	1,500
Unpredictable (randomized design)	4 channels	251	236	241	1,440
	6 channels	61	63	55	1,440
	8 channels	49	31	42	1,440

TABLE 3 | Estimated effects of the best fitting generalized (binomial logistic) mixed-effects model accounting for the correct word recognition.

Fixed effects	Estimate	Standard error	Value of *z*	Value of *p*
Intercept	5.09	0.24	21.38	<0.001
Channel condition (4 channels)	−2.87	0.22	−13.10	<0.001
Channel condition (6 channels)	−0.66	0.19	−3.42	0.001
Target word predictability (Low-Medium)	−0.52	0.27	−1.97	0.049
Target word predictability (High-Low)	2.18	0.30	7.21	<0.001
Channel condition × Target word predictability	−0.71	0.29	−2.44	0.015
Global channel context (Unpredictable - Predictable)	−0.27	0.14	−2.02	0.043

Optimal model:
glmer(response ~ 1 + 4ch + 6ch + Low-Medium + High-Low + 4ch: Low-Medium + ChannelContext +
(1 + 4ch + High-Low | subject) +
(1 + 4ch + 6ch + Low-Medium + High-Low + 4ch: Low-Medium || item).
NB: The minus sign is only a symbolic representation of the difference, between two factors, from the sliding difference contrasts; see **Supplementary Material** *for the details about the model description.*
"ch" is an abbreviation for "channels."

with Generalized Linear Mixed Models with lmerTest (Kuznetsova et al., 2017) and lme4 (Bates et al., 2015b) packages which operate on log-odds scale. Binary responses (correct responses coded as 1 and incorrect responses coded as 0) for all participants in both groups (predictable global noise context and unpredictable global noise context) were fit with a binomial mixed-effects model; i.e., response accuracy was a categorical-dependent variable in the model (Jaeger, 2006, 2008). Channel condition (categorical; 4 channels, 6 channels, and 8 channels), target word predictability (categorical; high, medium, and low predictability), global channel context (categorical; predictable channel context and unpredictable channel context), and the interaction of channel condition and target word predictability, and the main effect of global channel context were included in the fixed effects.

We first fitted a model with maximal random effects structure that included random intercepts for each participant and item (Barr et al., 2013). Both by-participant and by-item random slopes were included for channel condition, target word predictability, and their interaction. To find the optimal model for the data, non-significant higher-order interactions were excluded from the fixed-effects structure (and from the random-effects structure) in a stepwise manner. Model selection was based on Akaike Information Criterion (Grueber et al., 2011; Richards et al., 2011) unless otherwise stated. Random effects not supported by the data that explained zero variance according to singular value decomposition were excluded to prevent overparameterization. This gave a more parsimonious model (Bates et al., 2015a). Such a model was then extended separately with: (i) item-related correlation parameters, (ii) participant-related correlation parameter, and (iii) both item- and participant-related correlation parameters. The best fitting model among the parsimonious and extended models was then selected as the optimal model for our data. Note that the parsimonious model shows qualitatively the same effects as the maximal model.

We applied treatment contrast for channel condition (8 channels as the baseline; factor levels: 8 channels, 4 channels, and 6 channels) and sliding difference contrast for target word predictability (factor levels: medium predictability, low predictability, and high predictability) and global channel context (factor levels: unpredictable and predictable). We report the results from the optimal model (see **Table 3**).

RESULTS

We primarily wanted to test (i) whether predictability facilitates language comprehension only at a moderate level of spectral degradation and (ii) whether adaptation to degraded speech influences language comprehension. We observed that the mean response accuracy increased with an increase in number of noise-vocoding channels from 4 to 6 to 8, and with an increase in target word predictability from low to medium to high, as can be seen in **Figure 2**. This trend is consistent across both predictable channel context (blocked design) and unpredictable channel context (randomized design; see also **Tables 4** and **5**).

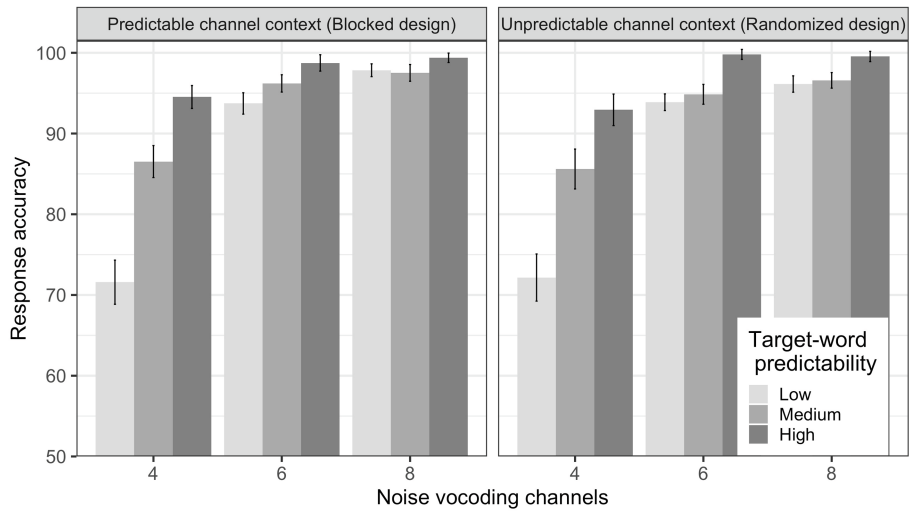

FIGURE 2 | Mean response accuracy for the blocked design (left) and the randomized design (right). Error bars represent standard error of the means.

TABLE 4 | Mean response accuracy in predictable channel context.

Noise-vocoding	Predictability	Accuracy	SE
4 channels	Low	71.59	2.74
	Medium	86.53	1.99
	High	94.53	1.42
6 channels	Low	93.73	1.33
	Medium	96.21	1.08
	High	98.75	1.02
8 channels	Low	97.84	0.8
	Medium	97.52	1.04
	High	99.38	0.59

TABLE 5 | Mean response accuracy in unpredictable channel context.

Noise-vocoding	Predictability	Accuracy	SE
4 channels	Low	72.16	2.93
	Medium	85.61	2.47
	High	92.94	1.96
6 channels	Low	93.88	1.04
	Medium	94.86	1.24
	High	99.81	0.62
8 channels	Low	96.14	1.02
	Medium	96.59	0.97
	High	99.55	0.64

The results of statistical analysis confirmed these observations. It showed that there was a main effect of channel condition indicating that the response accuracy was higher in the 8 channels than in the 4 channels [$\beta = -2.87$, SE = 0.22, z (6917) = −13.10, $p < 0.001$] and 6 channels [$\beta = -0.66$, SE = 0.19, z (6917) = −3.42, $p < 0.001$]. There was a main effect of target word predictability, suggesting that response accuracy was lower at low predictability than both high-predictability [$\beta = 2.18$, SE = 0.30, z (6917) = 7.21, $p < 0.001$] and medium-predictability sentences [$\beta = -0.52$, SE = 0.27, z (6917) = −1.97, $p = 0.049$]. There was also an interaction between channel condition and target word predictability [$\beta = -0.71$, SE = 0.29, z (6917) = −2.44, $p = 0.015$]. See **Figure 2** for the corresponding mean accuracies.

Subsequent subgroup analyses were performed following the same procedure as described above. The results are shown in **Table 6**. They revealed that the interaction was driven by the effect of predictability at 4 channels: The accuracy at high-predictability sentences was higher than medium-predictability sentences [$\beta = 1.14$, SE = 0.37, z (1608) = 3.10, $p = 0.002$], which in turn was also higher than low-predictability sentences [$\beta = 1.01$, SE = 0.24, z (1608) = 4.20, $p < 0.001$]. There was no significant difference in response accuracy between low- and medium-predictability sentences at both 6 [$\beta = 0.33$, SE = 0.32, z (2590) = 1.04, $p = 0.3$] and 8 channels [$\beta = -0.01$, SE = 0.32, z (2719) = −0.04, $p = 0.965$]. However, response accuracy was higher in high-predictability than in medium-predictability sentences at both 6 channels [$\beta = 1.83$, SE = 0.65, z (2590) = 2.83, $p < 0.005$] and 8 channels [$\beta = 1.54$, SE = 0.61, z (2719) = 2.54, $p = 0.011$].

We also found a main effect of global channel context showing that response accuracy was higher in predictable than in unpredictable channel context [$\beta = -0.27$, SE = 0.14, z (6917) = −2.02, $p = 0.043$].

To further test the effect of practice in the adaptation to noise, we added trial number as a fixed effect in the maximal model. Note that there were 30 trials in each block in the blocked design (predictable channel context). For comparability, we divided randomized design (unpredictable channel context) into four blocks; there were 30 consecutive trials in each block. Then, following the same procedure as above, an optimal model was obtained. The results showed that response accuracy did not change throughout the experiment [$\beta = -0.0001$, SE = 0.01, z (6917) = −0.02, $p = 0.985$]. It remained constant within each block in predictable channel context [$\beta = -0.02$, SE = 0.01, z (3626) = −1.43, $p = 0.152$] as well as in unpredictable channel context [$\beta = 0.01$, SE = 0.01, z (3291) = 1.05, $p = 0.292$].

TABLE 6 | Estimated effects of the generalized (binomial logistic) mixed-effects model accounting for the correct word recognition at 4 channels condition.

Fixed effects	Estimate	Standard error	Value of *z*	Value of *p*
Intercept	2.17	0.20	10.95	<0.001
Target word predictability (Medium-Low)	1.01	0.24	4.20	<0.001
Target word predictability (High-Medium)	1.14	0.37	3.10	0.002
Global channel context (Unpredictable - Predictable)	−0.27	0.18	−1.53	0.127

Optimal model:
glmer(response ~ 1 + 4ch + 6ch + Low-Medium + High-Low + 4ch: Low-Medium + ChannelContext +
(1 + 4ch + High-Low | subject) +
(1 + 4ch + 6ch + Low-Medium + High-Low + 4ch: Low-Medium || item)
NB: The minus sign is only a symbolic representation of the difference, between two factors, from the sliding difference contrasts; see **Supplementary Material** *for the details about the model description.*
"ch" is an abbreviation for "channels."

DISCUSSION

The present study had three goals: (i) to examine whether previously reported facilitatory effect of semantic predictability is restricted to only highly predictable sentence endings; (ii) to assess the role of perceptual adaptation on the facilitation of language comprehension by sentence predictability; and (iii) to use and establish a sensitive metric to measure language comprehension that takes into account whether listeners benefited from the semantic context of the sentence they have listened to.

Results of our study showed the expected interaction between predictability and degraded speech; that is, language comprehension was better for high-cloze than for low-cloze target words when the speech signal was moderately degraded by noise-vocoding through 4 channels, while the effect of predictability was absent when speech was not intelligible (noise-vocoding through 1 channel). These results are fully in line with Obleser and Kotz (2010); we partly included identical sentences from their study in the present study (see **Supplementary Material**). Importantly, in contrast to their study, we had also created sentences with medium-cloze target words (which were intermediate between high-cloze and low-cloze target words) and found that the effect of predictability was also significant when comparing sentences with medium-cloze target words against sentences with low-cloze target words at 4 channels noise-vocoding condition. Recognition of a target word was dependent on its level of predictability (measured by cloze probability), and correct recognition was not just limited to high-cloze target words. These significant differences in response accuracy between medium-cloze and low-cloze target words, and between medium-cloze and high-cloze target words at noise-vocoding through 4 channels show that the sentence-final word recognition is facilitated by semantic predictability in a *graded manner*. This is in line with the findings from the ERP literature where it has been observed that semantic predictability, in terms of cloze probability of target word of a sentence, modulates semantic processing, indexed by N400, in a graded manner (DeLong et al., 2005; Wlotko and Federmeier, 2012; Nieuwland et al., 2018).

The interpretation of the observed graded effect of semantic predictability at the moderate level of spectral degradation (i.e., at noise-vocoding through 4 channels) provides a novel insight into how listeners form prediction when the bottom-up input is compromised. That is, in an adverse listening condition, listeners rely more on top-down semantic prediction than on bottom-up acoustic-phonetic cues. However, such a reliance on top-down prediction is not an all-or-none phenomenon; instead, listeners form a probabilistic prediction of the target word. The effect of target word predictability on comprehension is not sharply focused solely on high-cloze target words like a "searchlight." But rather it is spread across a wide range, including low-cloze and medium-cloze target words. As the cloze probability of the target words decreases from high to low, the focus of the searchlight becomes less precise.

Obleser et al. (2007) and Strauß et al. (2013) reported an effect of predictability on language comprehension at noise-vocoding through 8 channels. On the other hand, we and Obleser and Kotz (2010) find a similar effect in 4 channels. This can be explained by the relative complexity of the stimuli used in this latter study. Obleser et al. (2007) took the sentences from G-SPIN (German version of Speech in Noise) test. Those sentences are longer and do not have uniform form and structure, while the sentences in Obleser and Kotz (2010) and the current study are shorter and have a uniform form and structure (Subject-Verb-Object). Owing to this fact, noise-vocoding through 4 channels was not intelligible enough for the G-SPIN test sentences, and the effect of predictability could be observed only at a higher number of vocoding channels (8 channels). This difference in number of channels (4 vs. 8) required for the effect of predictability to emerge is, therefore, due to the difference in stimuli complexity; the moderate-level spectral degradation could either be 8 channels or 4 channels depending on stimuli complexity. In the present study, moderate level of spectral degradation could be observed at noise-vocoding through 4 channels.

Previously reported facilitatory effect of semantic predictability comes from studies conducted in laboratory setups. The current experiment was conducted online. There is a possibility that different participants used different hearing devices. Such variability in hearing devices could not be controlled for in our experiment although the participants were restricted to using only desktop/laptop computers. However, we have no reason to believe that these variance sources are systematically correlated with our between-group manipulation (global channel context) and the effects are constant within subjects. Moreover, the main finding of this study, i.e., the graded effect of semantic predictability, is observed in both the groups.

We highlight the importance of considering participants' context use in experiments that attempt to answer the questions pertaining to the use of top-down predictive cues. Unless it is established that participants correctly understood the context, the findings are likely to be confounded by the trials in which the predictability condition is not controlled as participants did not understand the context correctly, and the target word is not predictable based on the misunderstood context. In those cases, correctly recognizing the target word does not necessarily mean that listeners made use of the context and it was top-down prediction that facilitated the comprehension. Similarly, in cases where the target word was wrong, and the context was also not understood, it does not mean that participants did not form prediction based on what they understood. Also, instructions can affect how participants direct their attention during the task – they might shift attention strategically only to the target word, if this is all that is required for the task; hence, task instructions can also be a confounding factor (Sanders and Astheimer, 2008; Astheimer and Sanders, 2009; Li et al., 2014).

An alternative explanation of our findings could be that the listeners "guessed" the verb after first correctly identifying the noun in a sentence, which a reviewer pointed out. We therefore conducted an additional analysis where we compared forward predictability effects (from verb to noun) to the size of backward predictability effects (correct identification of the noun based on the final verb). If the observed effect is simply a cloze guessing effect, then we would expect that both forward and backward predictability effects are similar in size. If, on the other hand, understanding the verb really helps to shape predictions of the upcoming noun, and this helps intelligibility, then the forward prediction effect should be larger. The results of this complementary analysis (see the **Supplementary Material**) support the findings of the main analysis reported in the Results. In the backward predictability analysis, there was no graded effect of predictability, and the backward effect of "guessing" the verb *jongliert* after recognizing the noun *Baelle*, if present at all, was smaller than the forward effect of predicting the noun after recognizing the verb in the sentence *Sie jongliert die Baelle*. This corroborates our argument that listeners in fact made use of the verb-evoked context to form predictions about upcoming noun, and not the other way around, in a graded manner when the speech was moderately degraded.

The results of the analyses of trial number on the effect of channel context to capture trial-by-trial perceptual adaptation showed that the response accuracy did not increase over the course of experiment. This suggests that listeners' performance remained constant over the course of experiment regardless of the predictability of next-trial spectral degradation. Perceptual adaptation occurs when the perceptual system of a listener retunes itself to the sensory properties of the auditory signal which can be facilitated by higher-level lexical information or feedback (Goldstone, 1998; Mattys et al., 2012). We speculate that the trial-by-trial variability in the spectral resolution of the speech signal in the unpredictable channel context prevented perceptual adaptation. Although there was certainty about the quality of speech signal within a block in the predictable channel context, we did not observe trial-by-trial perceptual adaptation in this condition either. This is contrary to previous studies showing that listeners adapt to degraded speech when the global context of speech quality is predictable (e.g., Davis et al., 2005; Erb et al., 2013). However, the crucial difference between those studies and our study is the manipulation of target word predictability. For example, Erb et al. (2013) presented sentences with only low-predictability target words from the G-SPIN test. We, on the other hand, parametrically varied target word predictability from low to medium and high. Note that we presented target words in a randomized order in both channel contexts. This alone introduces trial-by-trial uncertainty in the predictable channel context and possibly hinders trial-by-trial perceptual adaptation. As Goldstone (1998, p. 588) notes – "one way in which perception becomes adapted to tasks and environments is by increasing the attention paid to perceptual dimensions and features that are important, and/or by decreasing attention to irrelevant [perceptual] dimensions and features" (see also, Gold and Watanabe, 2010). In our study, listeners probably paid more attention to semantic properties of the sentences (i.e., contextual cues and target word predictability) than to the perceptual properties (i.e., spectral resolution or speech quality) as we had instructed. We speculate this might have resulted in the absence of trial-by-trial perceptual adaptation to degraded speech, even when next-trial channel condition was predictable. However, one noteworthy finding is the higher accuracy in the predictable channel context than in the unpredictable channel context. We interpret the differences in task performance in these two channel contexts in terms of general adaptation at a longer timescale. Adaptation to trial-by-trial variability of stimuli property reflects adaptation at a shorter timescale; at a longer timescale, however, listeners adapt to the experimental condition in which stimuli properties change slowly (Atienza et al., 2002; for neurobiological account, see Whitmire and Stanley, 2016; Zhang and Chacron, 2016; Weber et al., 2019). In both predictable and unpredictable channel contexts, adaptation in the short timescale was hindered by trial-by-trial variation of either one (target word predictability) or both properties (target word predictability and channel condition) of the speech stimuli. However, listeners adapted to the vocoded speech in the longer timescale when there was a certainty of channel condition (in predictable channel context) at the level of global channel context.

CONCLUSION

In conclusion, this study provides novel insights into predictive language processing when bottom-up signal quality is compromised and uncertain: We show that while processing moderately degraded speech, listeners form top-down predictions across a wide range of semantic space that is not restricted within highly predictable sentence endings. In contrast to the narrowed expectation view, comprehension of words ranging from low- to high-cloze probability, including medium-cloze probability, is facilitated in a graded manner while listening to a moderately degraded speech. We also found better speech comprehension when individuals were likely to have adapted to the noise condition in the blocked design compared to the randomized design.

We did not find learning effects at the trial-to-trial level of perceptual adaption – it may be that the adaptation was hampered by variation in higher-level semantic features (i.e., target word predictability). We also argue that for the examination of semantic predictability effects during language comprehension, the analyses of response accuracy should be based on the trials in which context evoking words are correctly identified in the first place to make sure that listeners make use of the contextual cues instead of analyzing general word recognition scores.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Deutsche Gesellschaft für Sprachwissenschaft (DGfS) Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

PB, VD, and JK were involved in planning and designing of the study. PB analyzed the data and wrote all parts of the manuscript. VD and JK made the suggestions to the framing, structuring, and presentation of findings as well as on the interpretation of findings. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

ACKNOWLEDGMENTS

We would like to thank Melanie Hiltz, Sarah Memeche, Katharina Stein, and Amélie Leibrock for their help in constructing the stimuli and assistance with the phonetic transcriptions, and Katja Häuser, Yoana Vergilova, and Marjolein van Os for helpful discussion. We also thank Jonas Obleser for sharing the stimuli used in their study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.714485/full#supplementary-material

REFERENCES

Altmann, G. T. M., and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* 73, 247–264. doi: 10.1016/S0010-0277(99)00059-1

Altmann, G. T. M., and Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: linking anticipatory (and other) eye movements to linguistic processing. *J. Mem. Lang.* 57, 502–518. doi: 10.1016/j.jml.2006.12.004

Amichetti, N. M., Atagi, E., Kong, Y. Y., and Wingfield, A. (2018). Linguistic context versus semantic competition in word recognition by younger and older adults with cochlear implants. *Ear Hear.* 39, 101–109. doi: 10.1097/AUD.0000000000000469

Astheimer, L. B., and Sanders, L. D. (2009). Listeners modulate temporally selective attention during natural speech processing. *Biol. Psychol.* 80, 23–34. doi: 10.1016/j.biopsycho.2008.01.015

Atienza, M., Cantero, J. L., and Dominguez-Marin, E. (2002). The time course of neural changes underlying auditory perceptual learning. *Learn. Mem.* 9, 138–150. doi: 10.1101/lm.46502

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015a). Parsimonious mixed models. [Preprint]. Available at: http://arxiv.org/abs/1506.04967 (Accessed September 22, 2019).

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

Brouwer, H., and Crocker, M. W. (2017). On the proper treatment of the N400 and P600 in language comprehension. *Front. Psychol.* 8:1327. doi: 10.3389/fpsyg.2017.01327

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204. doi: 10.1017/S0140525X12000477

Corps, R. E., and Rabagliati, H. (2020). How top-down processing enhances comprehension of noise-vocoded speech: predictions about meaning are more important than predictions about form. *J. Mem. Lang.* 113, 104–114. doi: 10.1016/j.jml.2020.104114

Dahan, D., and Magnuson, J. S. (2006). "Spoken word recognition," in *Handbook of psycholinguistics. Vol. 2.* eds. M. J. Traxler and M. A. Gernsbacher (Amsterdam: Academic Press), 249–283.

Darwin, C. (2005). Praat scripts for producing Shannon AM speech [Computer software]. Available at: http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts/ (Accessed March 08, 2021).

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *J. Exp. Psychol. Gen.* 134, 222–241. doi: 10.1037/0096-3445.134.2.222

DeLong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* 8, 1117–1121. doi: 10.1038/nn1504

Dupoux, E., and Green, K. (1997). Perceptual adjustment to highly compressed speech: effects of talker and rate changes. *J. Exp. Psychol. Hum. Percept. Perform.* 23, 914–927. doi: 10.1037//0096-1523.23.3.914

Eckert, M. A., Teubner-Rhodes, S., and Vaden, K. I. (2016). Is listening in noise worth it? The neurobiology of speech recognition in challenging listening conditions. *Ear Hear.* 37, 101S–110S. doi: 10.1097/AUD.0000000000000300

Ehrlich, S. F., and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *J. Verbal Learn. Verbal Behav.* 20, 641–655. doi: 10.1016/S0022-5371(81)90220-6

Erb, J. (2014). The neural dynamics of perceptual adaptation to degraded speech. doctoral dissertation. Leipzig: Max Planck Institute for Human Cognitive and Brain Sciences.

Erb, J., Henry, M. J., Eisner, F., and Obleser, J. (2013). The brain dynamics of rapid perceptual adaptation to adverse listening conditions. *J. Neurosci.* 33, 10688–10697. doi: 10.1523/JNEUROSCI.4596-12.2013

Federmeier, K. D., Wlotko, E. W., de Ochoa-Dewald, E., and Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Res.* 1146, 75–84. doi: 10.1016/j.brainres.2006.06.101

Fontan, L., Tardieu, J., Gaillard, P., Woisard, V., and Ruiz, R. (2015). Relationship between speech intelligibility and speech comprehension in babble noise. *J. Speech Lang. Hear. Res.* 58, 977–986. doi: 10.1044/2015_JSLHR-H-13-0335

Frisson, S., Rayner, K., and Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 862–877. doi: 10.1037/0278-7393.31.5.862

Garrido, M. I., Dolan, R. J., and Sahani, M. (2011). Surprise leads to noisier perceptual decisions. *Iperception* 2, 112–120. doi: 10.1068/i0411

Gold, J. I., and Watanabe, T. (2010). Perceptual learning. *Curr. Biol.* 20, R46–R48. doi: 10.1016/j.cub.2009.10.066

Goldstone, R. L. (1998). Perceptual learning. *Annu. Rev. Psychol.* 49, 585–612. doi: 10.1146/annurev.psych.49.1.585

Greenwood, D. D. (1990). A cochlear frequency-position function for several species – 29 years later. *J. Acoust. Soc. Am.* 87, 2592–2605. doi: 10.1121/1.399052

Grueber, C. E., Nakagawa, S., Laws, R. J., and Jamieson, I. G. (2011). Multimodel inference in ecology and evolution: challenges and solutions. *J. Evol. Biol.* 24, 699–711. doi: 10.1111/j.1420-9101.2010.02210.x

Guediche, S., Blumstein, S. E., Fiez, J. A., and Holt, L. L. (2014). Speech perception under adverse conditions: insights from behavioral, computational, and neuroscience research. *Front. Syst. Neurosci.* 7:126. doi: 10.3389/fnsys.2013.00126

Hakonen, M., May, P. J. C., Jääskeläinen, I. P., Jokinen, E., Sams, M., and Tiitinen, H. (2017). Predictive processing increases intelligibility of acoustically distorted speech: Behavioral and neural correlates. *Brain Behav.* 7:e00789. doi: 10.1002/brb3.789

Hartwigsen, G., Golombek, T., and Obleser, J. (2015). Repetitive transcranial magnetic stimulation over left angular gyrus modulates the predictability gain in degraded speech comprehension. *Cortex* 68, 100–110. doi: 10.1016/j.cortex.2014.08.027

Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., and de Lange, F. P. (2020). A hierarchy of linguistic predictions during natural language comprehension. [Preprint]. doi:10.1101/2020.12.03.410399.

Huettig, F., and Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Lang. Cognit. Neurosci.* 31, 19–31. doi: 10.1080/23273798.2015.1072223

Jachmann, T. K., Drenhaus, H., Staudte, M., and Crocker, M. W. (2019). Influence of speakers' gaze on situated language comprehension: evidence from Event-Related Potentials. *Brain Cogn.* 135:103571. doi: 10.1016/j.bandc.2019.05.009

Jaeger, T. F. (2006). Redundancy and syntactic reduction in spontaneous speech. doctoral dissertation. Stanford University.

Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434–446. doi: 10.1016/j.jml.2007.11.007

Kaiser, E., and Trueswell, J. C. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition* 94, 113–147. doi: 10.1016/j.cognition.2004.01.002

Kamide, Y., Altmann, G. T. M., and Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: evidence from anticipatory eye movements. *J. Mem. Lang.* 49, 133–156. doi: 10.1016/S0749-596X(03)00023-8

Knoeferle, P., Crocker, M. W., Scheepers, C., and Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition* 95, 95–127. doi: 10.1016/j.cognition.2004.03.002

Kochari, A. R., and Flecken, M. (2019). Lexical prediction in language comprehension: a replication study of grammatical gender effects in Dutch. *Lang. Cognit. Neurosci.* 34, 239–253. doi: 10.1080/23273798.2018.1524500

Kuperberg, G. R., and Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* 31, 32–59. doi: 10.1080/23273798.2015.1102299

Kutas, M., and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.* 62, 621–647. doi: 10.1146/annurev.psych.093008.131123

Kutas, M., and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307, 161–163. doi: 10.1038/307161a0

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed-effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13

Li, X., Lu, Y., and Zhao, H. (2014). How and when predictability interacts with accentuation in temporally selective attention during speech comprehension. *Neuropsychologia* 64, 71–84. doi: 10.1016/j.neuropsychologia.2014.09.020

Loizou, P. C., Dorman, M., and Tu, Z. (1999). On the number of channels needed to understand speech. *J. Acoust. Soc. Am.* 106, 2097–2103. doi: 10.1121/1.427954

Lupyan, G., and Clark, A. (2015). Words and the world: predictive coding and the language-perception-cognition interface. *Curr. Dir. Psychol. Sci.* 24, 279–284. doi: 10.1177/0963721415570732

Marques, T., Nguyen, J., Fioreze, G., and Petreanu, L. (2018). The functional organization of cortical feedback inputs to primary visual cortex. *Nat. Neurosci.* 21, 757–764. doi: 10.1038/s41593-018-0135-z

Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (2012). Speech recognition in adverse conditions: a review. *Lang. Cognit. Processes* 27, 953–978. doi: 10.1080/01690965.2012.705006

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., and Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *J. Mem. Lang.* 66, 545–567. doi: 10.1016/j.jml.2012.01.001

Nicenboim, B., Vasishth, S., and Rösler, F. (2020). Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia* 142:107427. doi: 10.1016/j.neuropsychologia.2020.107427

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., et al. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife* 7:e33468. doi: 10.7554/eLife.33468

Obleser, J. (2014). Putting the listening brain in context. *Lang. Ling. Compass* 8, 646–658. doi: 10.1111/lnc3.12098

Obleser, J., and Kotz, S. A. (2010). Expectancy constraints in degraded speech modulate the language comprehension network. *Cereb. Cortex* 20, 633–640. doi: 10.1093/cercor/bhp128

Obleser, J., and Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *NeuroImage* 55, 713–723. doi: 10.1016/j.neuroimage.2010.12.020

Obleser, J., Wise, R. J. S., Alex Dresner, M., and Scott, S. K. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. *J. Neurosci.* 27, 2283–2289. doi: 10.1523/JNEUROSCI.4663-06.2007

Peelle, J. E., Gross, J., and Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb. Cortex* 23, 1378–1387. doi: 10.1093/cercor/bhs118

Pickering, M., and Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychol. Bull.* 144, 1002–1044. doi: 10.1037/bul0000158

R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. Available at: https://www.R-project.org

Richards, S. A., Whittingham, M. J., and Stephens, P. A. (2011). Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework. *Behav. Ecol. Sociobiol.* 65, 77–89. doi: 10.1007/s00265-010-1035-8

Rosen, S., Faulkner, A., and Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: implications for cochlear implants. *J. Acoust. Soc. Am.* 106, 3629–3636. doi: 10.1121/1.428215

Samar, V. J., and Metz, D. E. (1988). Criterion validity of speech intelligibility rating-scale procedures for the hearing-impaired population. *J. Speech Hear. Res.* 31, 307–316. doi: 10.1044/jshr.3103.307

Samuel, A. G., and Kraljic, T. (2009). Perceptual learning for speech. *Atten. Percept. Psychophys.* 71, 1207–1218. doi: 10.3758/APP.71.6.1207

Sanders, L. D., and Astheimer, L. B. (2008). Temporally selective attention modulates early perceptual processing: event-related potential evidence. *Percept. Psychophys.* 70, 732–742. doi: 10.3758/PP.70.4.732

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends Cogn. Sci.* 17, 565–573. doi: 10.1016/j.tics.2013.09.007

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Shannon, R., Fu, Q.-J., and Galvin Iii, J. (2004). The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. *Acta Otolaryngol. Suppl.* 124, 50–54. doi: 10.1080/03655230410017562

Shannon, R. v., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science* 270, 303–304. doi: 10.1126/science.270.5234.303

Shannon, R. v., Zeng, F.-G., and Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *J. Acoust. Soc. Am.* 104, 2467–2476. doi: 10.1121/1.423774

Sheldon, S., Pichora-Fuller, M. K., and Schneider, B. A. (2008). Priming and sentence context support listening to noise-vocoded speech by younger and older adults. *J. Acoust. Soc. Am.* 123, 489–499. doi: 10.1121/1.2783762

Sommers, M. S., Nygaard, L. C., and Pisoni, D. B. (1994). Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *J. Acoust. Soc. Am.* 96, 1314–1324. doi: 10.1121/1.411453

Stadler, W., Ott, D. V. M., Springer, A., Schubotz, R. I., Schütz-Bosbach, S., and Prinz, W. (2012). Repetitive TMS suggests a role of the human dorsal premotor cortex in action prediction. *Front. Hum. Neurosci.* 6:20. doi: 10.3389/fnhum.2012.00020

Staub, A. (2011). The effect of lexical predictability on distributions of eye fixation durations. *Psychon. Bull. Rev.* 18, 371–376. doi: 10.3758/s13423-010-0046-9

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: critical review and theoretical interpretation. *Lang. Ling. Compass* 9, 311–327. doi: 10.1111/lnc3.12151

Stilp, C. (2020). Acoustic context effects in speech perception. *Wiley Interdiscip. Rev. Cogn. Sci.* 11:e1517. doi: 10.1002/wcs.1517

Strauß, A., Kotz, S. A., and Obleser, J. (2013). Narrowed expectancies under degraded speech: revisiting the N400. *J. Cogn. Neurosci.* 25, 1383–1395. doi: 10.1162/jocn_a_00389

Ueda, K., Araki, T., and Nakajima, Y. (2018). Frequency specificity of amplitude envelope patterns in noise-vocoded speech. *Hear. Res.* 367, 169–181. doi: 10.1016/j.heares.2018.06.005

Ueda, K., and Nakajima, Y. (2017). An acoustic key to eight languages/dialects: Factor analyses of critical-band-filtered speech. *Sci. Rep.* 7:42468. doi: 10.1038/srep42468

Vaden, K. I. Jr., Kuchinsky, S. E., Ahlstrom, J. B., Teubner-Rhodes, S., Dunbo, J. R., and Eckert, M. A. (2016). Cingulo-opercular function during word recognition in noise for older adults with hearing loss. *Exp. Aging Res.* 42, 67–82. doi: 10.1080/0361073X.2016.1108784

Vaden, K. I., Kuchinsky, S. E., Ahlstrom, J. B., Dubno, J. R., and Eckert, M. A. (2015). Cortical activity predicts which older adults recognize speech in noise and when. *J. Neurosci.* 35, 3929–3937. doi: 10.1523/JNEUROSCI.2908-14.2015

Vaden, K. I., Kuchinsky, S. E., Cute, S. L., Ahlstrom, J. B., Dubno, J. R., and Eckert, M. A. (2013). The cingulo-opercular network provides word-recognition benefit. *J. Neurosci.* 33, 18979–18986. doi: 10.1523/JNEUROSCI.1417-13.2013

Weber, A. I., Krishnamurthy, K., and Fairhall, A. L. (2019). Coding principles in adaptation. *Annu. Rev. Vision Sci.* 5, 427–449. doi: 10.1146/annurev-vision-091718-014818

Whitmire, C. J., and Stanley, G. B. (2016). Rapid sensory adaptation redux: a circuit perspective. *Neuron* 92, 298–315. doi: 10.1016/j.neuron.2016.09.046

Wingfield, A., McCoy, S. L., Peelle, J. E., Tun, P. A., and Cox, L. C. (2006). Effects and adult aging and hearing loss on comprehension of rapid speech varying in syntactic complexity. *J. Am. Acad. Audiol.* 17, 487–497. doi: 10.3766/jaaa.17.7.4

Wlotko, E. W., and Federmeier, K. D. (2012). Age-related changes in the impact of contextual strength on multiple aspects of sentence comprehension. *Psychophysiology* 49, 770–785. doi: 10.1111/j.1469-8986.2012.01366.x

Xia, J., Xu, B., Pentony, S., Xu, J., and Swaminathan, J. (2018). Effects of reverberation and noise on speech intelligibility in normal-hearing and aided hearing-impaired listeners. *J. Acoust. Soc. Am.* 143, 1523–1533. doi: 10.1121/1.5026788

Xiang, M., and Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Lang. Cognit. Neurosci.* 30, 648–672. doi: 10.1080/23273798.2014.995679

Zhang, Z. D., and Chacron, M. J. (2016). Adaptation to second order stimulus features by electrosensory neurons causes ambiguity. *Sci. Rep.* 6:28716. doi: 10.1038/srep28716

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

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

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 29 September 2021
doi: 10.3389/fnins.2021.696513

The Relationship Between Central Auditory Tests and Neurocognitive Domains in Adults Living With HIV

Christopher E. Niemczak[1†], Jonathan D. Lichtenstein[2,3†], Albert Magohe[4†], Jennifer T. Amato[2,3], Abigail M. Fellows[1], Jiang Gui[5], Michael Huang[1], Catherine C. Rieke[1,2], Enica R. Massawe[4], Michael J. Boivin[6], Ndeserua Moshi[4] and Jay C. Buckey[1,2]*

[1] Space Medicine Innovations Laboratory, Geisel School of Medicine, Dartmouth College, Hanover, NH, United States, [2] Dartmouth-Hitchcock Medical Center, Lebanon, NH, United States, [3] Department of Psychiatry, Geisel School of Medicine, Dartmouth College, Hanover, NH, United States, [4] Department of Otorhinolaryngology, Muhimibili University of Health and Allied Sciences, Dar es Salaam, Tanzania, [5] Department of Data Science, Geisel School of Medicine, Dartmouth College, Hanover, NH, United States, [6] Department of Psychiatry, Michigan State University, East Lansing, MI, United States

OPEN ACCESS

Edited by:
Howard Charles Nusbaum, University of Chicago, United States

Reviewed by:
Larry E. Humes, Indiana University Bloomington, United States
Eliane Schochat, University of São Paulo, Brazil

***Correspondence:**
Christopher E. Niemczak
Christopher.e.niemczak@dartmouth.edu

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

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

Received: 16 April 2021
Accepted: 07 September 2021
Published: 29 September 2021

Citation:
Niemczak CE, Lichtenstein JD, Magohe A, Amato JT, Fellows AM, Gui J, Huang M, Rieke CC, Massawe ER, Boivin MJ, Moshi N and Buckey JC (2021) The Relationship Between Central Auditory Tests and Neurocognitive Domains in Adults Living With HIV. Front. Neurosci. 15:696513. doi: 10.3389/fnins.2021.696513

Objective: Tests requiring central auditory processing, such as speech perception-in-noise, are simple, time efficient, and correlate with cognitive processing. These tests may be useful for tracking brain function. Doing this effectively requires information on which tests correlate with overall cognitive function and specific cognitive domains. This study evaluated the relationship between selected central auditory focused tests and cognitive domains in a cohort of normal hearing adults living with HIV and HIV– controls. The long-term aim is determining the relationships between auditory processing and neurocognitive domains and applying this to analyzing cognitive function in HIV and other neurocognitive disorders longitudinally.

Method: Subjects were recruited from an ongoing study in Dar es Salaam, Tanzania. Central auditory measures included the Gap Detection Test (Gap), Hearing in Noise Test (HINT), and Triple Digit Test (TDT). Cognitive measures included variables from the Test of Variables of Attention (TOVA), Cogstate neurocognitive battery, and Kiswahili Montreal Cognitive Assessment (MoCA). The measures represented three cognitive domains: processing speed, learning, and working memory. Bootstrap resampling was used to calculate the mean and standard deviation of the proportion of variance explained by the individual central auditory tests for each cognitive measure. The association of cognitive measures with central auditory variables taking HIV status and age into account was determined using regression models.

Results: Hearing in Noise Tests and TDT were significantly associated with Cogstate learning and working memory tests. Gap was not significantly associated with any cognitive measure with age in the model. TDT explained the largest mean proportion of variance and had the strongest relationship to the MoCA and Cogstate tasks. With age in the model, HIV status did not affect the relationship between central auditory tests and cognitive measures. Age was strongly associated with multiple cognitive tests.

Conclusion: Central auditory tests were associated with measures of learning and working memory. Compared to the other central auditory tests, TDT was most strongly related to cognitive function. These findings expand on the association between auditory processing and cognitive domains seen in other studies and support evaluating these tests for tracking brain health in HIV and other neurocognitive disorders.

Keywords: HIV, cognition, central auditory processing, attention, auditory disease, cognitive processing speed, Africa South of the Sahara

INTRODUCTION

Central auditory function, reflected in tests of speech perception in background noise, correlates with cognition (Watson, 1991; Hallgren et al., 2001; Anderson and Kraus, 2010; Hoover et al., 2017; Panza et al., 2018; Danielsson et al., 2019; Humes, 2020, 2021), including cognitive dysfunction due to mild cognitive impairment (MCI), Alzheimer's disease (Gates et al., 1996; Idrizbegovic et al., 2011), and HIV (Zhan et al., 2017b; Buckey et al., 2019; White-Schwoch et al., 2020; Niemczak et al., 2021). This suggests central auditory tests might be useful for tracking cognitive dysfunction in populations with disordered neuro-cognitive processing. Yet, the relationship between central auditory function and cognition is multifactorial and questions remain regarding the correlation of central auditory tests with specific cognitive domains (i.e., processing speed, working memory, etc.). The ability to perceive, understand, and respond to a conversational partner in background noise encompasses a variety of neurocognitive domains (Gates et al., 1996; Anderson et al., 2013). While conversing in a noisy environment may seem like a simple, common task, detecting quick acoustic changes in an auditory stream and understanding the content, requires fast and accurate processing in the auditory processing pathway–including higher-level cognitive processing. The correlation between central auditory function and cognition suggests that auditory information is not only processed by the cochlea and auditory pathway, but also by other associative cortical areas (Harris et al., 2012; Palmer and Musiek, 2014; Sardone et al., 2019; Song et al., 2020). Our previous work has shown a relationship between central auditory test performance and cognitive performance, which suggested that central auditory tests might be useful for tracking cognitive performance in people living with HIV (PLWH) (Maro et al., 2014; Zhan et al., 2018; Buckey et al., 2019; Niemczak et al., 2021). The goal of this study was to evaluate the relationship between central auditory tests and neurocognitive domains in adults living with HIV and HIV-negative (HIV–) controls. Specifically, we examined how three central auditory tests relate to the specific cognitive domains of processing speed, learning, and working memory. The aim was to provide focused results on specific relationships of auditory processing and distinct neurocognitive domains to inform multifactorial longitudinal analyses in HIV and other neurocognitive disorders.

The worldwide prevalence of HIV/AIDS is approximately 37.9 million, with sub-Saharan African countries accounting for two thirds of the global HIV burden (Prevention, February 5, Centers for Disease Control and Prevention, 2020). Advances in understanding HIV replication, tracking immunologic progression, and using combination antiretroviral therapy (cART), have resulted in reduced viral loads and a drastic reduction in HIV mortality (Saylor et al., 2016). While rates of asymptomatic and mild neurocognitive dysfunction remain stable, cART has resulted in a significant decline in HIV-related dementia (Saylor et al., 2016). Despite this reduction in the severest forms of cognitive impairment, neurocognitive dysfunction persists in a subset of PLWH, even among those consistently taking cART and those with suppressed viral loads (Heaton et al., 2015). Identifying, tracking, and potentially predicting the development of neurocognitive dysfunction in this population would provide crucial benefits for PLWH. Identifying biomarkers for brain health in PLWH is important for reducing mortality, morbidity, and disease progression (Saylor et al., 2016). Cognitive impairment can lower treatment adherence and quality of life (Ettenhofer et al., 2010). Variability in how neurocognitive dysfunction presents in HIV relates to the diffuse nature of the disease's impact upon the central nervous system (Zhan et al., 2017a). Per recommendations from the National Institute of Neurological Diseases and Stroke, comprehensive neuropsychological assessment is considered the gold standard in assessing and monitoring neurocognition in PLWH (Antinori et al., 2007). This form of assessment covers a variety of cognitive domains, both broad and specific, providing an understanding of global function and particular domain-based skills (e.g., language, attention, memory). Neuropsychological assessment, however, is costly, time consuming, and requires specialized training for interpretation. It is also sensitive to education and cultural background. Access to such evaluations in resource limited settings, where the burden of HIV and other neurocognitive diseases tends to be significantly higher than in developed countries, is not always feasible. Additionally, finding culturally and linguistically appropriately normed measures is challenging (Robertson et al., 2009). Central auditory focused tests offer an attractive option in these settings because the tests are easy to explain, unlikely to be sensitive to education or cultural background, and can be repeated over time.

People living with HIV have differences in brain regions and functions necessary for auditory processing including gray matter atrophy, axonal injury, loss of axonal density, and diffuse white matter abnormalities in the internal capsule, thalamus, and corpus collosum (Zhan et al., 2017a; Kuhn et al., 2018). The auditory system provides a useful tool for assessing brain function because processing auditory information is neurologically demanding. Most people are familiar with tests of peripheral hearing (e.g., a threshold audiogram), but centrally

focused auditory tests involve much more than just assessing the quietest tone an individual can detect. After the cochlea converts sound waves into nerve signals the brain must quickly perform a series of complex functions to determine the meaning of the content. Speech perception, particularly interpreting speech in noise, engages several cortical and subcortical centers (Kotz and Schwartze, 2010; Specht, 2014). This involves neural pathways throughout the brainstem and into the cortex that integrate with high-level linguistic and cognitive systems, such as processing speed and working memory (Rudner and Lunner, 2014; Pichora-Fuller et al., 2016). Previous studies have shown that more complex auditory tasks relate to cognition beyond peripheral hearing sensitivity (Danielsson et al., 2019). A recent study by Danielsson et al. found associations between auditory processing tasks of temporal-order identification and gap detection with semantic long-term memory and working memory. Auditory thresholds had no significant effect on any of the cognitive measures. What makes central auditory focused tests so appealing is that they are relatively short (a gap detection test takes 5 min), easy to explain (the hearing-in-noise test and triple digit task involve identifying words or numbers in background noise), can be repeated, and do not require trained administrators. This is particularly important for deploying these tests in the developing world where normative cognitive data often do not exist, and education levels are variable. These measures would be a major advance for following PLWH in the developing world. Further understanding the relationship of central auditory tests with specific neurocognitive domains could also provide detailed knowledge for targeted longitudinal analyses to better assess, track, and potentially predict neurodegeneration in those with HIV and other neurocognitive diseases.

Our group has shown that PLWH with normal peripheral hearing are more likely to report trouble understanding speech in noise compared to controls (Maro et al., 2014; Zhan et al., 2017b). These studies suggest worse performance on measures of speech-in-noise identification may reflect damage to the central nervous system resulting in cognitive impairment (Buckey et al., 2019). Previous studies have shown a relationship between cognition and auditory processing ability such as speech perception in noise (Pichora-Fuller et al., 1995; Wong et al., 2009; Zekveld et al., 2011), auditory temporal ordering (Szymaszek et al., 2009; Danielsson et al., 2019; Humes, 2020;2021), and gap detection testing (Harris et al., 2010). In our previous work, we have shown a significant relationship between the ability to understand speech in noise and cognitive status in PLWH (Zhan et al., 2017b). Also, using fMRI we have shown activation in frontal areas during a challenging speech-in-noise task (Song et al., 2020) showing that challenging speech-in-noise tasks involve areas beyond the auditory cortex. In addition, a growing body of literature suggests central auditory processing deficits in PLWH on self-report, neurophysiological, and audiological measures (White-Schwoch et al., 2020; Niemczak et al., 2021). The relationship between measures of auditory processing and cognition likely relates to the similarities in specific neurocognitive domains used for advanced processes of auditory perception and higher-order cognition (Danielsson et al., 2019). The goal of this study was to examine a selection of central auditory tests and their relationship to specific cognitive domains to help select and inform which tests might be most sensitive in a longitudinal study to help detect cognitive changes over time. To do this, we focused on tests we believed had a strong central auditory component. We deliberately included individuals who had normal peripheral auditory function (i.e., normal audiograms) and selected tests that involve temporal processing and speech perception in noise. We hypothesized that central auditory tests would be associated with cognitive function in three domains of processing speed, learning, and working memory in PLWH. By using cognitive measures with various neurocognitive domains (speed of processing, learning, working memory), this analysis would also examine the specific domains of cognition that most strongly relate to central auditory performance. If central auditory tests are related to certain aspects of cognitive function, these measures might also provide a quantitative, time efficient, and repeatable measure of cognitive function in the developing world. In addition, once these relationships are identified, longitudinal analyses could be used to develop a predictive auditory screening tool related to multiple domains of cognitive function.

MATERIALS AND METHODS

Study Design and Participants

This prospective cohort study design was derived from a longitudinal study of HIV-infected adults (all taking cART) in Dar es Salaam, Tanzania. At the time of data collection for this analysis, we had gathered auditory and cognitive information from 259 HIV+ (mean 40.9 years) to 76 HIV– (mean 32.3 years) individuals from the larger study database. See **Table 1** for a complete review of sample demographics.

Procedures

The institutional review boards of both Dartmouth College and the Muhimbili University of Health and Allied Sciences approved this study's research protocol. All research was completed in

TABLE 1 | Summary of study demographics between HIV groups.

		HIV+(*n* = 259)	HIV–(*n* = 76)	*P*-value
Age (mean, SD)		40.9 (12.0)	32.3 (10.6)	<0.001
Gender (n, %)	*Male*	74 (29%)	31 (40%)	-
	Female	185 (71%)	39 (60%)	-
Years of education (mean, SD)		8.92 (2.61)	10.3 (2.71)	<0.001
Pure tone average (dB HL mean, SD)	*Right*	7.33 (4.86)	4.82 (5.50)	0.001
	Left	6.48 (5.09)	4.74 (5.49)	0.002
CD4 counts (mean, SD)		681.6 (318.2)	787.1 (212.1)	<0.001
Lowest CD4 counts (mean, SD)		406.7 (205.8)	659.9 (210.0)	<0.001

Mean and standard deviation (SD) of each demographic variable for HIV groups. P-values were calculated using the Mann-Whitney U-tests due to differences in distribution of groups, which were found to be significant across all tests (all ≤0.002).

accordance with the Helsinki Declaration. All participants were obligated to provide written informed consent. Participants consisted of PLWH and HIV– adults who were tested at the Infectious Disease Center in Dar es Salaam, Tanzania. To ensure accuracy of analysis and control for variables that could affect central auditory and cognitive function, we used a series of data selection techniques. Individuals were excluded if they had abnormal hearing sensitivity (>25 dB HL from 0.5 to 4 kHz) or abnormal middle ear function. Individuals were also excluded if they had a positive history of ear drainage, concussion, significant noise or chemical exposure, neurological disease, mental illness, ototoxic antibiotics (e.g., gentamycin), or chemotherapy. This selection technique resulted in 385 individuals, but only 335 individuals with complete central auditory variables (i.e., no missing auditory data) were selected for the study. We wanted every subject to have completed all the central auditory focused tests. The demographics of this sample population are provided in **Table 1**. To reduce the variation from using a single measurement from a single experimental session (cross-sectional sample), we used the mean from multiple visits over time. To do this, we plotted each subject's cognitive and central auditory tests over time and identified outliers by plotting a regression line to the data and removing values that were >2 standardized residuals from the regression line. After removing outliers, we took the average of each subject's test scores over time to create one data point for each subject.

Audiological testing was performed using a hearing assessment system built by the Creare, LLC. Creare's wireless noise attenuating headset (CWNAH) has the device speakers mounted in highly noise attenuating ear cups. The attenuation provided by this headset is better than any currently available commercial hearing test device as measured by an independent laboratory according to the relevant ANSI standards (Meinke et al., 2017). Before starting the audiological evaluation, patients had an otoscopic exam and cerumen was removed as needed. All subjects completed a health history questionnaire that asked about health conditions that might affect their hearing or central nervous system (e.g., head trauma, other central nervous system infection). To verify normal middle ear status, tympanometry at 226 Hz was performed on both ears using a Madsen Otoflex 100 (GN Otometrics, Denmark). Individuals with abnormal tympanograms (Type B or C) were referred for treatment and subsequent reevaluation.

Hearing thresholds were measured at frequencies 500, 1000, 2000, and 4000 Hz using a Békésy-like tracking procedure as described previously (Maro et al., 2014). Thresholds of 25 dB HL or better for each ear across the aforementioned frequencies were considered normal. Pure tone averages (PTA) were calculated by taking the mean of all measured frequencies. Measures of central auditory processing included gap detection, speech-in-noise, and digits-in-noise testing. The Kiswahili language version of the Hearing in Noise Test (HINT) and the Kiswahili Triple Digit Test (TDT) were used to assess speech perception in noise. The HINT was administered in three test conditions: Noise Front, Noise Right, and Noise Left. In each HINT test, a different list of 20 sentences was presented in random order in the presence of the masking noise spectrally matched to the long-term average of the target material. The presentation level of the noise remained fixed at 65 dB (A-weighting), and the test instrument adjusted the level of each sentence adaptively depending on whether the test administrator indicated that the previous sentence was repeated correctly. The presentation level of the sentence was reduced if the previous sentence was repeated correctly and increased if the previous sentence was repeated incorrectly. This adaptive procedure was used to determine the presentation level of each sentence in the list. The average presentation level of all sentences after the first four sentences defined the speech reception threshold (SRT) for the test condition expressed as a signal to noise ratio (SNR). A composite SNR of all three noise conditions was calculated and used as the primary variable of interest for the HINT.

In the TDT, recordings of natural productions of three-digit triplets such as 3-5-9 (spoken as "tatu-tano-tisa" in Kiswahili) were used as target stimuli (Kiswahili numbers below 10 have the same number of syllables). All digit triplets were produced and recorded by a male speaker in a soundproof booth. Digit triplet recognition was tested in the presence of competing Schroeder-phase masking noise. The test included 30 total presentations of pseudorandom digit triplets with 6 practice presentations. Presentations were presented in pairs of positive and negative phase maskers. Each pair was presented at the same SNR. The ordering of the masker was randomized for each pair. The test started at a 0-dB initial SNR with the masker fixed at 75 dB SPL. SNRs were then adjusted after each presentation or pair of presentations by varying the target level. 2.0 dB SPL was added to the target level for each incorrect digit and 2.0 dB SPL was subtracted for each correct digit from the previous positive-phase presentation. A speech reception threshold was calculated as the SNR of the last 7 positive-phase presentations, which was used as the primary variable of interest. Completing one list of 30-digit triplets took 3-6 min.

The Gap Detection Test (GAP) determines a participant's ability to identify short gaps in noise by pushing a button when they first identify a break in noise. The details of the gap detection test have been published previously (Maro et al., 2014). We implemented an adaptive gap detection algorithm using a single staircase. Gaps occur randomly in the middle portion of 4.5 s of white noise delivered at 65 dB SPL. No gaps were presented in the first or last second. If the subject misidentified 2 gaps, in a row or 3 gaps overall, the staircase "reversed," and the gap length increased. In this way, the staircase algorithm converged to the subject's gap threshold. The test started at a gap length of 20 m sec and continued until the subject completed 10 reversals or a total of 120 presentations. From these gap tests a plot of the percentage of time a gap was correctly detected vs. gap length was produced. This curve can be fit using the Hill equation to calculate the gap length where 50% of the gaps were detected correctly (Grigoryan et al., 2013). These values were used in the analysis. The subjects received training in the gap test with both a training video and a screen that provided both auditory and visual feedback. The operator presented gaps to the subject until the subject comprehended the task.

The Kiswahili version of the Montreal Cognitive Assessment (MoCA) was used as a screening measure of cognitive

function. The MoCA has been used in diverse populations and has demonstrated reliability and validity with the Tanzanian population (Vissoci et al., 2019). This measure assesses short-term and long-term memory recall, visual spatial abilities, executive functioning, attention, concentration, and working memory. The maximum score achievable is 30 points and a correction was implemented based on years of education.

Computerized neurocognitive assessments were conducted using the Cogstate and Test of Variables of Attention (TOVA). All cognitive assessments utilized visual stimuli, not auditory stimuli, to test neurocognitive function (apart from the instructions). The Cogstate[1] is comprised of multiple tests assessing a variety of domains. The Cogstate battery, was chosen because it uses culturally neutral stimuli (e.g., playing cards) to ensure that the assessment is not limited by a participant's level of education. Card games are popular in Tanzania, so the playing card approach was familiar to the cohort. The Cogstate tasks are computer-based and designed for repeated administration. The Cogstate battery has been used to assess cognitive function in patients with HIV and has been shown to correlate well with standard neuropsychological test batteries (Cysique et al., 2006; Overton et al., 2011; Bloch et al., 2016; Kamminga et al., 2017). In addition, the Cogstate includes tasks known to be sensitive to cognitive domains affected in adults (Hammers et al., 2011, 2012; Bloch et al., 2016; Boivin et al., 2016). We chose outcomes measures of working memory and learning from the Cogstate to include in our analysis. The Test of Variables of Attention (TOVA. Version 8.0) was also administered to all subjects as it has been used with sub-Saharan African populations across a wide range of studies (Bangirana et al., 2015; Boivin et al., 2019). The TOVA has several advantages because it uses visual stimuli, measures response times precisely (±1 m sec), is language- and culture-free, and has a history of use in resource-challenged areas (Ruel et al., 2012). **Table 2** has a list of the measures, their cognitive domains, and the outcome variables used in our analyses.

[1]www.Cogstate.com

Statistical Analysis

Data were analyzed and plotted using MATLAB® 2020b. As stated above, a set of cognitive outcome variables were determined *a priori* based on the sensitivity of specific domains, their correlation to standard neuropsychological test batteries, and cognitive domains likely to be affected by HIV (see **Table 2**). We first tested simple group mean differences using non-parametric procedure (Mann-Whitney U-test). A bootstrapping method was used to examine the relationship of central auditory tests to cognitive measures. This resampling technique was used to estimate the mean and standard deviation of the proportion of variance (R^2) and overall significance (p-Value) of the linear relationship between each central auditory test and cognitive measure. We randomly sampled ∼60% of the cohort (200 individuals) 5000 times for each central auditory/cognitive relationship. To examine the effects of age and HIV, we used general linear models to assess the association between central auditory tests and cognitive measures with age and HIV included in the model [model specification - ($Cognitive\ measure \sim Gap + HINT + TDT + HIV\ status + Age$)]. Previous studies have shown a strong relationship between age, auditory tests, and cognitive measures (see Humes et al., 2012 for review). This analysis method resulted in 7 models, one for each cognitive measure. To adjust for multiple comparisons, we used an α level

TABLE 2 | Descriptions of cognitive measures.

Cognitive measure	Domain assessed	Test description	Outcome measure
MoCA			
Kiswahili MoCA	Neurocognitive Screening	Screens short- and long-term memory recall, visual spatial abilities, executive functioning, attention, concentration, and working memory.	A subject is able to obtain a total of 30 points. Higher scores represent better performance.
TOVA			
Response Time *(RT)*	Attention, processing speed *(RT-ms)*	Correct response time mean. Measures quickness of response time in milliseconds.	The mean response time of the correct responses. Lower = better.
Ex-Gaussian μ *(ExGμ)*	Attention, processing speed *(ExGμ-ms)*	The mean response time (in milliseconds) of the correct responses, modeled using the Ex-Gaussian distribution.	This score takes into account the skew in distribution of reaction time scores. Lower = better.
Cogstate			
One Card Learning *(OCL)*	Visual leaning accuracy* *(OCL-acc)*	Accuracy of performance; arcsine square root proportion correct.	Accuracy is the primary outcome for OCL. Higher scores = better.
	Visual learning speed *(OCL-lmn)*	Speed of performance; mean of the $\log_{10}$ transformed reaction times for correct responses.	Speed of processing is the secondary outcome measure for OCL. Lower score = better
One Back Test *(ONB)*	Working memory accuracy *(ONB-acc)*	Accuracy of performance; arcsine square root proportion correct.	Accuracy is the secondary outcome for ONB. Higher scores = better.
	Working memory speed* (ONB-lmn)	Speed of performance; transformed reaction times for correct responses.	Speed of processing is the primary measure for ONB. Lower score = better

*All cognitive measures, the domain they assess, a description of the test, and the outcome measures are listed for the entire study. MoCA, Montreal Cognitive Assessment; TOVA, Test of Variables of Attention. *Indicates the primary variable of interest for Cogstate subtest.*

of <0.005. For all model analyses, a variance inflation factor was calculated to assess collinearity among the predictor variables (Gap, HINT, and TDT). All variance inflation factors were <1.8 indicating low levels of collinearity between predictor variables. This provided an alternative way to examine the sampling distribution of proportion of variance and overall significance.

RESULTS

Overall Mean HIV+ and HIV– Group Comparisons

Overall demographic differences between groups showed that HIV + individuals were older, less educated, and had pure tone averages that were higher than the HIV– group. See **Table 2** for demographic differences. Mean differences between central auditory and cognitive measures also revealed significant HIV group differences. See **Table 3** for mean and standard deviations of cognitive and central auditory processing outcome variables. All central auditory variables revealed highly significant differences between HIV groups ($p \leq 0.012$). MoCA, TOVA, and Cogstate also revealed significant differences between groups ($p \leq 0.037$). Importantly, these differences do not show the association between auditory and cognitive variables.

Relationship of Specific Central Auditory Tests to Cognitive Measures

To evaluate the relationship of central auditory variables to specific cognitive domains, we used a bootstrapping method to assess the relationship between each individual auditory variable and each cognitive measure. The most significant results were the relationships of the TDT to the MoCA, OCL-lmn, and ONB-lmn (all mean $p < 0.001$) with mean R^2 values of 0.142, 0.113, and 0.120, respectively (**Table 4**). All central auditory focused tests were significantly related to the MoCA, but the TDT relationship was much stronger than for either the HINT or Gap and was the only significant measure when multiple comparisons were considered. The Gap test was significantly related to the TOVA measures, but neither the HINT nor TDT were related to the TOVA. The strongest results were for the TDT with the OCL and ONB speed of processing measures. The HINT was only weakly related to these measures. Interestingly, the HINT was significantly related to the OCL and ONB accuracy measures, while the TDT and Gap were not.

Association of Cognitive Measures to Central Auditory Measures Including Age and HIV Status

We evaluated the association of cognitive test domains to central auditory tests, including age and HIV status in the model. Overall, we found significant associations between multiple cognitive and central auditory test variables. **Table 5** shows the results of all linear regression models. MoCA scores were significantly associated with the TDT ($p < 0.001$), but not with any other variable. TOVA response time (RT) and ex-gaussian μ (ExGμ) were significantly associated with age ($p < 0.001$) but no central auditory/cognitive relationship was found to be significant. Cogstate OCL accuracy (OCL-acc) was only significantly associated with HINT ($p = 0.004$), while the speed of learning (OCL-lmn) was significantly associated with age ($p < 0.001$) and TDT ($p < 0.001$). Cogstate ONB accuracy (ONB-acc) was significantly associated with HINT ($p = 0.003$), while working memory speed (ONB-lmn) was significantly associated with age ($p = 0.001$) and TDT ($p = 0.002$). Together, these results suggest that: (1) Gap was not significantly related to any of the cognitive measures independent of age; (2) The HINT was significantly related to OCL-acc and ONB-acc, and (3) The TDT showed the strongest associations with significant relationships to the MoCA, OCL-lmn, and ONB-lmn. Age also showed strong associations with TOVA response time, ExGμ, OCL-lmn, and ONB-lmn. R^2 values were also strongest for the OCL-lmn (0.253) and ONB-lmn (0.257) models. Surprisingly, HIV status was not significant in any model and age was not significantly related to the MoCA. **Figure 1**. shows plots of OCL-lmn and ONB-lmn for

TABLE 3 | Mean central auditory and cognitive performance differences between HIV groups.

		Mean performance scores		
Measure	**Subtest**	**HIV + (SD)**	**HIV– (SD)**	***P*-value**
Central auditory processing	Gap Detection Test (*GAP* - ms)	3.65 (1.37)	3.00 (1.18)	<0.001
	Hearing in Noise Test (*HINT* - dB SNR)	–10.7 (0.99)	–11.1 (1.01)	0.012
	Triple Digit Test (*TDT* - dB SNR)	–18.1 (4.70)	–20.1 (4.63)	<0.001
MoCA	Total Score (# correct)	26.6 (2.86)	27.5 (2.30)	0.012
TOVA	Reaction Time *(RT - ms)*	395.1 (62.9)	364.7 (63.2)	<0.001
	Ex-Gaussian μ *(ExGμ - ms)*	324.7 (59.0)	296.1 (51.5)	<0.001
Cogstate	*One Card Learning – accuracy *(OCL-acc)*	0.94 (0.10)	0.96 (0.10)	0.013
	One Card Learning – speed of processing *(OCL-lmn)*	3.17 (0.11)	3.10 (0.09)	<0.001
	*One Back Test – speed of processing *(ONB-lmn)*	3.08 (0.10)	3.01 (0.09)	<0.001
	One Back Test – accuracy *(ONB-acc)*	1.07 (0.20)	1.12 (0.17)	0.037

*Mean and standard deviation (SD) are displayed for HIV + and HIV– groups. Significant differences were found on all central auditory variables and cognitive measures. P-values were calculated using Mann-Whitney U-tests (non-parametric) due to differences in distribution between HIV groups. *Indicates the primary outcome measures for Cogstate subtest. Each measure is displayed in raw score units along with the applicable abbreviation in italic.*

TABLE 4 | Bootstrap results.

			Bootstrap results			
Cognitive variable	**Response variable**	**Predictor variable**	**Mean R^2**	**Std. R^2**	**Mean *p*-Value**	**Std. *p*-Value**
MoCA	*Total score*	Gap	0.036	0.016	0.033	0.056
		HINT	0.042	0.022	0.034	0.072
		TDT	0.142	0.028	<0.001	<0.001
TOVA	*Response time*	Gap	0.039	0.017	0.021	0.043
		HINT	0.012	0.004	0.498	0.205
		TDT	0.019	0.010	0.158	0.169
	ExGμ	Gap	0.030	0.016	0.045	0.076
		HINT	0.002	0.004	0.582	0.260
		TDT	0.011	0.008	0.234	0.196
Cogstate	OCL-acc*	Gap	0.011	0.009	0.271	0.023
		HINT	0.037	0.017	0.035	0.073
		TDT	0.018	0.010	0.132	0.144
	OCL-lmn	Gap	0.029	0.017	0.056	0.089
		HINT	0.038	0.018	0.024	0.048
		TDT	0.113	0.025	<0.001	0.001
	ONB-acc	Gap	0.011	0.009	0.254	0.232
		HINT	0.096	0.018	0.012	0.032
		TDT	0.032	0.017	0.045	0.076
	*ONB-lmn**	Gap	0.027	0.014	0.061	0.093
		HINT	0.032	0.017	0.044	0.079
		TDT	0.120	0.022	<0.001	<0.001

*The results of resampling between individual central auditory and cognitive measures are displayed below. 200 random samples of the 335 subject cohort were chosen over 5000 iterations to calculate the mean and standard deviation (Std.) R^2 and p-Value. The relationship of TDT and MoCA, OCL-lmn, and ONB-lmn showed the largest R^2 with mean p-Values below 0.001. *Indicates the primary outcome measures for Cogstate subtest. The colored values indicate significant of $p < 0.005$.*

the TDT. The graphs show a significant positive trend for the entire cohort and for the HIV + group to perform slightly worse on both measures.

DISCUSSION

The findings of the current study support and build upon previous work in understanding how central auditory measures and cognition are associated. The results showed that central auditory tests were significantly associated with cognitive abilities across areas of learning and working memory on the Cogstate, but not response time as measured by the TOVA. The HINT and TDT showed the strongest associations with learning and working memory on the Cogstate. Specifically, the TDT showed the strongest relationship to OCL-lmn and ONB-lmn in the linear models and with the bootstrapping resampling method. Surprisingly, HIV status was not significantly associated with any cognitive measure. These findings support the idea that central auditory tests assess cognitive functions related to the domains of learning and working memory. Longitudinal results are needed to determine whether central auditory focused tests could be used to track cognitive function over time in patients with particular conditions that affect the central nervous system, such as HIV and other neurocognitive diseases.

With speech-in-noise measures, we found that subjects who performed better on HINT and TDT also performed better on OCL and ONB tasks related to learning and working memory. Speech perception in noise has been linked to multiple cognitive functions (i.e., executive and attentive functions) (Anderson and Kraus, 2010; Anderson et al., 2013; Moore et al., 2014). The rationale for this link is based on the complex acoustic processing necessary to perceive word representations and phonemes accurately, extract the meaning of the message within background noise successfully, and remember the message from beginning to end to respond correctly. This process requires cognitive-linguistic abilities, including working memory to attend to and recall what is being said (Craik, 2007; Anderson and Kraus, 2010; Rönnberg et al., 2013). Success in the learning domain, as evidenced by OCL tasks, also requires recruitment of complex neural networks across a broad range of cortical and subcortical areas (Koziol and Budding, 2009; Du Plessis et al., 2017). Recently, using fMRI we found increased activation in the right superior and middle frontal gyrus during speech perception in noise in a cohort of normal hearing Mandarin listeners (Song et al., 2020). These two frontal sub-regions are relevant for tone perception, phonological working memory, and orientation of attention (Husain et al., 2006; Japee et al., 2015). Therefore, this provides evidence that speech-in-noise tests are tapping into areas of the brain also associated with tests such as the ONB and OCL.

The strongest relationships were seen with the TDT. The TDT is a closed set task (numbers 1-9) with no spatial adjustment and uses Schroder-phase masking noise (positive Schroder-phase

TABLE 5 | Linear regression results.

				Linear regression results			
Cognitive variable	**Response variable**	**Predictor variable**	**Beta**	**Std. Err.**	**t-Stat**	***p*-Value**	**R^2**
MoCA	*Total score*	HIV	0.060	0.387	0.156	0.876	0.177
		Age	−0.025	0.014	−1.769	0.078	
		TDT	−0.173	0.036	−4.780	<0.001	
		HINT	−0.310	0.151	−2.052	0.041	
		Gap	−0.174	0.118	−1.478	0.140	
TOVA	*Response time*	HIV	−11.168	8.771	−1.273	0.204	0.145
		Age	1.629	0.323	5.049	<0.001	
		TDT	−0.171	0.814	−0.210	0.834	
		HINT	−3.052	3.513	−0.869	0.386	
		Gap	3.239	2.729	1.187	0.236	
	ExGμ	HIV	−9.722	7.741	−1.256	0.210	0.198
		Age	1.982	0.284	6.984	<0.001	
		TDT	−0.661	0.718	−0.921	0.358	
		HINT	−2.589	3.073	−0.842	0.400	
		Gap	0.466	2.397	0.195	0.084	
Cogstate	*OCL-acc**	HIV	0.008	0.014	0.579	0.563	0.069
		Age	0.000	0.001	−0.853	0.394	
		TDT	−0.001	0.001	−0.634	0.526	
		HINT	−0.016	0.006	−2.792	0.004	
		Gap	−0.004	0.004	−0.806	0.421	
	OCL-lmn	HIV	−0.015	0.015	−0.976	0.330	0.253
		Age	0.003	0.001	5.813	<0.001	
		TDT	0.005	0.001	3.358	0.001	
		HINT	0.011	0.006	1.757	0.080	
		Gap	−0.001	0.005	−0.265	0.791	
	ONB-acc	HIV	0.005	0.026	0.187	0.852	0.097
		Age	−0.002	0.001	−2.132	0.034	
		TDT	−0.003	0.002	−1.258	0.209	
		HINT	−0.030	0.010	−2.870	0.003	
		Gap	−0.003	0.008	−0.341	0.733	
	*ONB-lmn**	HIV	−0.019	0.013	−1.489	0.137	0.257
		Age	0.003	0.000	6.076	0.001	
		TDT	0.004	0.001	3.109	0.002	
		HINT	0.006	0.005	1.255	0.211	
		Gap	−0.002	0.004	−0.472	0.637	

*Results display the association between central auditory and cognitive measures as well as HIV status and age. Models show decreased performance on central auditory tests are significantly associated with degraded cognitive measures (Model specification: Cognitive measure∼ HIV status + Age + TDT + HINT + Gap). Significant values are highlighted in gray. *Indicates the primary outcome measures for Cogstate subtest. T-stat was calculated by dividing the Beta (coefficient) by the standard error.*

noise is used to calculate the SNR). TDT yielded the largest R^2 values and lowest *p*-Values and showed the strongest associations to OCL-lmn, ONB-lmn, and MoCA. One possible reason for these findings is the global cognitive nature of both the TDT and -lmn/MoCA measures. The OCL and ONB-lmn measures are the mean of the $\log_{10}$ transformed reaction times for correct responses to tasks related to playing cards. The -lmn measures could be interpreted as combination of accuracy, speed, and memory. This is similar to the findings of Humes et al. (2013) and others, who showed the importance of global sensory-processing performance to global cognitive function (Humes et al., 2013; Humes, 2015; Deal et al., 2016; Glick and Sharma, 2020). The working memory components of the ONB and OCL tasks could also be related to auditory working memory, which has been shown to be an important component of language comprehension, even in the absence of background noise (Daneman and Merikle, 1996; Wingfield and Tun, 2007). The addition of background noise in the TDT may reduce auditory working memory capacity, resulting in the decreased ability to rehearse and recall a target, further compromising the perception of the signal already degraded by noise (Pichora-Fuller et al., 1995; Parbery-Clark et al., 2011). Age effects were also observed on ONB-lmn and OCL-lmn measures. The degree of age-related function on working memory has shown to be dependent of task difficult with larger age effects with higher cognitive demands (Salthouse and Meinz, 1995; Nilsson, 2003; Danielsson et al., 2019). The association of TDT to MoCA also provides evidence for a global hypothesis. The MoCA is a popular screening test for cognitive impairment that covers key cognitive domains including episodic memory, language, attention, orientation, visuospatial ability and executive functions (Nasreddine et al., 2005). In those with mild cognitive impairment speech-in-noise processing may increase the recruitment of neural networks involved in memory, attention, and learning to compensate for this dysfunction (Iliadou et al., 2017; Jalaei et al., 2019). Another hypothesized reason for the association between TDT and cognitive measures is the cross-cultural simplicity. While the Gap does not require any linguistic interpretation and the HINT sentences have been translated accurately into Swahili, the TDT provides a simple closed-set speech-in-noise task with minimal cultural or educational influences.

A surprising finding was that although both the HINT and TDT are speech-in-noise tasks, the TDT showed much stronger relationships with cognitive test results. Also, the unique pattern of the HINT relating to accuracy, and TDT relating to speed of working memory and visual learning on the ONB and OCL tasks was unexpected. Both the HINT and TDT are speech-in-noise tasks that use a SNR metric as the outcome variable. But the HINT is a composite score of three spatial orientations of the speaker and background noise. The HINT also uses speech-shaped noise to the long-term average of the presented sentences. Also, the HINT requires repeating back entire sentences, where the overall meaning of the sentence can provide some context to help the subject repeat it correctly. The TDT, however, requires remembering three unrelated numbers. The reasons for the differences between the tests are not clear, but does show that not all measures including speech in noise are the same.

Overall, our findings support the study of central auditory tests for measuring function on neurocognitive domains of learning and working memory regardless of HIV status. In the current study, we did not find strong evidence for HIV affecting the association between central auditory performance and cognitive measures. We hypothesized that PLWH would show differences in auditory processing due the diffuse effect of HIV on the central nervous system such as gray matter atrophy, diffuse

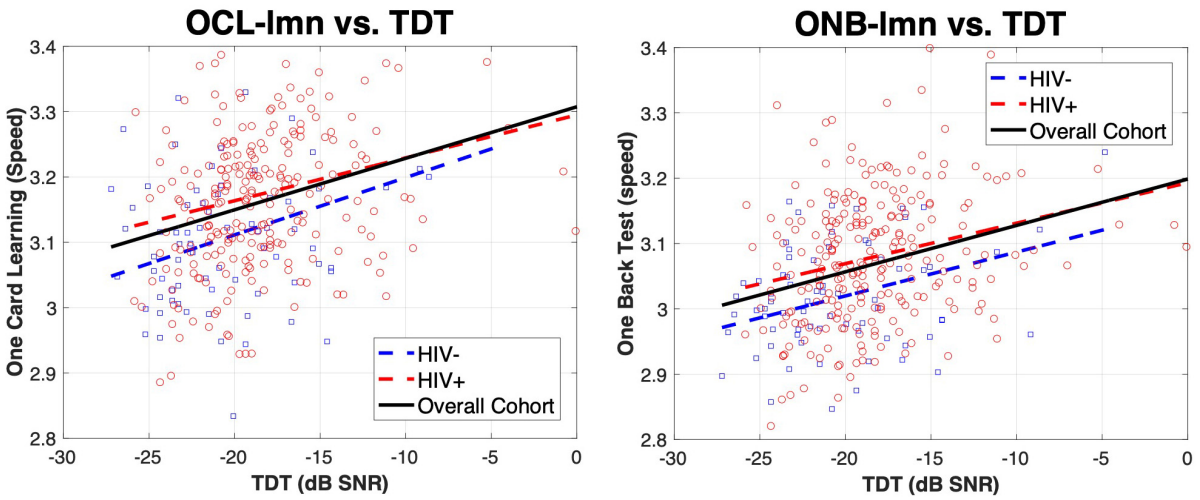

FIGURE 1 | Relationship of OCL-Imn and ONB-Imn with TDT for HIV– and HIV + individuals. Panels show the relationship of OCL-Imn (left) and ONB-Imn (right) for HIV– (blue squares) and HI + (red circles). Lines indicate a least-squares fit for HIV– (blue dotted line), HIV + (red dotted line), and the overall cohort (black solid line). OCL-Imn is the mean of the $\log_{10}$ transformed reaction times for correct responses during the One Card Learning task. This task was to quickly and accurately recognize if a playing card has been presented before throughout the duration test. Lower scores equate to better performance. ONB-Imn is also the $\log_{10}$ transformed reaction times for the One Back Test using playing cards. Results show significant positive relationships between TDT and OCL-Imn and ONB-Imn.

white matter abnormalities in the internal capsule, thalamus, and corpus collosum, axonal injury, and loss of axonal density (Kuhn et al., 2018). We did find overall mean differences between HIV groups (**Table 3**), and differences in distribution of the scatter points for HIV– vs. HIV + individuals (**Figure 1**), but no significant difference in the relationship between central auditory and cognitive measures when age was in the model. These findings could be because many PLWH are performing to similar levels of those without the disease due to modern antiretroviral therapy. Additionally, it may be that only a subset of those with HIV show a difference in the relationship of central auditory tests and cognitive measures and this is not apparent in the overall measures. The CD4 levels of our cohort are consistent with HIV + individuals with well-controlled HIV on cART indicating that HIV was mostly well-controlled in this group. Also, it may be that although there were more PLWH with cognitive difficulties, the relationship between cognition and the central auditory variables is not different from what occurs due to aging. Our results support that age is a significant factor in the auditory-cognitive relationship. PLWH may be experiencing "accelerated aging" and may just be at a higher point on the curve relating cognition to central auditory test performance. Longitudinal analysis of this relationship is warranted to better assess this "accelerated aging" hypothesis. Regardless, if deficits in auditory function precede cognitive decline and ultimately brain health, early detection of cognitive deficits may be possible by evaluating aspects of central auditory processing.

We found central auditory tests were significantly associated with learning and working memory on the Cogstate, but not response time on the TOVA. The results are interesting because, except for providing instructions in performing the tests, none of the cognitive tests used in the study have an auditory component for their execution. By examining temporal acuity in an auditory gap detection paradigm, we found that age was significantly related to TOVA response time and Ex-Gaussian Mu with older adults having poorer response time compared to young adults. These observed age effects are consistent with previous studies that have suggested that age-related differences in complex measures of auditory temporal processing may be explained, in part, by age-related deficits in processing speed and attention (Snell, 1997; Humes et al., 2009, 2012; Harris et al., 2010; Humes, 2021). Age has a large effect on cognitive speed, which declines earlier and at a higher rate than memory (Salthouse, 2009). Humes et al. (2009) found a significant correlation between auditory and visual gap detection that was associated with general cognitive function, but not with processing speed on the Wechsler Adult Intelligence Scale (WAIS-III) in older adults (60-88 years). While there was a difference in age between cohorts and we used the TOVA instead of the WAIS-III, our results support previous findings that show gap detection is not related to processing speed when age is included in the model. Continuing to study gap detection and processing speed is important because they have been linked to age related changes in speech recognition, especially in acoustically complex conditions (Snell, 1997; Palmer and Musiek, 2014). But with the TOVA used in this study, we cannot endorse a significant relationship between processing speed and gap detection accounting for age in our experimental cohort.

The central auditory relationships to cognition are particularly important because central auditory tests can be easy to administer and do not require extensive training. Measures of central auditory processing are easy to train people on how to use, not complicated to administer, and not restricted to English. This makes them particularly appealing for use in resource-limited settings, such as developing nations. In contrast, cognitive tests can be challenging to train people

how to use, are resource-intensive (e.g., materials), are often sensitive to education, and have typically been created for Western cultures. These are all barriers to their use in resource-limited, international settings. Most central auditory tests can be completed faster than traditional full cognitive test batteries (NIH toolbox takes approximately 2 h to complete) and some take less time than a cognitive screening test (10 min for the MoCA). The entire central auditory test battery (GAP, HINT, and TDT) used in this study took approximately 15 min, while the TOVA alone took over 20 min. These factors provide compelling arguments for exploring the use of central auditory tests to track cognitive performance in resource limited settings.

This study has some limitations. Previous work has demonstrated that age is an independent predictor of cognitive performance and speech in noise ability in HIV + adults (Zhan et al., 2017b). Processing speed is one of the strongest predictors of performance across cognitive tasks in adults (Salthouse and Ferrer-Caja, 2003) and age-related changes in processing speed are well established (Verhaeghen and Salthouse, 1997) and supported in this study. While age-related changes were not the focus of the current study, their influence on the auditory/cognitive relationship cannot be ignored. We observed an age effect on both TOVA measures, OCL-lmn, and ONB-lmn. Interestingly, all of these measures have a speed of processing component. Yet, even with age effects included in the model, central auditory tests were still significantly associated with cognitive function on the Cogstate.

Regarding gap detection testing and cognitive speed of processing, we only used gap detection thresholds as our primary outcome variable. There may be additional data within the test that were not fully utilized in this study. For example, measuring response time at the gap detection thresholds or even plotting a curve to each individual's gap detection test from 0% correct at short gaps to 100% at longer gaps. Further development of these methods may produce a more sensitive test in detecting cognitive decline in HIV or other population with neurocognitive decline. Another limitation is that the present study does not address whether central auditory tests can be used to predict or track cognitive function in individuals over time. Examining the relationship between cognition and central auditory measures and whether this affected by HIV requires a longitudinal study. Instead, this study was focused on what cognitive domains were most strongly related to central auditory tests in this mixed cohort of normal hearing individuals that spanned a large age range. Result from this study do not prove causal relations, but only significant associations between central auditory and cognitive measures. Also, our test battery was limited in its scope. With a different set of cognitive tests (and additional domains), relationships might have been more robust. In addition, exploring the relationship of central auditory processing and cognition within HIV-infected individuals should be completed in populations outside of sub-Saharan Africa. This will allow for a better understanding of the generalizability of our findings. Further analysis of longitudinal CD4 count history may also provide a continuous metric of infection severity, which may be more predictive of cognitive dysfunction. With modern cART, lower CD4 counts have been exceedingly rare in our cohort and the general health of all subjects is consistent with well-managed HIV infection. Further longitudinal analysis of those who develop lower CD4 counts (e.g., <200 cell/μl) warranted. Examination of central auditory processing over time in HIV + individuals is also needed, as such work might provide further insight into the concept of accelerated aging in the brain in HIV-infected persons, and whether this responds to interventions.

CONCLUSION

The overall results from the study suggest central auditory focused tests are positively related to cognitive function in our cohort, particularly in the areas of learning and working memory. The Gap test was not related to any cognitive measure with age in the model, while the HINT and TDT were related to learning and working memory. TDT scores were also found to be significantly related to the MoCA. We did not find any evidence of a HIV effect on the association of central auditory and cognitive measures. With longitudinal confirmation, central auditory tests, specifically speech-in-noise tests such as the TDT could provide an easy-to-use, quick method for assessing and potentially predicting cognitive dysfunction in those with HIV and other related cognitive deficits.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Dartmouth College. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CN completed the final data and statistical analyses, created the tables and figures, and completed the final draft of the manuscript. JL participated in the analysis plan, statistical analyses, interpretation of findings, and completed the initial complete draft of the manuscript. AM supervised the assessment team, coordinated the participant scheduling and follow-up, provided quality assurance, reviewed the data, and supported institutional review board approvals. JA completed preliminary data analyses, tables, and co-authored the completed first draft of the manuscript. AF verified data integrity, trained the assessment team, help to lead quality assurance for testing and data management, coordinated standard operating procedures, coordinated monthly conference call meetings for the study team, and supported institutional review board approvals. MH assisted in data analysis. CR verified data integrity and assisted with data interpretation. EM assisted with study management, and data review. MB participated in study design, test selection, and

consulted to the project. NM assisted with study design and was primarily responsible for all aspects of study oversight in Tanzania. JB was study primary investigator, and participated in all phases of study conceptualization, study design, proposal writing, analysis plan, study implementation of protocols, interpretation, and manuscript editing. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the National Institutes of Health (NIH), grant number 5R01DC009972-10 to principal investigator JB. The content of this report is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

REFERENCES

Anderson, S., and Kraus, N. (2010). Sensory-Cognitive Interaction in the Neural Encoding of Speech in Noise: A Review. *J. Am. Acad. Audiol.* 21, 575–585. doi: 10.3766/jaaa.21.9.3

Anderson, S., White-Schwoch, T., Parbery-Clark, A., and Kraus, N. (2013). A dynamic auditory-cognitive system supports speech-in-noise perception in older adults. *Hear. Res.* 300, 18–32. doi: 10.1016/j.heares.2013.03.006

Antinori, A., Arendt, G., Becker, J. T., Brew, B. J., Byrd, D. A., Cherner, M., et al. (2007). Updated research nosology for HIV-associated neurocognitive disorders. *Neurology* 69, 1789–1799. doi: 10.1212/01.WNL.0000287431.88658.8b

Bangirana, P., Sikorskii, A., Giordani, B., Nakasujja, N., and Boivin, M. J. (2015). Validation of the CogState battery for rapid neurocognitive assessment in Ugandan school age children. *Child Adolesc. Psychiatry Ment. Health* 9:38. doi: 10.1186/s13034-015-0063-6

Bloch, M., Kamminga, J., Jayewardene, A., Bailey, M., Carberry, A., Vincent, T., et al. (2016). A Screening Strategy for HIV-Associated Neurocognitive Disorders That Accurately Identifies Patients Requiring Neurological Review. *Clin. Infect. Dis.* 63, 687–693. doi: 10.1093/cid/ciw399

Boivin, M. J., Chernoff, M., Fairlie, L., Laughton, B., Zimmer, B., Joyce, C., et al. (2019). African Multi-Site 2-Year Neuropsychological Study of School-Age Children Perinatally Infected, Exposed, and Unexposed to Human Immunodeficiency Virus. *Clin. Infect. Dis.* 71, e105–e114. doi: 10.1093/cid/ciz1088

Boivin, M. J., Nakasujja, N., Sikorskii, A., Opoka, R. O., and Giordani, B. (2016). A Randomized Controlled Trial to Evaluate if Computerized Cognitive Rehabilitation Improves Neurocognition in Ugandan Children with HIV. *Aids Res. Hum. Retrovir.* 32, 743–755. doi: 10.1089/aid.2016.0026

Buckey, J. C., Fellows, A. M., Magohe, A., Maro, I., Gui, J., Clavier, O., et al. (2019). Hearing complaints in HIV infection originate in the brain not the ear. *AIDS* 33, 1449–1454. doi: 10.1097/QAD.0000000000002229

Centers for Disease Control and Prevention (2020). *Global HIV & Tuberculosis*. Available Online at: https://www.cdc.gov/globalhivtb/ [Accessed February 20, 2020 2020]

Craik, F. I. (2007). The role of cognition in age-related hearing loss. *J. Am. Acad. Audiol.* 18, 539–547. doi: 10.3766/jaaa.18.7.2

Cysique, L. A., Maruff, P., Darby, D., and Brew, B. J. (2006). The assessment of cognitive function in advanced HIV-1 infection and AIDS dementia complex using a new computerised cognitive test battery. *Arch. Clin. Neuropsychol.* 21, 185–194. doi: 10.1016/j.acn.2005.07.011

Daneman, M., and Merikle, P. M. (1996). Working memory and language comprehension: a meta-analysis. *Psychon. Bull. Rev.* 3, 422–433. doi: 10.3758/BF03214546

Danielsson, H., Humes, L. E., and Rönnberg, J. (2019). Different Associations between Auditory Function and Cognition Depending on Type of Auditory Function and Type of Cognition. *Ear. Hear.* 40, 1210–1219. doi: 10.1097/AUD.0000000000000700

Deal, J. A., Betz, J., Yaffe, K., Harris, T., Purchase-Helzner, E., Satterfield, S., et al. (2016). Hearing Impairment and Incident Dementia and Cognitive Decline in Older Adults: the Health ABC Study. *J. Gerontol. Ser. A Biol. Sci. Med. Sci.* 72:glw069. doi: 10.1093/gerona/glw069

Du Plessis, L., Paul, R. H., Hoare, J., Stein, D. J., Taylor, P. A., Meintjes, E. M., et al. (2017). Resting-state functional magnetic resonance imaging in clade C HIV: within-group association with neurocognitive function. *J. Neurovirol.* 23, 875–885. doi: 10.1007/s13365-017-0581-5

Ettenhofer, M. L., Foley, J., Castellon, S. A., and Hinkin, C. H. (2010). Reciprocal prediction of medication adherence and neurocognition in HIV/AIDS. *Neurology* 74, 1217–1222. doi: 10.1212/WNL.0b013e3181d8c1ca

Gates, G. A., Cobb, J. L., Linn, R. T., Rees, T., Wolf, P. A., and D'agostino, R. B. (1996). Central Auditory Dysfunction, Cognitive Dysfunction, and Dementia in Older People. *Arch. Otolaryngol. Head Neck Surg.* 122, 161–167. doi: 10.1001/archotol.1996.01890140047010

Glick, H. A., and Sharma, A. (2020). Cortical Neuroplasticity and Cognitive Function in Early-Stage, Mild-Moderate Hearing Loss: evidence of Neurocognitive Benefit From Hearing Aid Use. *Front. Neurosci.* 14:93. doi: 10.3389/fnins.2020.00093

Grigoryan, G., Fellows, A. M., Musiek, F., Chambers, R., Clavier, O. H., and Buckey, J. C. (2013). "Mathematical modeling of gap detection" in *Association for Research in Otolaryngology Midwinter Meeting*. (Baltimore: Association for Research in Otolaryngology).

Hallgren, M., Larsby, B., Lyxell, B., and Arlinger, S. (2001). Cognitive effects in dichotic speech testing in elderly persons. *Ear Hear.* 22, 120–129. doi: 10.1097/00003446-200104000-00005

Hammers, D., Spurgeon, E., Ryan, K., Persad, C., Barbas, N., Heidebrink, J., et al. (2012). Validity of a Brief Computerized Cognitive Screening Test in Dementia. *J. Geriatr. Psychiatry Neurol.* 25, 89–99. doi: 10.1177/0891988712447894

Hammers, D., Spurgeon, E., Ryan, K., Persad, C., Heidebrink, J., Barbas, N., et al. (2011). Reliability of repeated cognitive assessment of dementia using a brief computerized battery. *Am. J. Alzheimer. Dis. Other. Demen.* 26, 326–333. doi: 10.1177/1533317511411907

Harris, K. C., Eckert, M. A., Ahlstrom, J. B., and Dubno, J. R. (2010). Age-related differences in gap detection: effects of task difficulty and cognitive ability. *Hear. Res.* 264, 21–29. doi: 10.1016/j.heares.2009.09.017

Harris, K. C., Wilson, S., Eckert, M. A., and Dubno, J. R. (2012). Human Evoked Cortical Activity to Silent Gaps in Noise. *Ear Hear.* 33, 330–339. doi: 10.1097/AUD.0b013e31823fb585

Heaton, R. K., Franklin, D. R. Jr., Deutsch, R., Letendre, S., Ellis, R. J., Casaletto, K., et al. (2015). Neurocognitive change in the era of HIV combination antiretroviral therapy: the longitudinal CHARTER study. *Clin. Infect. Dis.* 60, 473–480. doi: 10.1093/cid/ciu862

Hoover, E. C., Souza, P. E., and Gallun, F. J. (2017). Auditory and Cognitive Factors Associated with Speech-in-Noise Complaints following Mild Traumatic Brain Injury. *J. Am. Acad. Audiol.* 28, 325–339. doi: 10.3766/jaaa.16051

Humes, L. E. (2015). Age-Related Changes in Cognitive and Sensory Processing: focus on Middle-Aged Adults. *Am. J. Audiol.* 24, 94–97. doi: 10.1044/2015_AJA-14-0063

Humes, L. E. (2020). Associations Between Measures of Auditory Function and Brief Assessments of Cognition. *Am. J. Audiol.* 29, 825–837. doi: 10.1044/2020_AJA-20-00077

Humes, L. E. (2021). Longitudinal Changes in Auditory and Cognitive Function in Middle-Aged and Older Adults. *J. Speech Lang. Hear. Res.* 64, 230–249. doi: 10.1044/2020_JSLHR-20-00274

Humes, L. E., Busey, T. A., Craig, J. C., and Kewley-Port, D. (2009). The effects of age on sensory thresholds and temporal gap detection in hearing, vision, and touch. *Atten. Percept. Psychophys.* 71, 860–871. doi: 10.3758/APP.71.4.860

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

Husain, F. T., Fromm, S. J., Pursley, R. H., Hosey, L. A., Braun, A. R., and Horwitz, B. (2006). Neural bases of categorization of simple speech and nonspeech sounds. *Hum. Brain Map.* 27, 636–651. doi: 10.1002/hbm.20207

Idrizbegovic, E., Hederstierna, C., Dahlquist, M., Kampfe Nordstrom, C., Jelic, V., and Rosenhall, U. (2011). Central auditory function in early Alzheimer's disease and in mild cognitive impairment. *Age Ageing* 40, 249–254. doi: 10.1093/ageing/afq168

Iliadou, V. V., Bamiou, D.-E., Sidiras, C., Moschopoulos, N. P., Tsolaki, M., Nimatoudis, I., et al. (2017). The Use of the Gaps-In-Noise Test as an Index of the Enhanced Left Temporal Cortical Thinning Associated with the Transition between Mild Cognitive Impairment and Alzheimer's Disease. *J. Am. Acad. Audiol.* 28, 463–471. doi: 10.3766/jaaa.16075

Jalaei, B., Valadbeigi, A., Panahi, R., Nahrani, M. H., Arefi, H. N., Zia, M., et al. (2019). Central Auditory Processing Tests as Diagnostic Tools for the Early Identification of Elderly Individuals with Mild Cognitive Impairment. *J. Audiol. Otol.* 23, 83–88. doi: 10.7874/jao.2018.00283

Japee, S., Holiday, K., Satyshur, M. D., Mukai, I., and Ungerleider, L. G. (2015). A role of right middle frontal gyrus in reorienting of attention: a case study. *Front. Syst. Neurosci.* 9:23. doi: 10.3389/fnsys.2015.00023

Kamminga, J., Bloch, M., Vincent, T., Carberry, A., Brew, B. J., and Cysique, L. A. (2017). Determining optimal impairment rating methodology for a new HIV-associated neurocognitive disorder screening procedure. *J. Clin. Exp. Neuropsychol.* 39, 753–767. doi: 10.1080/13803395.2016.1263282

Kotz, S. A., and Schwartze, M. (2010). Cortical speech processing unplugged: a timely subcortico-cortical framework. *Trends Cogn. Sci.* 14, 392–399. doi: 10.1016/j.tics.2010.06.005

Koziol, L. F., and Budding, D. E. (2009). *Subcortical Structures and Cognition : implications for Neuropsychological Assessment*. New York: Springer. doi: 10.1007/978-0-387-84868-6

Kuhn, T., Kaufmann, T., Doan, N. T., Westlye, L. T., Jones, J., Nunez, R. A., et al. (2018). An augmented aging process in brain white matter in HIV. *Hum. Brain Mapp.* 39, 2532–2540. doi: 10.1002/hbm.24019

Maro, I., Moshi, N., Clavier, O. H., Mackenzie, T. A., Kline-Schoder, R. J., Wilbur, J. C., et al. (2014). Auditory impairments in HIV-infected individuals in Tanzania. *Ear Hear.* 35, 306–317. doi: 10.1097/01.aud.0000439101.07257.ed

Meinke, D. K., Norris, J. A., Flynn, B. P., and Clavier, O. H. (2017). Going wireless and booth-less for hearing testing in industry. *Int. J. Audiol.* 56, 41–51. doi: 10.1080/14992027.2016.1261189

Moore, D. R., Edmondson-Jones, M., Dawes, P., Fortnum, H., Mccormack, A., Pierzycki, R. H., et al. (2014). Relation between Speech-in-Noise Threshold, Hearing Loss and Cognition from 40–69 Years of Age. *PLoS One* 9:e107720. doi: 10.1371/journal.pone.0107720

Nasreddine, Z. S., Phillips, N. A., BéDirian, V. R., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal Cognitive Assessment, MoCA: a Brief Screening Tool For Mild Cognitive Impairment. *J. Am. Geriatr. Soc.* 53, 695–699. doi: 10.1111/j.1532-5415.2005.53221.x

Niemczak, C., Fellows, A., Lichtenstein, J., White-Schwoch, T., Magohe, A., Gui, J., et al. (2021). Central Auditory Tests to Track Cognitive Function in People With HIV: longitudinal Cohort Study. *JMIR Form Res.* 5:e26406. doi: 10.2196/26406

Nilsson, L.-G. (2003). Memory function in normal aging. *Acta Neurol. Scand.* 107, 7–13. doi: 10.1034/j.1600-0404.107.s179.5.x

Overton, E. T., Kauwe, J. S., Paul, R., Tashima, K., Tate, D. F., Patel, P., et al. (2011). Performances on the CogState and standard neuropsychological batteries among HIV patients without dementia. *AIDS Behav.* 15, 1902–1909. doi: 10.1007/s10461-011-0033-9

Palmer, S. B., and Musiek, F. E. (2014). Electrophysiological gap detection thresholds: effects of age and comparison with a behavioral measure. *J. Am. Acad. Audiol.* 25, 999–1007. doi: 10.3766/jaaa.25.10.8

Panza, F., Quaranta, N., and Logroscino, G. (2018). Sensory Changes and the Hearing Loss–Cognition Link. *JAMA Otolaryngol. Head Neck Surg.* 144:127. doi: 10.1001/jamaoto.2017.2514

Parbery-Clark, A., Strait, D. L., Anderson, S., Hittner, E., and Kraus, N. (2011). Musical Experience and the Aging Auditory System: implications for Cognitive Abilities and Hearing Speech in Noise. *PLoS One* 6:e18082. doi: 10.1371/journal.pone.0018082

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., et al. (2016). Hearing Impairment and Cognitive Energy: the Framework for Understanding Effortful Listening (FUEL). *Ear Hear.* 37, 5S–27S. doi: 10.1097/AUD.0000000000000312

Pichora-Fuller, M. K., Schneider, B. A., and Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *J. Acoust. Soc. Am.* 97, 593–608. doi: 10.1121/1.412282

Robertson, K., Liner, J., and Heaton, R. (2009). Neuropsychological assessment of HIV-infected populations in international settings. *Neuropsychol. Rev.* 19, 232–249. doi: 10.1007/s11065-009-9096-z

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., et al. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Front. Syst. Neurosci.* 7:31. doi: 10.3389/fnsys.2013.00031

Rudner, M., and Lunner, T. (2014). Cognitive spare capacity and speech communication: a narrative overview. *Biomed. Res Int.* 2014:869726. doi: 10.1155/2014/869726

Ruel, T. D., Boivin, M. J., Boal, H. E., Bangirana, P., Charlebois, E., Havlir, D. V., et al. (2012). Neurocognitive and motor deficits in HIV-infected Ugandan children with high CD4 cell counts. *Clin. Infect. Dis.* 54, 1001–1009. doi: 10.1093/cid/cir1037

Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiol. Aging* 30, 507–514. doi: 10.1016/j.neurobiolaging.2008.09.023

Salthouse, T. A., and Ferrer-Caja, E. (2003). What needs to be explained to account for age-related effects on multiple cognitive variables? *Psychol. Aging* 18, 91–110. doi: 10.1037/0882-7974.18.1.91

Salthouse, T. A., and Meinz, E. J. (1995). Aging, inhibition, working memory, and speed. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 50, 297–306. doi: 10.1093/geronb/50B.6.P297

Sardone, R., Battista, P., Panza, F., Lozupone, M., Griseta, C., Castellana, F., et al. (2019). The Age-Related Central Auditory Processing Disorder: silent Impairment of the Cognitive Ear. *Front. Neurosci.* 13:619. doi: 10.3389/fnins.2019.00619

Saylor, D., Dickens, A. M., Sacktor, N., Haughey, N., Slusher, B., Pletnikov, M., et al. (2016). HIV-associated neurocognitive disorder - pathogenesis and prospects for treatment. *Nat. Rev. Neurol.* 12:309. doi: 10.1038/nrneurol.2016.27

Snell, K. B. (1997). Age-related changes in temporal gap detection. *J. Acoust. Soc. Am.* 101, 2214–2220. doi: 10.1121/1.418205

Song, F., Zhan, Y., Ford, J. C., Cai, D. C., Fellows, A. M., Shan, F., et al. (2020). Increased Right Frontal Brain Activity During the Mandarin Hearing-in-Noise Test. *Front. Neurosci.* 14:614012. doi: 10.3389/fnins.2020.614012

Specht, K. (2014). Neuronal basis of speech comprehension. *Hear. Res.* 307, 121–135. doi: 10.1016/j.heares.2013.09.011

Szymaszek, A., Sereda, M., Pöppel, E., and Szelag, E. (2009). Individual differences in the perception of temporal order: the effect of age and cognition. *Cogn. Neuropsychol.* 26, 135–147. doi: 10.1080/02643290802504742

Verhaeghen, P., and Salthouse, T. A. (1997). Meta-analyses of age-cognition relations in adulthood: estimates of linear and nonlinear age effects and structural models. *Psychol. Bull.* 122, 231–249. doi: 10.1037/0033-2909.122.3.231

Vissoci, J. R. N., De Oliveira, L. P., Gafaar, T., Haglund, M. M., Mvungi, M., Mmbaga, B. T., et al. (2019). Cross-cultural adaptation and psychometric properties of the MMSE and MoCA questionnaires in Tanzanian Swahili for a traumatic brain injury population. *BMC Neurol.* 19:57. doi: 10.1186/s12883-019-1283-9

Watson, B. U. (1991). Some Relationships between Intelligence and Auditory-Discrimination. *J. Speech Hear. Res.* 34, 621–627. doi: 10.1044/jshr.3403.621

White-Schwoch, T., Magohe, A. K., Fellows, A. M., Rieke, C. C., Vilarello, B., Nicol, T., et al. (2020). Auditory neurophysiology reveals central nervous system dysfunction in HIV-infected individuals. *Clin. Neurophysiol.* 131, 1827–1832. doi: 10.1016/j.clinph.2020.04.165

Wingfield, A., and Tun, P. A. (2007). Cognitive supports and cognitive constraints on comprehension of spoken language. *J. Am. Acad. Audiol.* 18, 548–558. doi: 10.3766/jaaa.18.7.3

Wong, P. C. M., Jin, J. X., Gunasekera, G. M., Abel, R., Lee, E. R., and Dhar, S. (2009). Aging and cortical mechanisms of speech perception in noise. *Neuropsychologia* 47, 693–703. doi: 10.1016/j.neuropsychologia.2008.11.032

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

Zhan, Y., Fellows, A. M., Qi, T., Clavier, O. H., Soli, S. D., Shi, X., et al. (2017b). Speech in Noise Perception as a Marker of Cognitive Impairment in HIV Infection. *Ear Hear.* 39:1. doi: 10.1097/AUD.0000000000000508

Zhan, Y., Buckey, J. C., Fellows, A. M., and Shi, Y. (2017a). Magnetic Resonance Imaging Evidence for Human Immunodeficiency Virus Effects on Central Auditory Processing: a Review. *J. Aids Clin. Res.* 8:708. doi: 10.4172/2155-6113.1000708

Zhan, Y., Fellows, A. M., Qi, T., Clavier, O. H., Soli, S. D., Shi, X., et al. (2018). Speech in Noise Perception as a Marker of Cognitive Impairment in HIV Infection. *Ear. Hear.* 39, 548–554.

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

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

frontiers in Psychology

BRIEF RESEARCH REPORT
published: 04 October 2021
doi: 10.3389/fpsyg.2021.729037

The Emotional Effect of Background Music on Selective Attention of Adults

Éva Nadon[1,2,3], Barbara Tillmann[4], Arnaud Saj[2,3†] and Nathalie Gosselin[1,2*†]*

[1] International Laboratory for Brain, Music and Sound Research (BRAMS), Music, Emotions, and Cognition Research Laboratory (MUSEC), Center for Research on Brain, Language and Music (CRBLM), Montreal, QC, Canada, [2] Department of Psychology, University of Montreal, Montreal, QC, Canada, [3] Center for Interdisciplinary Research in Rehabilitation of Metropolitain Montreal (CRIR), Montreal, QC, Canada, [4] Lyon Neuroscience Research Center, CNRS, UMR 5292, INSERM, U1028, University Lyon 1, Lyon, France

Daily activities can often be performed while listening to music, which could influence the ability to select relevant stimuli while ignoring distractors. Previous studies have established that the level of arousal of music (e.g., relaxing/stimulating) has the ability to modulate mood and affect the performance of cognitive tasks. The aim of this research was to explore the effect of relaxing and stimulating background music on selective attention. To this aim, 46 healthy adults performed a Stroop-type task in five different sound environments: relaxing music, stimulating music, relaxing music-matched noise, stimulating music-matched noise, and silence. Results showed that response times for incongruent and congruent trials as well as the Stroop interference effect were similar across conditions. Interestingly, results revealed a decreased error rate for congruent trials in the relaxing music condition as compared to the relaxing music-matched noise condition, and a similar tendency between relaxing music and stimulating music-matched noise. Taken together, the absence of difference between background music and silence conditions suggest that they have similar effects on adult's selective attention capacities, while noise seems to have a detrimental impact, particularly when the task is easier cognitively. In conclusion, the type of sound stimulation in the environment seems to be a factor that can affect cognitive tasks performance.

Keywords: selective attention, inhibition, Stroop task, background music, musical emotion, background noise, arousal, neuropsychology

OPEN ACCESS

Edited by:
Stephen Charles Van Hedger, Huron University College, Canada

Reviewed by:
Bruno Gingras, University of Vienna, Austria
McNeel Gordon Jantzen, Western Washington University, United States

***Correspondence:**
Arnaud Saj
Arnaud.saj@umontreal.ca
Nathalie Gosselin
Nathalie.gosselin@umontreal.ca

† These authors have contributed equally to this work

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

Received: *22 June 2021*
Accepted: *13 September 2021*
Published: *04 October 2021*

Citation:
Nadon É, Tillmann B, Saj A and Gosselin N (2021) The Emotional Effect of Background Music on Selective Attention of Adults. Front. Psychol. 12:729037. doi: 10.3389/fpsyg.2021.729037

INTRODUCTION

Music is considered among the most enjoyable and satisfying human activities (Dubé and Le Bel, 2003). The recent development of portable players with unlimited access to musical libraries means that people's access to music has never been greater than in the last decade (Krause et al., 2014). Adults listen to music for an average time of 17.8 h per week (International Federation of the Phonographic Industry, 2018). It is therefore possible to infer that most adults perform a large part of their daily tasks in the presence of background music (cooking, driving, working, studying, etc.). The efficient accomplishment of these tasks recruits the capacities of selective attention, also referred as attentional control; the cognitive ability to select, among a considerable load of information, relevant stimuli while inhibiting others (Murphy et al., 2016; Bater and Jordan, 2020).

Due to their front-line role in information processing, selective attention capacities represent the gateway to other executive and memory functions, the latter allowing us to adapt to the demands of daily life (Cohen, 2011; Nobre et al., 2014). According to the preceding definitions, the presence of inattention would cause a deleterious effect on overall cognitive performance due to the processing of information irrelevant to the accomplishment of a task, at the expense of relevant information (Baldwin, 2012). Therefore, with a growing body of research showing that the presence of background music influences cognitive functioning (for review, see Kämpfe et al., 2010), it is important to better understand the influence of background music on selective attention. Particularly, this would allow for the development of recommendations aiming to optimize efficient performance in everyday life.

Research investigating the effects of background music on selective attention performance has shown mixed results; sometimes showing neutral (Petrucelli, 1987; Cassidy and MacDonald, 2007; Wallace, 2010; Speer, 2011; Cloutier et al., 2020; Deng and Wu, 2020), beneficial (Amezcua et al., 2005; Darrow et al., 2006; Cassidy and MacDonald, 2007; Masataka and Perlovsky, 2013; Slevc et al., 2013; Fernandez et al., 2019), or deleterious (Masataka and Perlovsky, 2013; Slevc et al., 2013; Fernandez et al., 2019; Deng and Wu, 2020; Cloutier et al., 2020) effects on performance. However, multiple factors can influence this variability, such as the methodological limits observed within this literature. Several studies present small samples of adult participants, making it difficult to generalize the results to the general adult population ($\leq$24 adult participants; Amezcua et al., 2005; Speer, 2011; Giannouli, 2012; Fernandez et al., 2019; Cloutier et al., 2020). In addition, most of the time, non-auditory (e.g., silence) and auditory (e.g., noises with sound characteristics similar to those of music) control conditions were lacking (Darrow et al., 2006; Marchegiani and Fafoutis, 2019; Deng and Wu, 2020; Cloutier et al., 2020). There were also methodological limitations regarding the choice of the sound material used. For example, some studies have presented music with words (e.g., Darrow et al., 2006; Speer, 2011; Marchegiani and Fafoutis, 2019; Deng and Wu, 2020), which has generally resulted in a deleterious effect on performance. However, several studies have previously shown that the presence of speech or words in a sound environment tends to negatively affect cognitive performance in comparison with a speechless sound environment (e.g., Salamé and Baddeley, 1989; Szalma and Hancock, 2011). Therefore, the effect of language processing is confounded with the effect of background music in these studies.

Another element that could explain the variability between the results of previous studies is the lack of control over the emotional characteristics of the sound stimuli being utilized (as discussed in Kämpfe et al., 2010; Schellenberg, 2013). Indeed, different sound environments can induce different emotions. Particularly for musical stimuli, musical parameters, such as tempo, can be modulated to induce different musical emotions, like the level of arousal (Gabrielsson and Juslin, 2003); music with fast tempi are usually considered as stimulating, while music with slow tempi are considered as relaxing (Västjäll, 2001; Bigand et al., 2005; Vieillard et al., 2008). The emotional characteristics of a sound stimuli, like its level of arousal, are important to consider as studies have shown links between them and performance on cognitive tasks (e.g., spatial skills, Thompson et al., 2001; selective attention, Ghimire et al., 2019). Indeed, according to the arousal-mood hypothesis (Nantais and Schellenberg, 1999; Thompson et al., 2001, 2011; Husain et al., 2002; Schellenberg and Hallam, 2005; Schellenberg et al., 2008), cognitive performance can be promoted by sound stimulation, notably by increasing physiological activation and improving mood. Both music and noise can induce emotions (Hunter and Schellenberg, 2010), but there is a general agreement that music is efficient to induce positive emotions, and therefore it can be employed to positively modulate mood (Thompson et al., 2001). The previously cited research by Schellenberg and Weiss (2013) has shown that when participants listen to music that positively alters their mood before performing a cognitive task, like a stimulating and pleasant music, their performance in this cognitive task was improved. The arousal-mood hypothesis has been built on data that are based on listening to a stimulus before the accomplishment of a cognitive task.

The objective of this current study was to investigate the effect of the arousal level of background music on adults' selective attention. To do so, we compared the effect of stimulating and relaxing music on performance at a Stroop-type task, with two music-matched noise conditions (stimulating music-matched-noise and relaxing music-matched noise), and a silence condition. Based on the arousal-mood hypothesis, we hypothesized that the sound environment judged to be the most stimulating and pleasant—the stimulating music condition—would be the most beneficial environment to optimize cognitive performance.

MATERIALS AND METHODS

Participants

46 participants [27 females (58.7% of the sample), mean age: 25.57 years $\pm$ 4.33, mean years of education: 17.1 years $\pm$ 2.24]. All participants were native French speakers, had normal hearing (measured by a brief hearing test done with an audiometer AC40 Interacoustics; participants had pure tone thresholds under 40 dB SPL; World Health Organization, 1980) and normal or corrected-to-normal vision. None of the participants had color blindness or a history of neurological/psychiatric/neurodevelopmental disorders. None of them were taking drugs or medication that affected the central nervous system during the study. In addition, participants were excluded if they presented at least one of the following criteria: (i) music perception deficits (i.e., performance below 73% at the scale subtest, 70% at the off-beat subtest, and 68% at the out-of-key subtest of the online Montreal Battery of Evaluation of Amusia (MBEA; Peretz and Vuvan, 2017); (ii) presence of mood disorders, evaluated with the Beck Depression Inventory-II (BDI-II; Beck et al., 1996) and the Beck Anxiety Inventory (BAI; Beck et al., 1988) and (iii) musicians. Individuals were considered musicians if they completed equal to or more than 5 years of formal music lessons or were self-taught under that time frame in learning/practicing an instrument, and were practicing a musical instrument equal to or more than 2 h

per week (Zhang et al., 2020). The average number of years of musical training/practice of the participants (calculated by taking the number of years of formal music training added to the number of years of autodidactic learning or practicing an instrument) was 1.95 years ± 2. All participants gave their written informed consent in accordance with regulation of the local ethics committee at the University of Montreal.

Auditory Materials

The 16 auditory stimuli encompassed eight musical excerpts and eight acoustically music-matched noise stimuli. The eight musical stimuli (four highly pleasant and stimulating excerpts and four highly pleasant and relaxing excerpts) were selected from our lab database of 42 short instrumental classical music excerpts, all in major mode. The selection was made based on valence (i.e., 0 = very unpleasant—100 = very pleasant), arousal (i.e., 0 = very relaxing—100 = very stimulating) and familiarity judgments (i.e., 0 = unknown—100 = very familiar) obtained by 46 non-musicians who did not participate in the current study; using visual analog scales (Nadon et al., 2016; for more information see **Supplementary Material**). The original excerpts from our database consisted of the first 30 s of each piece of music. In order to be able to accumulate enough data for each musical excerpt during the experimental task, the excerpts for the current study were made up of the first 60 s of the same musical pieces. Using data from Nadon et al. (2016), independent-samples *t*-tests revealed that excerpts in the relaxing condition differ significantly from the ones in the stimulating condition in terms of arousal (respectively, $M = 11.73 \pm 11.1$, $M = 79.18 \pm 18.75$; $t_{(366)} = -42$, $p < 0.001$, $\eta^2 = 0.82$) and familiarity (respectively, $M = 44.35 \pm 36.72$, $M = 91.35, \pm 19.25$; $t_{(366)} = -15.38$, $p < 0.001$, $\eta^2 = 0.39$). No difference of valence was found between the relaxing and stimulating excerpts (respectively, $M = 80.61 \pm 18.72$, $M = 78.43 \pm 21.11$; $t_{(366)} = -1.1$, $p = 0.3$, $\eta^2 = 0.01$ (see **Supplementary Material** for more information).

For auditory control conditions, acoustically music-matched noises were created based on the signal-processing procedure used in previous research (Zatorre et al., 1994; Blood et al., 1999). The spectral envelope of each music stimulus was exported and applied to a synthesized white noise. This generated "noise melody" was thus different for each matched music stimuli. To ensure that participants would not recognize the rhythmic patterns from the matched music piece while listening to the music-matched noise stimulus, each noise stimulus was played in reverse to create the final music-matched noise stimulus.

The final 16 stimuli (i.e., eight musical excerpts and eight acoustically music-matched noise excerpts) were normalized in amplitude, had a duration of 60 s, with 1 s ms fade-in and 2 secfade-out. All above sound processing was performed using Adobe Audition 3.0 software (Adobe Systems Inc., San Jose, CA, United States).

Experimental Stroop Task

Participants performed the task in a soundproof room. Visual information was displayed on a computer monitor at a distance of 60 cm, while auditory information was presented binaurally using headphones (DT770 Pro, Beyerdynamic) at a sound level ranging approximately around 60 decibels. This decision was made based on the results of Thompson et al. (2011) in which music demonstrated a deleterious effect on reading comprehension performance when presented at around 72 decibels, mainly for fast tempo music, compared to when presented at 60 decibels. Based on these results, 60 decibels appears to be the ideal sound level to perform a cognitive task simultaneously. Participants had access to a keyboard and a mouse, all of which were connected to the computer (HP ProDesk 600 G1, Windows 7) located outside the room, on which the task was run. Communication between inside and outside the soundproof room was done using microphones.

Selective attention was measured using a computerized Stroop-type task (Stroop, 1935; customized scripted and inspired by the Double trouble task from the *Cambridge Brain Sciences team*[1]). Each trial presented a target word (RED or GREEN) that appeared above two response words (RED and GREEN, see **Figure 1**). The color of the target word was either congruent (e.g., the word RED presented in red ink) or incongruent (e.g., the word RED presented in green ink) to the meaning of the word. To add a level of difficulty, when the trial was incongruent, the ink color in which the response words were presented was also incongruent. Participants therefore had to identify the ink color of the target word by selecting (with the keyboard arrows left and right) the correct response word presented underneath. The presentation of stimuli, and the recording of the type of stimuli presented, response time and accuracy, were carried out using the Psychtoolbox-3.0.13 (developed by Matlab and GNU Octave) implemented in Matlab 2015b (Mathworks Inc., Natick, MA, United States).

Participants were instructed to perform Stroop trials while being as fast and accurate as possible. Each Stroop trial consisted of a fixation cross (presented 500 ms, see **Figure 1**) followed by one of the eight possible color-word stimulus options, presented in a pseudo-randomized order. Participants had a maximum of 2,000 ms to give their answer. If participants answered before this given time, another trial began and so on. Past this time, the trial ended, a missed trial was recorded, the words "Too late" appeared on the screen (for 400 ms), and the next trial began.

Procedure

Participants practiced performing the task in three blocks of 30 s, with a possibility to take a break between the blocks in order to clarify instructions if needed. Each block was performed, respectively, in silence; accompanied by a music stimulus previously judged to have a neutral level of activation and high level of valence (see **Supplementary Table 1**; Nadon et al., 2016), and with the matched noise stimulus. Practice blocks were similar to experimental blocks, except that the participants responded to Stroop trials for only 16 s. During these practice blocks, participants received feedback for their answers (correct/incorrect; for 800 ms). After completing the three practice blocks, participants could choose to receive the instructions specific to the experimental part, or to continue practicing (by performing all three blocks again).

[1] www.cambridgebrainsciences.com/tests/double-trouble

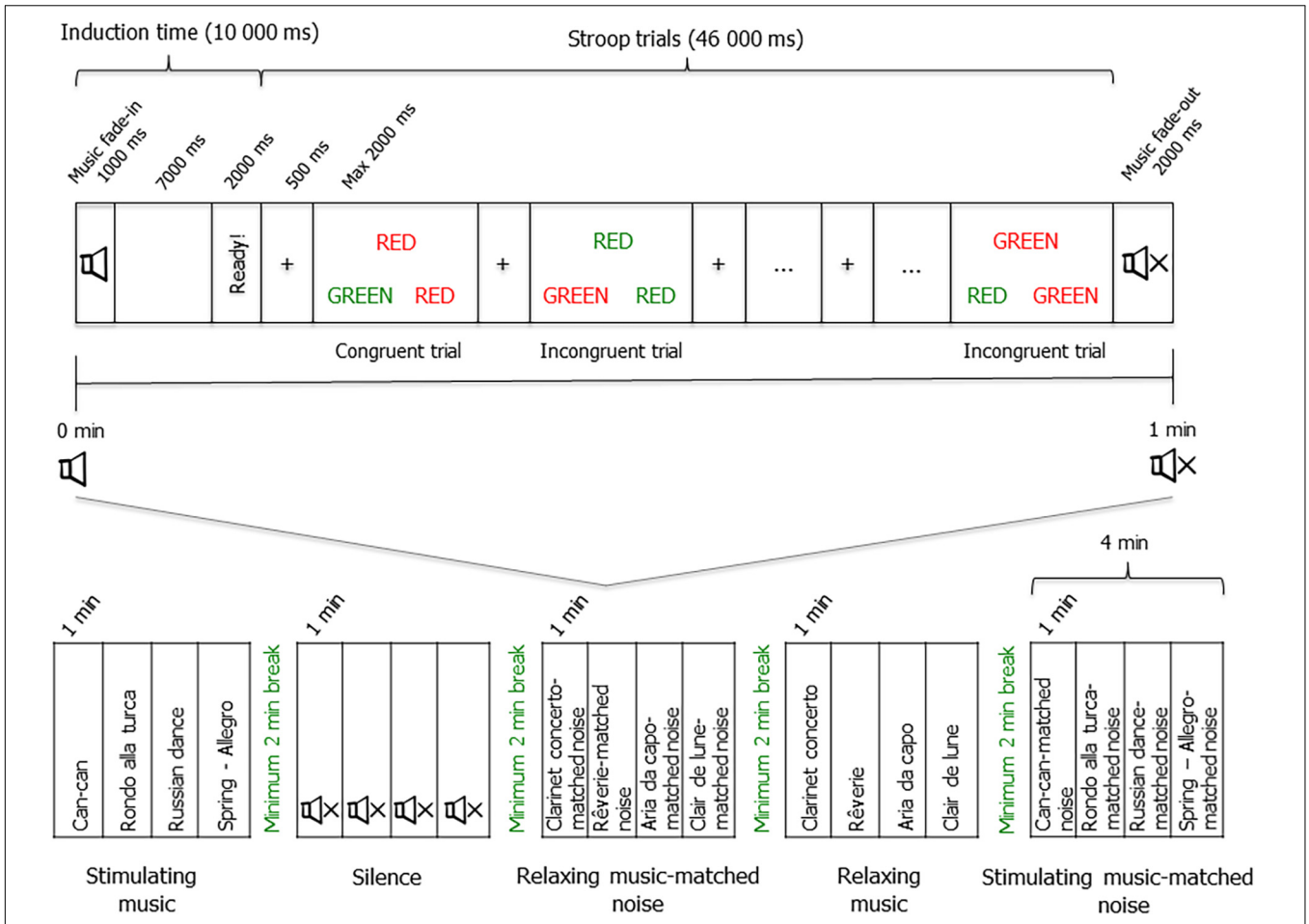

FIGURE 1 | Schematic representation of the experimental procedure and experimental Stroop block.

For the experimental testing, participants performed the Stroop task in five sound conditions: Silence (S), relaxing music (RM), relaxing music-matched noise (RMN), stimulating music (SM), and stimulating music-matched noise (SMN; see **Figure 1**). The order in which participants performed the sound conditions was counterbalanced across participants and the order of presentation of musical or noise stimuli inside the same sound condition was randomized across participants using the Matlab script. Each sound condition consists of four consecutive blocks of 60,000 ms. Each block began with an induction phase (for 8,000 ms) presenting a blank screen while the participant either listened to the music or noise played, or remained in silence, depending on the sound condition that was performed (see **Figure 1**). Then, the word "Ready!" was presented (for 2,000 ms), followed by the beginning of a 46,000 ms sequence of Stroop task trials. Participants therefore performed their last Stroop trial just before the sound fade-out, when applicable. When participants completed a sound condition (total of 4 min), they had to take a break of at least 2 min, during which they left the soundproof room to fill out the questionnaires, until they were asked to return to the room to perform the next condition.

After completing the task, participants were asked to listen to each auditory stimulus they heard during the task. Stimuli were presented in a randomized order and visual analog scales were showed on the screen. Participants were asked to evaluate the level of arousal [very relaxing (0) to very stimulating (100)], valence [very unpleasant (0) to very pleasant (100)], and familiarity [unknown (0) to very familiar (100)] for each auditory stimulus.

Data Analyses

Accuracy (error rate (ER); percentage of incorrect responses excluding missed trials) and mean response times (RT) of correct responses trials were computed for each participant, for each sound condition (i.e., RM, NRM, SM, NSM, and S) and Stroop congruence trial type (i.e., congruent and incongruent). A trial was considered correct when the participant was able to accurately identify the ink color of the target word within the imposed time limit (2,000 ms). Of these correct trials, a first mean and standard deviation were calculated, and only RT between −1.97 and 1.97 standard deviation from the participant's mean were used to calculate mean RT. The Stroop interference effect was calculated by subtracting mean

RT of congruent from incongruent conditions (i.e., mean RT incong.—mean RT cong.) for each sound condition. Mean ER and RT were entered into separate repeated measures analyses of variance (ANOVAs) with Sound Conditions and Stroop Congruence trial type as within-subject factors. Mean Stroop interference scores were entered into another repeated measures ANOVA with Sound Conditions as within-subject factor. When interactions or a principal effect were significant, *t*-test analysis were performed.

Paired-sample *t*-tests were performed to evaluate differences between judgments of arousal, valence and familiarity for the musical stimuli and the music-matched noise stimuli (see **Supplementary Table 2**). All data were analyzed using IBM SPSS Statistics 26 (IBM Corp., 2019). The alpha levels were set at.05 for all analyses.

RESULTS

Auditory Stimuli Evaluation

As expected, judgments of arousal from participants were significantly higher for the stimulating music (SM) compared to the relaxing music (RM). The arousal was judged significantly higher for the two noise conditions (RMN and SMN) compared to the RM. Similarly, the arousal was judged significantly lower for the two noise conditions (RMN and SMN) than for the SM. SMN was considered significantly more stimulating than RMN (see **Figure 2** and **Supplementary Table 3** for details on arousal's results).

For the evaluation of valence, as expected, participants considered that the two music conditions (RM and SM) were significantly more pleasant than the two noise conditions (RMN and SMN). There was no significant difference between the two music conditions (RM and SM) and the two noise conditions (RMN and SMN) in terms of valence (see **Figure 2** and **Supplementary Table 3** for details on valence's results).

The most familiar condition was SM, followed by the RM condition, with the other two noise conditions being significantly less familiar (RMN and SMN). There was no difference between the level of familiarity among the two noise conditions (RMN and SMN; see **Figure 2** and **Supplementary Table 3** for more details).

Stroop Task

The correct response time (RT) and error rate (ER) analyzes supported the observation of a Stroop interference effect as RTs were significantly slower and ERs were higher on incongruent trials compared to congruent trials [effect of congruence on RTs: $F_{(1,\ 45)} = 253.93$, $p < 0.005$, $\eta^2 = 0.85$; effect of congruence on error rate: $F_{(1,\ 45)} = 104.158$, $p < 0.005$, $\eta^2 = 0.70$]. In terms of RT on incongruent and congruent trials, there were no significant differences in performance between the different sound conditions [$F_{(1,\ 45)} = 1.01$, $p = 0.405$, $\eta^2 = 0.02$]. In the analysis of ERs for incongruent and congruent trials in each sound condition, the ER for congruent trials in the RMN condition was significantly higher than in the RM condition [$t_{(45)} = 2.10$, $p < 0.05$, $\eta^2 = 0.09$]. A similar tendency is noted between the ER for congruent trials in the SMN condition compared to the ER in RM condition [$t_{(45)} = 1.81$, $p = 0.077$, $\eta^2 = 0.07$]. Regarding the ER for incongruent trials, there was a trend toward a higher ER in the SM condition compared to the silence condition [$t_{(45)} = -1.69$, $p = 0.097$, $\eta^2 = 0.06$, see **Figure 3** and **Supplementary Table 4** for more details on Stroop's task results]. No significant effect was found in the analysis with the mean Stroop interference effect scores for each sound condition [$F_{(1,\ 45)} = 0.394$, $p = 0.813$, $\eta^2 = 0.009$].

DISCUSSION

The aim of this study was to investigate the effect of the arousal level of background music on selective attention of adults. The results there did not reveal any significant differences in attentional performance depending on whether the task was performed in silence, accompanied by relaxing music or stimulating music. Even though the results showed that participants tended to make a greater number of errors when listening to stimulating music compared to silence, this difference was not significant. However, when comparing on-task performance in the presence of music or noise, performance is more affected by the presence of noise given that there is a significant difference in error rate for congruent trials between relaxing music (RM) and relaxing music-matched noise (RMN), and a trend between relaxing music and relaxing music-matched noise (SMN).

These results are somewhat encouraging as they showed that the addition of to-be processed cognitive information (e.g., background music/noise) does not necessarily have deleterious effects on attentional performance as some theories suggests (e.g., Kahneman (1973) limited capacity model). With these results, it is possible to assume that performing a task requiring attention in the presence of instrumental music should not have a negative effect on the level of selective attention demand in order to perform the task optimally.

The arousal-mood hypothesis (Nantais and Schellenberg, 1999; Thompson et al., 2001, 2011; Husain et al., 2002; Schellenberg and Hallam, 2005; Schellenberg et al., 2008) suggests that a sound environment judged to be stimulating and pleasant would be a beneficial environment to optimize cognitive performance (for details, see Schellenberg, 2013). It was therefore expected that the stimulating music condition would be the sound environment in which we would see the lowest error rate and weakest Stroop interference. In contrast to the hypotheses, the presence of pleasant and stimulating music during the accomplishment of the task did not significantly improve task performance. A small tendency to make more errors on incongruent trials in this sound environment was also noted. These results differ from those of previous work studying the effect of the arousal level of background music upon selective attention (Fernandez et al., 2019; Cloutier et al., 2020). However, these studies mainly aimed to make comparisons between groups (elderly vs. young adults), while the present study had an objective of generalization to the adult population. Furthermore, the tasks involved were different: while previous studies employed the Flanker task to assess selective visual

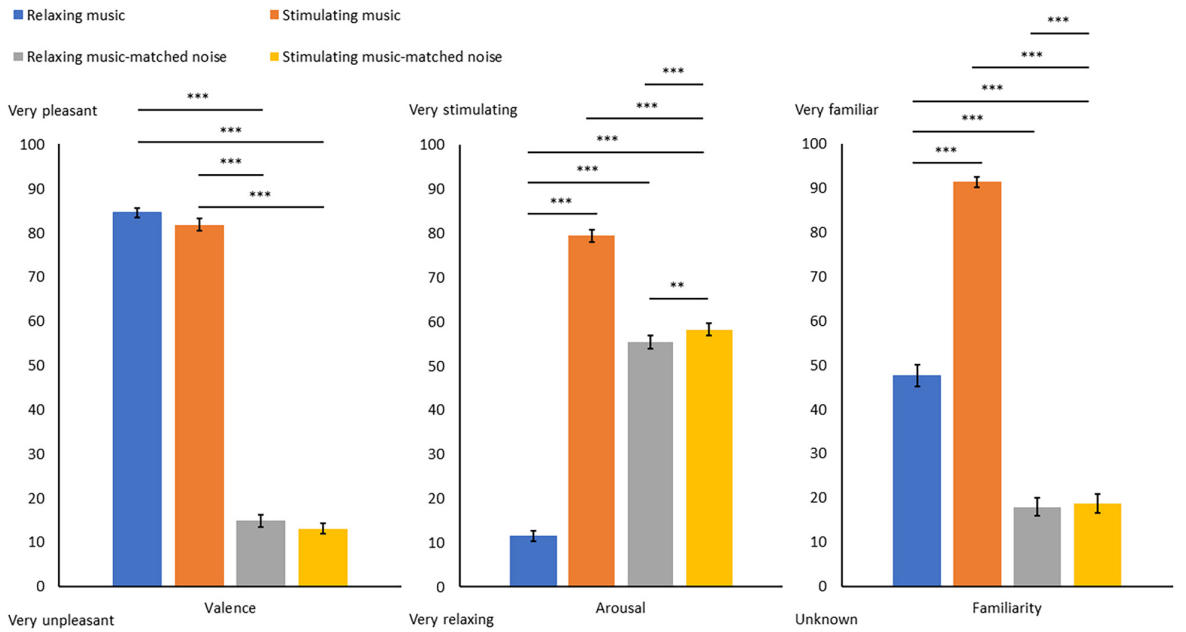

FIGURE 2 | Mean scores for the emotional judgments of valence and arousal and mean scores for the evaluation of familiarity for each sound condition. Graph shows standard errors. $^{**}p < 0.01$. $^{***}p < 0.001$.

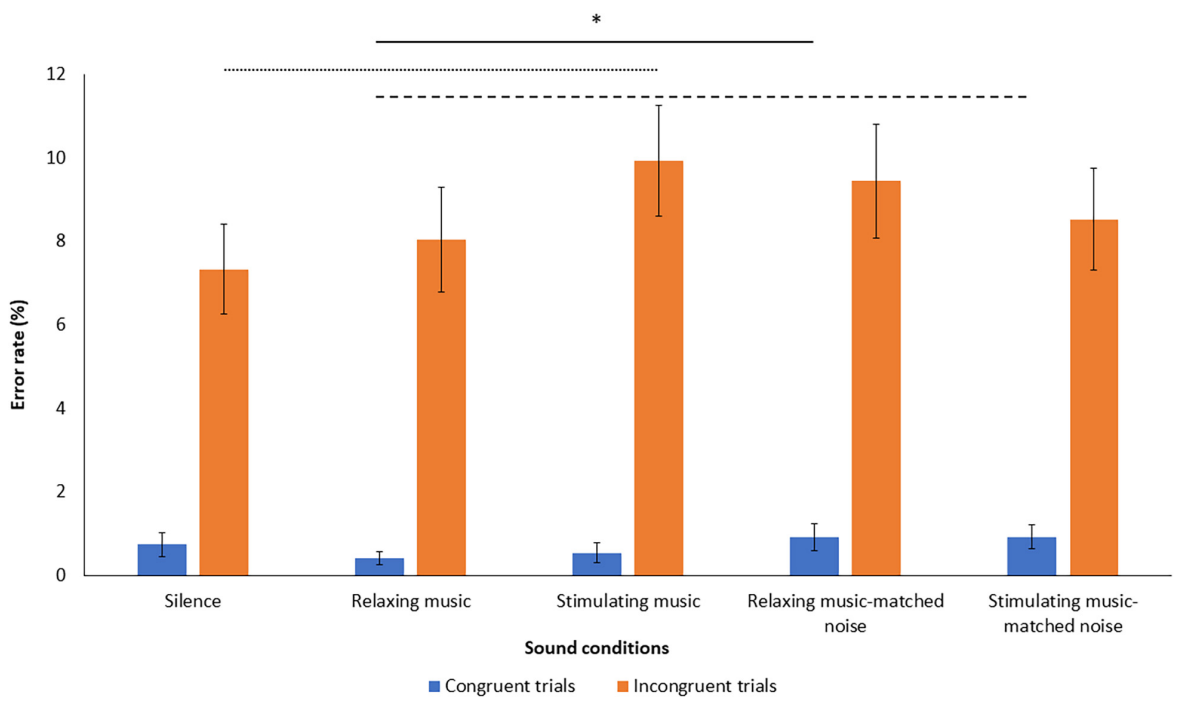

FIGURE 3 | Mean error rate on incongruent and congruent trials for each sound condition. Graph shows standard errors. $^{*}p < 0.05$. - - - $p = 0.07$. . . . . $p = 0.09$.

attention, the current study utilized the Stroop task which involves language processing. On the other hand, the number of sound conditions in this study may affect the statistical power of the results. It would therefore be interesting to investigate whether the results would be the same with even a larger sample-size in future studies (even though our sample-size was larger than in previous studies).

A key finding of this study is a negative effect of music-matched noise stimuli (low pleasantness) on attentional performance. These results converge with previous work by

Masataka and Perlovsky (2013) and Slevc et al. (2013) showing lower performance on a similar Stroop task in the presence of dissonant music (sound pairings perceived as generally unpleasant or possessing low-pleasantness valence). Interestingly in these studies, greater consonance (sound pairings perceived as generally pleasant or possessing a high-pleasantness valence) led to better performance on the Stroop task. It would therefore be interesting to investigate further to assess which factor, the level of valence/pleasantness or the degree of consonance, had the greater influence upon the results of this and previous studies (Masataka and Perlovsky, 2013; Slevc et al., 2013). This could be done by integrating stimuli that are both consonant and unpleasant, such as scary or sad music, or by specially composed music material. In previous research, the relationship between background music and cognitive performance seems to be affected by the degree of familiarity of the musical stimulus (if the music was already known to the participant). Higher familiarity has a positive effect on performance for cognitive tasks (Darrow et al., 2006; Speer, 2011; Giannouli, 2012). One potential limitation of this study is that, despite an attempt to select equally familiar music of similar valence, the stimulating musical stimuli were rated as more familiar by the participants than the relaxing musical stimuli (see **Supplementary Materials** for details). It is then surprising that the present findings did not support an effect of stimulating music on task performance given that the stimulating music condition was biased toward higher familiarity.

Judgments of valence can be influenced by the familiarity of a musical piece. Some studies have shown that perceivers tend to find a stimulus that they already know (e.g., a piece of music) more pleasant (Parente, 1976; Schellenberg et al., 2008; Van Den Bosch et al., 2013). Familiar background music has also been associated with increased pleasure in the process of completing a task without compromising task performance (Pereira et al., 2011; Feng and Bidelman, 2015). In this regard, Darrow et al. (2006); Speer (2011), Giannouli (2012), and Kiss and Linnell (2020) asked their participants to bring their favorite music into the lab, which then was used as background music to perform a selective attention task or a sustained-attention task. In these studies, the music selected by the participants held characteristics of high emotional valence and familiarity. However, the other characteristics of the music utilized were heterogeneous between participants (e.g., style, complexity of the music pieces, presence of lyrics, or the level of arousal). The results of these studies indicate that participants consistently performed better in the familiar music conditions. As we know little about the characteristics of the different pieces of music used in these studies and that a great variability is present between them, it is difficult to identify whether the results are generalizable to listening to background music in general or whether they are specifically attributable to a modulation of mood and/or arousal due to the emotional characteristics and familiarity of the music used. Future research should combine this approach with systematic acoustic as well as musical and linguistic structure analyses of the used material to further our understanding of the potential characteristics involved in the observed effects.

Taken together, our findings suggest that it is not sufficient for background music to be arousing, pleasant and familiar in order to enhance attentional performance as suggested by the Arousal-mood theory, and that factors related to individual musical taste may be driving the effects found in previous studies.

Finally, based on the results of this study, we can therefore recommend that tasks requiring selective attention can be performed in an environment of silence as well as with pleasant instrumental music. Findings from this study can be extended to practical use in environments with loud or unpleasant intermittent noises (for example open-plan offices or when space for telework must be shared). According to Szalma and Hancock (2011), intermittent unpredictable short noise bursts are the most disturbing forms of noise; these could be the sound of a horn outside, the laughter of a colleague in a nearby open-plan office, a family member shutting a door nearby, etc. Listening to music in the background may be an efficient tool, equal to working in silence, for masking unpleasant intermittent noises while maintaining a similar level of selective attention on a given task. In this light, future work comparing the presence of music with pleasant noises (such as waves or waterfall noises) would be interesting to investigate given their potential for masking intermittent noises.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Comité d'éthique de la recherche en arts et en sciences (CÉRAS), University of Montreal, Montreal, QC, Canada. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ÉN, BT, and NG contributed equally to the project's conception and the study design. ÉN performed the literature search. ÉN and AS drafted the manuscript and performed the statistical analysis. BT, NG, and ÉN contributed to the establishment of the Stroop task parameters. AS, NG, and BT provided the critical revisions. All authors contributed to the article and approved the submitted version.

FUNDING

The study was funded by the Fonds de recherche du Québec société et culture (FRQSC) -Établissement de nouveaux professeurs-chercheurs grant 2016-NG-189171 to NG; and 2019-BIZ-256735, the Fonds de recherche du Québec en santé

(FRQS), 2018-35294, and the Center for Research on Brain, Language and Music (CRBLM). The team Auditory Cognition and Psychoacoustics is part of the LabEx CeLva (Centre Lyonnais d'Acoustique, ANR-10-LABX-60). The CRBLM was funded by the Fonds de recherche du Québec nature et technologie (FRQNT) and by the FRQSC. A special thanks to the Erasmus Auditory Neuroscience Network for having favored the collaboration between NG and BT to set up the idea of this project.

ACKNOWLEDGMENTS

We would like to thank the research assistants who contributed to the data collection: Maude Picard and Idir Saggadi. We also want to thank Isabelle Peretz, for her ideas and support that contributed to the project's conception and design. We would like to thank the International Laboratory for Brain, Music and Sound Research (BRAMS) employees and students for their constant support, especially Bernard Bouchard, Eva Best, Johanne David, Luis Alexander Nieva Chavez, Natalia Fernandez, Simon Rigoulot, and Mihaela Felezeu.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.729037/full#supplementary-material

REFERENCES

Amezcua, C., Guevara, M. A., and Ramos-Loyo, J. (2005). Effects of musical tempi on visual attention ERPs. *Internat. J. Neurosci.* 115, 193–206. doi: 10.1080/00207450590519094

Baldwin, C. L. (2012). *Auditory cognition and human performance: Research and applications*. Boca Raton: CRC Press.

Bater, L. R., and Jordan, S. S. (2020). "Selective attention," in *Encyclopedia of Personality and Individual Differences*, eds V. Zeigler-Hill and T. K. Shackelford (Cham: Springer), doi: 10.1007/978-3-319-24612-3

Beck, A. T., Steer, R. A., and Brown, G. (1996). *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.

Beck, A. T., Epstein, N., Brown, G., and Steer, R. A. (1988). An Inventory for Measuring Clinical Anxiety: Psychometric properties. *J. Consult. Clin. Psychol.* 56, 893–897. doi: 10.1037/0022-006X.56.6.893

Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., and Dacquet, A. (2005). Multidimensional scaling of emotional response to music: The effect of musical expertise and of the duration of the excerpts. *Cogn. Emot.* 19, 1113–1139. doi: 10.1080/02699930500204250

Blood, A. J., Zatorre, R. J., Bermudez, P., and Evans, A. C. (1999). Emotional responses to pleasant and unpleasant music correlate with activity in paralimbic brain regions. *Nat. Neurosci.* 2, 382–387. doi: 10.1038/7299

Van Den Bosch, I., Salimpoor, V., and Zatorre, R. J. (2013). Familiarity mediates the relationship between emotional arousal and pleasure during music listening. *Front. Hum. Neurosci.* 7:534. doi: 10.3389/fnhum.2013.00534

Cassidy, G., and MacDonald, R. A. (2007). The effect of background music and background noise on the task performance of introverts and extraverts. *Psychol. Music* 35, 517–537. doi: 10.1177/0305735607076444

Cloutier, A., Fernandez, N. B., Houde-Archambault, C., and Gosselin, N. (2020). Effect of background music on attentional control in older and young adults. *Front. Psychol.* 11:2694. doi: 10.3389/fpsyg.2020.557225

Cohen, R. A. (2011). "Attention," in *Encyclopedia of clinical Neuropsychology*, eds J. S. Kreutzer, J. DeLuca, and B. Caplan (New Yrok, NY: Springer), doi: 10.1007/978-0-387-79948-3_1267

Darrow, A.-A., Johnson, C., Agnew, S., Fuller, E. R., and Uchisaka, M. (2006). Effect of preferred music as a distraction on music majors' and nonmusic majors' selective attention. *Bulletin of the Council for Research in Music Education* 170, 21–31.

Deng, M., and Wu, F. (2020). Impact of background music on reaction test and visual pursuit test performance of introverts and extraverts. *Internat. J. Industr. Erg.* 78:102976. doi: 10.1016/j.ergon.2020.102976

Dubé, L., and Le Bel, J. L. (2003). The content and structure of laypeople's concept of pleasure. *Cogn. Emot.* 17, 263–295. doi: 10.1080/02699930302295

Feng, S., and Bidelman, G. (2015). Music familiarity modulates mind wandering during lexical processing. *CogSci* 2015:1125.

Fernandez, N., Trost, W. J., and Vuilleumier, P. (2019). Brain networks mediating the influence of background music on selective attention. *Soc. Cogn. Affect. Neurosci.* 14, 1441–1452. doi: 10.1093/scan/nsaa004

Gabrielsson, A., and Juslin, P. N. (2003). *Emotional expression in music*. Oxford: Oxford University Press.

Ghimire, N., Yadav, R., and Mukhopadhyay, S. (2019). Comparative study of heart rate, blood pressure and selective attention of subjects before and after music. *Birat. J. Health Sci.* 4, 625–628. doi: 10.3126/bjhs.v4i1.23935

Giannouli, V. (2012). Attention and Music. Paper presented at the Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music. Available online at: http://icmpc-escom2012.web.auth.gr/files/papers/345_proc.pdf (accessed March 27, 2021).

Hunter, P. G., and Schellenberg, E. G. (2010). "Music and emotion," in *Music Perception*, eds M. R. Jones, R. R. Fay, and A. N. Popper (New York, NY: Springer), 129–164. doi: 10.1007/978-1-4419-6114-3_5

Husain, G., Thompson, W. F., and Schellenberg, E. G. (2002). Effects of musical tempo and mode on arousal, mood, and spatial abilities. *Music Percept.* 20, 151–171. doi: 10.1525/mp.2002.20.2.151

IBM Corp. (2019). *IBM SPSS Statistics for Windows, Version 26.0*. Armonk, NY: IBM Corp.

International Federation of the Phonographic Industry (2018). *Music consumer insight report*. London: International Federation of the Phonographic Industry.

Kahneman, D. (1973). *Attention and effort*, Vol. 1063. Englewood Cliffs, NJ: Prentice-Hall, 218–226.

Kämpfe, J., Sedlmeier, P., and Renkewitz, F. (2010). The impact of background music on adult listeners: A meta-analysis. *Psychol. Music* 39, 424–448. doi: 10.1177/0305735610376261

Kiss, L., and Linnell, K. J. (2020). The effect of preferred background music on task-focus in sustained attention. *Psycholog. Res.* 2020, 1–13. doi: 10.1007/s00426-020-01400-6

Krause, A., North, A., and Hewitt, L. (2014). Music selection behaviors in everyday listening. *J. Broadcast. Electr. Media* 58, 306–323. doi: 10.1080/08838151.2014.906437

Parente, J. A. (1976). Music preference as a factor of music distraction. *Percept. Motor Skills* 43, 337–338. doi: 10.2466/pms.1976.43.1.337

Pereira, C. S., Teixeira, J., Figueirodo, P., Xavier, J., Castro, S. L., and Brattico, E. (2011). Music and emotions in the brain: Familiarity matters. *PLoS One* 6:e27241. doi: 10.1371/journal.pone.0027241

Peretz, I., and Vuvan, D. T. (2017). Prevalence of congenital amusia. *Eur. J. Hum. Genet.* 25, 625–630. doi: 10.1038/ejhg.2017.15

Petrucelli, J. (1987). *The effects of noise and stimulative and sedative music on Stroop performance variables*. Ph.D. dissertation, Jamaica, NY: St John's University, NY.

Marchegiani, L., and Fafoutis, X. (2019). Word spotting in background music: A behavioural study. *Cogn. Comput.* 11, 711–718. doi: 10.1007/s12559-019-09649-9

Masataka, N., and Perlovsky, L. (2013). Cognitive interference can be mitigated by consonant music and facilitated by dissonant music. *Sci. Rep.* 3:2028. doi: 10.1038/srep02028

Murphy, G., Groeger, J. A., and Greene, C. M. (2016). Twenty years of load theory – Where are we now, and where should we go next? *Psychon. Bull. Rev.* 23, 1316–1340. doi: 10.3758/s13423-015-0982-5

Nadon, E., Cournoyer Lemaire, E., Bouvier, J., and Gosselin, N. (2016). "The effect of musical expertise in the emotional judgment of music," in *Poster presented at 26th Annual Meeting from the Canadian Society for Brain, Behaviour and Cognitive Science (CSBBCS)*, (Ottawa).

Nantais, K. M., and Schellenberg, E. G. (1999). The Mozart effect: An artifact of preference. *Psycholog. Sci.* 10, 370–373. doi: 10.1111/1467-9280.00170

Nobre, K., Nobre, A. C., and Kastner, S. (eds) (2014). *The Oxford handbook of attention*. Oxford: Oxford University Press. doi: 10.1093/oxfordhb/9780199675111.001.0001

Salamé, P., and Baddeley, A. (1989). Effects of background music on phonological short-term memory. *Q. J. Exp. Psychol.* 41, 107–122. doi: 10.1080/14640748908402355

Schellenberg, E. G., and Hallam, S. (2005). Music listening and cognitive abilities in 10- and 11-year-olds: The Blur effect. *Ann. N Y Acad. Sci.* 1060, 202–209. doi: 10.1196/annals.1360.013

Schellenberg, E. G., Peretz, I., and Vieillard, S. (2008). Liking for happy-and sad-sounding music: Effects of exposure. *Cogn. Emot.* 22, 218–237. doi: 10.1080/02699930701350753

Schellenberg, E. G. (2013). Music and cognitive abilities. *Curr. Direct. Psycholog. Sci.* 14, 317–320. doi: 10.1111/j.0963-7214.2005.00389.x

Slevc, L., Reitman, J., and Okada, B. (2013). "Syntax in music and language: the role of cognitive control," in *Proceedings of the Annual Meeting of the Cognitive Science Society* (Austin, TX: Cognitive Science Society), 3414–3419.

Speer, S. (2011). *Effect of background music, speech and silence on office workers' selective attention*, Master's thesis, Tallahassee, FL: The Florida State University, USA.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *J. Exp. Psychol.* 18, 643–662. doi: 10.1037/h0054651

Szalma, J. L., and Hancock, P. A. (2011). Noise effects on human performance: a meta-analytic synthesis. *Psychol. Bull.* 137:682. doi: 10.1037/a0023987

Thompson, W. F., Schellenberg, E. G., and Husain, G. (2001). Arousal, mood, and the Mozart effect. *Psycholog. Sci.* 12, 248–251. doi: 10.1111/1467-9280.00345

Thompson, W., Schellenberg, G., and Letnic, A. (2011). Fast and loud background music disrupt reading comprehension. *Psychol. Music* 40, 700–708. doi: 10.1177/0305735611400173

Västjäll, D. (2001). Emotion induction through music: a review of the musical mood induction procedure. *Musicae Sci.* 5, 173–211. doi: 10.1177/10298649020050S107

Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., and Bouchard, B. (2008). Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cogn. Emot.* 22, 720–752. doi: 10.1080/02699930701503567

Wallace, M. (2010). *The influence of acoustic background on visual Stroop task performance*. Master's thesis, Winnipeg: University of Manitoba.

World Health Organization (1980). *International classification of impairments, disabilities and handicaps (ICIDH): a manual of classification relating to the consequences of diseases*. Geneva: World Health Organization.

Zatorre, R. J., Evans, A. C., and Meyer, E. (1994). Neural mechanisms underlying melodic perception and memory for pitch. *J. Neurosci.* 14, 1908–1919. doi: 10.1523/JNEUROSCI.14-04-01908.1994

Zhang, J. D., Susino, M., McPherson, G. E., and Schubert, E. (2020). The definition of a musician in music psychology: a literature review and the six-year rule. *Psychol. Music* 48, 389–409. doi: 10.1177/0305735618804038

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

frontiers in Psychology

ORIGINAL RESEARCH
published: 15 November 2021
doi: 10.3389/fpsyg.2021.740363

Looming Effects on Attentional Modulation of Prepulse Inhibition Paradigm

Zhemeng Wu, Xiaohan Bao, Lei Liu and Liang Li**

School of Psychological and Cognitive Sciences, Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China

In a hazardous environment, it is fundamentally important to successfully evaluate the motion of sounds. Previous studies demonstrated "auditory looming bias" in both macaques and humans, as looming sounds that increased in intensity were processed preferentially by the brain. In this study on rats, we used a prepulse inhibition (PPI) of the acoustic startle response paradigm to investigate whether auditory looming sound with intrinsic warning value could draw attention of the animals and dampen the startle reflex caused by the startling noise. We showed looming sound with a duration of 120 ms enhanced PPI compared with receding sound with the same duration; however, when both sound types were at shorter duration/higher change rate (i.e., 30 ms) or longer duration/lower rate (i.e., more than 160 ms), there was no PPI difference. This indicates that looming sound–induced PPI enhancement was duration dependent. We further showed that isolation rearing impaired the abilities of animals to differentiate looming and receding prepulse stimuli, although it did not abolish their discrimination between looming and stationary prepulse stimuli. This suggests that isolation rearing compromised their assessment of potential threats from approaching objects and receding objects.

Keywords: auditory looming bias, prepulse inhibition, attention, isolation rearing, rats

Edited by:
Stephen Charles Van Hedger, Huron University College, Canada

Reviewed by:
Stephanie Clarke, Centre Hospitalier Universitaire Vaudois (CHUV), Switzerland
Susanne Schmid, Western University, Canada
Cathrin Rohleder, The University of Sydney, Australia

****Correspondence:***
Zhemeng Wu
zhemeng.wu@utoronto.ca
Liang Li
liangli@pku.edu.cn

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

Received: *13 July 2021*
Accepted: *11 October 2021*
Published: *15 November 2021*

Citation:
Wu Z, Bao X, Liu L and Li L (2021) Looming Effects on Attentional Modulation of Prepulse Inhibition Paradigm. Front. Psychol. 12:740363. doi: 10.3389/fpsyg.2021.740363

INTRODUCTION

The detection of an approaching object is fundamentally important to the survival of an organism. For instance, the decision by an animal in the wild to forage for food requires an evaluation of predation by assessing their approaching behavior (e.g., faster pace) and their associated probabilities of occurrence and magnitudes. In auditory field, looming sounds with rising intensity and receding sounds with falling intensity are primary cues to aid in judging the motion of objects (Seifritz et al., 2002; Hall and Moore, 2003; Baumgartner et al., 2017). The phenomenon of looming sounds containing an intrinsic and unconditioned warning value therefore being more salient than receding sounds, is termed as "auditory looming bias" (Seifritz et al., 2002; Hall and Moore, 2003; Maier et al., 2004; Baumgartner et al., 2017; Glatz and Chuang, 2019; Bidelman and Myers, 2020). Previous literature reveals that sound spectrum affects the perception of auditory looming bias. Compared with looming white noise with equal increasing intensity, complex pure tones (Neuhoff, 1998, 2001) or harmonic tones (Deneux et al., 2016) elicited stronger auditory looming bias. For example,

macaques were more attracted to the approaching harmonic tones by orienting them to longer time than receding tones (Ghazanfar et al., 2002).

Prepulse inhibition (PPI) of the acoustic startle response paradigm is the suppression of the startle reflex when an intense startling stimulus is preceded by a weaker sensory stimulus (the prepulse) (Fendt et al., 2001; Du et al., 2009; Li, 2009; Li et al., 2009). Graham (1975) proposed a "protection of processing" theory for justifying the function of PPI: the weaker prepulse preceding the startling noise triggers a gating mechanism that dampens the disruptive effects caused by the startling noise (Graham, 1975). Therefore PPI has been recognized as an operational cross-species measure of sensorimotor gating mechanism to help humans and animals adapt to the complex environment (Swerdlow et al., 2001, 2006). The "protection of processing" theory of PPI (Graham, 1975) makes PPI a good behavioral paradigm for studying how the salient value of auditory looming sounds as a prepulse would protect the organism from interference by disruptive stimuli. In this study, we used PPI of the acoustic startle response paradigm to investigate the effects of looming sounds on sensorimotor gating in rats.

Previous literature shows that emotional or spatial attention to the prepulse can enhance PPI of the acoustic startle response. In rats, when a prepulse is paired with the foot shock and becomes emotionally salient, PPI induced by this fear-conditioned prepulse is enhanced (Du et al., 2009, 2010, 2011; Li, 2009; Li et al., 2009). Additionally, PPI can be further enhanced by a spatially separated conditioned prepulse from background noise masker than a spatially co-located prepulse with the masker (Du et al., 2009, 2010, 2011; Li, 2009; Li et al., 2009), which illustrates that spatial attention to the prepulse enhances PPI. Both the emotional fear conditioning–induced PPI enhancement and perceptual spatial separation–induced PPI enhancement reveal that PPI can be modulated by attention (Li et al., 2009). These attentional enhancements of PPI are achieved by acquiring salient valence of prepulse (e.g., fear conditioning), and it is of interest to know whether a prepulse with natural and intrinsic salience (e.g., looming sounds) can also draw the attention of rats, therefore enhancing the PPI. The main aim of this study was to investigate how the auditory looming sound as a prepulse affects PPI compared with the receding sound.

Looming sounds are perceived as spanning a longer duration compared with receding sounds with the same presentation duration (Schlauch et al., 2001; Grassi and Darwin, 2006; DiGiovanni and Schlauch, 2007; Grassi, 2010). This subjective different temporal judgment between looming and receding sounds indicates that looming sounds are attended for longer duration than receding sounds (Glatz and Chuang, 2019). As the duration of prepulse has an impact on PPI (Swerdlow et al., 2001; Li et al., 2009), in our study, we manipulated the duration/rate of both looming and receding sounds as prepulse and kept the total amount of intensity change the same across the two prepulse stimuli (i.e., 50 dB SPL), to investigate whether the auditory looming effects on PPI are duration/rate dependent. As the total amount of intensity change was equal to the duration multiplied by the change rate of prepulse, in our study, change of duration also caused change of intensity rate for looming and receding sounds.

Isolation rearing starting from early life is one of the commonly used animal models, to mimic the behavioral and cognitive impairments in many mental disorders, such as schizophrenia (Li et al., 2009; Wu et al., 2018), depression and anxiety (Thorsell et al., 2006; Mosaferi et al., 2015; Ma et al., 2017; Harvey et al., 2019), attention-deficit hyperactivity disorder (ADHD) (Yates et al., 2012; Ouchi et al., 2013), and other general early-life stress effects upon the brain and behavior during adulthood (Weiss et al., 2000, 2004). Among the cognitive/behavioral deficits caused by isolation rearing, attentional impairment is found in several behavioral paradigms (Wu et al., 2018). For example, it has been found that isolation rearing abolished both emotional and spatial attentional enhancements of PPI (Du et al., 2009, 2010; Li et al., 2009; Lim et al., 2012; Wu et al., 2016), declined attentional processing to novelty (Atmore et al., 2020), and impaired attentional set shifting (McLean et al., 2010). As looming sounds contain an intrinsic salient value for organism to avoid danger, it is of interest to investigate how isolation rearing would preserve or abolish this intrinsic attention to approaching objects. The second aim of this study was to investigate how isolation rearing affects auditory looming effects on attentional modulation of PPI. As previous literature documented an attentional deficit caused by isolation rearing (Du et al., 2009, 2010; Li et al., 2009; McLean et al., 2010; Lim et al., 2012; Wu et al., 2016, 2018; Atmore et al., 2020), we hypothesized that isolation rearing might impair the abilities of animals to differentiate looming and receding sounds; therefore no PPI difference between the two prepulse stimuli would be observed.

MATERIALS AND METHODS

Animals

Twelve Sprague–Dawley rats participated in Experiment 1. Another eighteen Sprague–Dawley pups were involved in Experiment 2. All the rats and the pregnant female rats were purchased from the Vital-River Experimental Animals Technology Ltd., Beijing, China. They were transported in a special vehicle to our laboratory in cages of five to six animals each. All animal training and experimental procedures were performed in accordance with the guidelines of the Beijing Laboratory Animal Center, and the Policies on the Use of Animals and Humans in Neuroscience Research approved by the Society for Neuroscience (2006).

The twelve adult Sprague–Dawley rats in Experiment 1 were socially housed and tested after approximately 2 months of their arrival to the laboratory. The eighteen pups remained in the litter along with their lactating mothers until weaning on postnatal day (PND) 21. After weaning, each of the pups was randomly assigned to either the social rearing group (reared in pairs, $N = 10$) or the isolation rearing group (reared individually, $N = 8$) for 8 weeks. The pups assigned to either the social or isolation rearing group were tested in Experiment 2. The eighteen pups in Experiment 2 were tested on their PND 77. Each individual rat

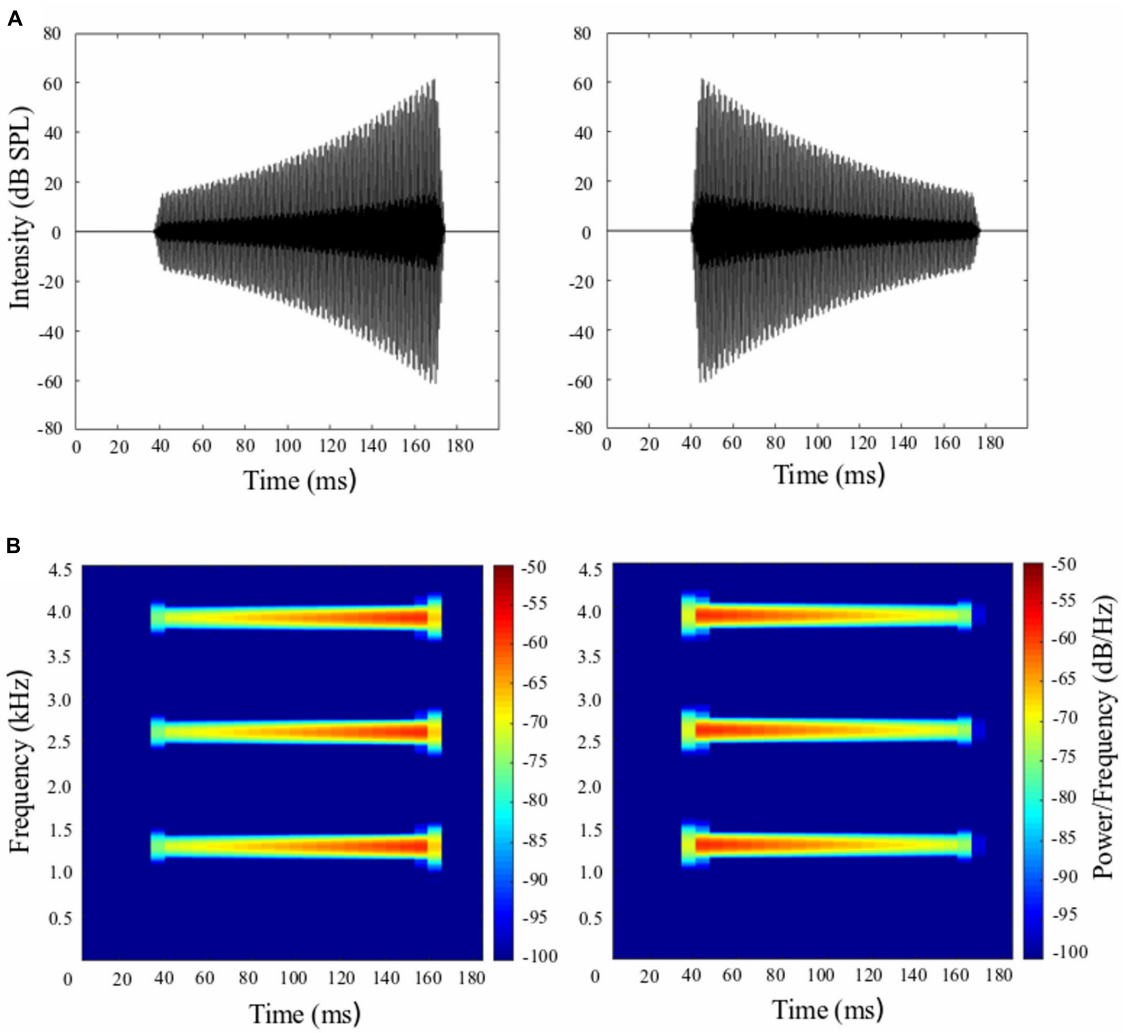

FIGURE 1 | Rising- and falling-intensity harmonic tone stimuli used in the prepulse inhibition (PPI) paradigm. **(A)** The waveform of looming prepulse with rising intensity from 17 to 67 dB SPL (left panel) and receding prepulse with falling intensity from 67 to 17 dB SPL (right panel) with a duration of 120 ms. **(B)** The frequency spectrum for looming prepulse (left panel) and receding prepulse (right panel).

(isolation-reared) and three individual rats (socially reared) were housed in a single transparent plastic cage (48 × 30 × 18 cm, the cage size was the same in socially reared and isolation-reared rats), under a temperature of 24 ± 2°C and a 12-h light/dark cycle with food and water freely available.

Stimuli and Apparatus

The whole-body startle reflex of rat was induced by an intense 10-ms broadband noise burst (0–10 kHz, 100-dB SPL) delivered by a loudspeaker above the rat's head. The electrical voltage signals were sampled at 16 kHz and collected for 500 ms. In a single trial, as the startling stimulus could reliably induce a distinct waveform complex of the startle response (Zou et al., 2007; Du et al., 2011), the startle response was digitized and measured as the peak-to-peak amplitude between the primary peak component and the subsequent peak component. The prepulse stimulus, which ended 50 ms prior to the startling noise, was a 30, 120, 160, or 200-ms three-harmonic tone complex (1.3, 2.6, 3.9 kHz) with rising intensity (i.e., looming prepulse) and falling intensity (i.e., receding prepulse) (**Figure 1**). The sound level of looming prepulse rose from 17 to 67 dB SPL; and the sound level of receding prepulse dropped from 67 to 17 dB SPL. The stationary prepulse was the same three-harmonic tone complex with a stable intensity of 67 dB SPL. We controlled the amount of intensity change the same for looming and receding prepulse stimuli (i.e., 50 dB SPL) and changed the duration, therefore the corresponding rate for 30, 120, 160, and 200 ms prepulse could be calculated as (67–17)/30 = 1.67 dB SPL/ms; (67–17)/120 = 0.42 dB SPL/ms; (67–17)/160 = 0.31 dB SPL/ms; and (67–17)/200 = 0.025 dB SPL/ms. We could see shorter duration was accompanied with higher change rate and *vice versa*. The root mean square (RMS) level of looming and receding prepulses were identical: the RMS value was 7.0, 24.5, 32.3, and 40.1 for durations of 30, 120, 160, and 200 ms, respectively. This indicates that the

energy was the same for all prepulses at a given duration. The Sprague–Dawley albino rats have the greatest hearing sensitivity, which is between 8 and 38 kHz. The best hearing points are 8 kHz and 32–38 kHz. Between 8 and 32 kHz, the sensitivity levels slightly decline (Kelly and Masterton, 1976; Borg, 1982; Heffner et al., 1994). Although the prepulse stimuli used in our study are not within the range of the highest hearing sensitivity, the rat's hearing threshold for 1–2 kHz tones is approximately 25 dB SPL and for 3–4 kHz tones is approximately 15 dB SPL (Kelly and Masterton, 1976; Borg, 1982; Heffner et al., 1994). Thus, our rats could detect the stationary prepulse stimuli at 67 dB SPL, looming and receding prepulse stimuli at 3.9 kHz above 15 dB SPL (i.e., within the range of 17–67 dB SPL), and at 1.3 and 2.6 kHz above 25 dB SPL. The 17dB SPL is more likely not loud enough to detect the 1.3 and 2.6 kHz tone of the three-harmonic tone complex. This is one of the weaknesses of our study when choosing the frequency and intensity of three-harmonic tone complex as looming and receding prepulse stimuli. The prepulse was delivered by each of the two horizontal and spatially separated loudspeakers, which were placed horizontally in the frontal field with a 100° separation angle and 52 cm away from the rat's head position. All the sound stimuli were digitally generated by Adobe Audition software and converted by a custom-developed sound delivery system (National Key Laboratory on Machine Perception, Peking University). Calibration of sound intensity was conducted with a Larson Davis Audiometer Calibration and Electro-acoustic Testing System (AUDit & System 824, Larson Davis, Depew, NY, United States).

Testing Procedures

On the first three successive days, the rat was placed inside a restraining cage, whose dimensions matched the size of the rat body so that the rat could not reorient its head and body positions, and the rat was exposed to the broadband noise (60 dB SPL) delivered for 30 min. The broadband noise was continuously presented by each of the two horizontal loudspeakers. Neither the prepulse nor the startling noise was presented during the acclimation. This procedure was to adapt the rat to the restraining cage and testing chamber. The enclosure was wiped down with ethanol solution after each animal.

On the fourth day (test day), the rat was tested in the same environment embedded with the background noise. The rat received 10 presentations of startling stimulus without prepulse presentation for the first 5 min. In Experiment 1, the testing block included 45 trials: 5 trials containing startling noise alone, 20 trials containing a looming prepulse with rising intensity preceding startling noise (with 5 trials per each prepulse duration), and 20 trials containing another receding prepulse with falling intensity preceding startling noise (with 5 trials per each prepulse duration). In Experiment 2, the duration of prepulse was fixated (i.e., 120 ms) based on the result of Experiment 1, the testing block included 20 trials: 5 trials containing startling noise alone, 5 trials containing a looming prepulse preceding startling noise, and 5 trials containing a stationary prepulse preceding startling noise, and 5 trials containing a receding prepulse preceding startling noise. The three types of trials within a testing block were presented in a random order, and this order of the presentation was randomized across rats. The randomization of the presentation order and rat order was created using a "RAND" formula in Excel. The inter-stimulus onset interval between a prepulse and the startling stimulus was fixated at 100 ms. The interval between each trial varied between 25 and 35 s (mean = 30 s).

Data Analysis

The startle response was measured as the peak-to-peak amplitude between the primary peak component (latency mainly between 15 and 20 ms) and the subsequent peak component (latency mainly between 20 and 25 ms), and were baseline corrected. The startle responses were reliable across all the rats. All the trials, including startling noise-only trials and prepulse trials, were included in the analysis. The value of PPI of the acoustic startle response was calculated with the below formula:

$$PPI\,(\%) = \frac{(amplitude\ to\ startling\ noise\ alone - amplitude\ to\ startling\ noise\ preceded\ by\ prepulse)}{amplitude\ to\ startling\ noise\ alone} \times 100\%$$

In Experiment 1, a two-way repeated-measures ANOVA with the factors prepulse duration (four levels) and prepulse type (two levels) was conducted. In Experiment 2, a two-way mixed-design ANOVA with the factors rearing conditions (independent factor, two levels) and prepulse type (dependent factor, three levels) was conducted. All the statistical analyses were performed using SPSS 15.0 software. The null hypothesis rejection level was set at 0.05. The data were presented using raincloud plots (Allen et al., 2019).

RESULTS

Experiment 1: Looming Effects on Attentional Modulation of Prepulse Inhibition at Different Prepulse Durations in Socially Reared Rats

We found a difference of PPI induced by looming prepulse with rising intensity and receding prepulse with falling intensity that depended on the duration of prepulse (**Figure 2A**). A two-way repeated measures ANOVA was conducted on the PPI data, with prepulse type (2 levels: looming and receding) and prepulse duration (4 levels: 30, 120, 160, and 200 ms) as the within-subject factor. We found a marginal significant quadratic trend interaction for prepulse duration ($F_{(1,11)} = 3.441$, $p = 0.091$, $\eta^2 = 0.238$). The pairwise t-tests further showed that PPI elicited by looming prepulse was marginally significantly larger than the PPI induced by receding prepulse when the duration of prepulse was 120 ms ($t_{(11)} = 2.715$, $p = 0.020$, Bonferroni corrected $p = 0.080$). However, no significant difference of PPI occurred between looming and receding prepulse at 30 ms ($t_{(11)} = -1.764$, $p = 0.105$, Bonferroni corrected $p = 0.420$), 160 ms ($t_{(11)} = 0.296$, $p = 0.773$, Bonferroni corrected $p = 1.000$), and 200 ms ($t_{(11)} = -0.387$, $p = 0.706$, Bonferroni corrected $p = 1.000$).

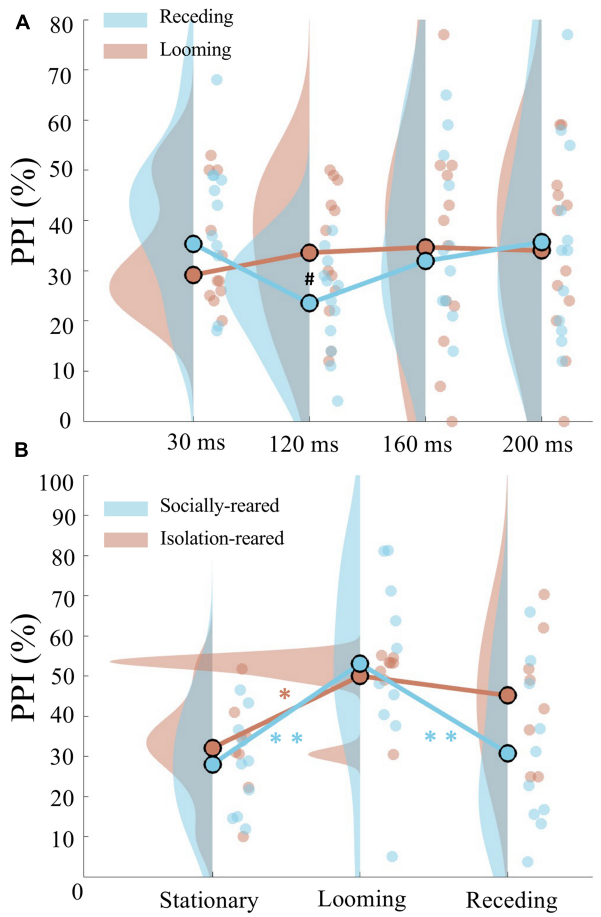

FIGURE 2 | Prepulse duration and rearing type on looming-induced attentional modulation of PPI. **(A)** Raincloud plots of PPI induced by looming prepulse (red) and receding prepulse (blue) when the prepulse duration was 30, 120, 160, and 200 ms. Note that a marginal significant difference of PPI between the looming and receding prepulses was observed when both the prepulses were 120 ms. **(B)** Different modulated PPI by two rearing types. Not that in socially reared rats (blue), PPI was significantly increased by looming prepulse than PPI induced by stationary and receding prepulse; in isolation-reared rats (red), PPI was only increased by looming prepulse than PPI induced by stationary prepulse. No difference of PPI between looming and receding prepulses was found. Circles are individual data; circles with black edge color show the average of group data; distributions show probability density function of data points. *Indicates $p < 0.05$, **indicates $p < 0.01$, and #indicates $p < 0.1$.

Experiment 2: Different Looming Effects on Attentional Modulation of Prepulse Inhibition in Socially Reared and Isolation-Reared Rats

In Experiment 1, the difference of looming-induced PPI enhancement occurred when the duration of prepulse was 120 ms, thus we chose 120 ms prepulse and further investigated whether the looming-induced PPI enhancement would be different between socially reared and isolation-reared rats in Experiment 2.

A two-way mixed-design repeated measures ANOVA was conducted on the PPI data, with prepulse type (three levels: stationary, looming, and receding) as the within-subject factor and group (two levels: socially rearing, isolation rearing) as the between-subject factor. A significant main effect of prepulse type ($F_{(2,32)} = 16.872$, $p = 0.00001$, $\eta^2 = 0.513$) was found, in addition to a significant linear trend interaction between prepulse type and group ($F_{(1,16)} = 5.787$, $p = 0.029$, $\eta^2 = 0.266$). The latter prompted further analysis on the two animal groups. In socially reared rats, repeated measures ANOVA showed a main effect of prepulse type ($F_{(2,18)} = 16.416$, $p = 0.000088$, $\eta^2 = 0.646$). The *post hoc* pairwise *t*-tests with Bonferroni correction revealed a larger PPI induced by looming prepulse than stationary prepulse ($p = 0.002$) and a larger PPI induced by looming prepulse than receding prepulse ($p = 0.004$). In isolation-reared rats, another repeated measures ANOVA showed a main effect of prepulse type ($F_{(2,14)} = 5.030$, $p = 0.023$, $\eta^2 = 0.418$), which prompted further scrutiny by *post hoc* pairwise *t*-tests with Bonferroni correction. These revealed a significant larger PPI induced by looming prepulse compared with PPI induced by stationary prepulse ($p = 0.034$), although no significant difference of PPI between the looming and receding prepulses was found ($p = 1.000$) (**Figure 2B**).

DISCUSSION

The results of our study demonstrated that the looming sounds with an adequate duration (i.e., 120 ms) induced PPI enhancement compared with receding sounds with the same duration, suggesting that approaching sounds serve an intrinsic warning cue for individuals to dampen the startle response elicited by a sudden and interfering stimuli. The "auditory looming bias" shown in previous literature (Neuhoff, 2001; Ghazanfar et al., 2002; Seifritz et al., 2002; Hall and Moore, 2003; Maier et al., 2004; Maier and Ghazanfar, 2007; Grassi, 2010; Baumgartner et al., 2017; Glatz and Chuang, 2019; Bidelman and Myers, 2020) was also found in PPI of the acoustic startle response paradigm in rats. We further showed that this looming effect–induced PPI enhancement was time or rate dependent. When the duration of looming prepulse was either too short (e.g., 30 ms, a faster rate) or too long (e.g., longer than 160 ms, a slower rate), PPI enhancements induced by the looming sounds disappeared for the fact that there was no difference of PPI elicited by the looming and receding sounds. This illustrates that an adequate processing time of approaching objects is needed to capture the attention of the rats and therefore enhanced PPI. It is also noted that recency effect of looming sounds could not explain our results. If rats differentiated looming and stationary prepulse based on recency effect, we should have found no difference of PPI between stationary (67dB SPL) and looming prepulse (the last tail intensity was 67 dB SPL). However, we found a significant difference between these two prepulse stimuli in both the animal groups. Previous literature documented various duration of approaching sounds–induced auditory looming bias in animals and humans: macaques demonstrated behavioral bias for looming sounds at

750 ms (Ghazanfar et al., 2002) and their neural discharge rate for rising and falling intensity sounds differed at a shorter duration (i.e., 25 ms) (Lu et al., 2001); humans judged longer for approaching than receding sounds when they were in the duration of 200 ms (Schlauch et al., 2001), or as long as 1,000 ms (Grassi and Darwin, 2006). How the temporal information is maintained in looming sounds and how the duration of looming sound affects the auditory looming bias are of interest for future research to advance.

The cognitive mechanism of different perceptions of looming and receding sounds has been studied in previous literature. One of the mechanisms is that looming sounds induced a loudness bias whereby its highest intensity of looming sounds was perceived in the end, therefore looming sounds were perceived louder than receding sounds and more easily captured attention (Neuhoff, 2001; Susini et al., 2010; Aumond et al., 2017). Another explanation is that looming sounds contained the first unattenuated segment which received the highest weight and therefore captured attention, while receding sounds had their first fade-in segments which were ignored (Oberfeld and Plank, 2011). For example, the decay portions or the tails of receding sounds were easy to be ignored, especially in a reverberant environment (e.g., the background broadband noise in our experiment), thus the tails may not be considered a meaningful part for the judgment of sound motion (DiGiovanni and Schlauch, 2007).

The neural mechanisms of perceiving looming and receding sounds differed. In animals, auditory looming sounds activated primary auditory cortex stronger than receding sounds in marmosets (Lu et al., 2001) and macaques (Maier and Ghazanfar, 2007). In humans, neuroimaging evidence showed that compared with receding sounds, looming sounds activated a broader neural network subserving motion perception, including the motor and premotor cortices and the cerebellum; and subserving attention, compromising the superior temporal sulci, the middle temporal gyri, and the temporoparietal junction (Seifritz et al., 2002). Another magnetoencephalography study in humans also confirmed that rising intensity complex sounds induced a stronger activation in bilateral inferior temporal gyrus and right temporoparietal junction compared with falling intensity complex (Bach et al., 2015). Taken together, looming sounds were more salient than receding sounds and involved more activations in temporal and frontal areas in animals and humans.

The neural circuitry mediating PPI is mostly located in the midbrain, comprising inferior colliculus (Li et al., 1998a,b, 2009), deeper layers, and intermediate layers of the superior colliculus (Fendt et al., 2001; Li et al., 2009), and primary auditory cortex (Du et al., 2011). PPI can be enhanced when the prepulse is endowed with emotional attention. For example, when the salient valence of prepulse was acquired by pairing it with foot shock, it has been found that this emotional attentional modulation of PPI relied on the lateral amygdala of rats (Du et al., 2011). Furthermore, this fear-conditioned prepulse induced larger PPI when this prepulse was spatially separated from background noise masker than the same prepulse was co-located with the masker. This spatial attentional modulation of PPI depended on the posterior parietal cortex of the rats (Du et al., 2011). One of the potential areas in mediating the approaching sound–induced PPI enhancement may be the primary auditory cortex of rats, as it was found to be involved in both the neural circuit of regulating PPI (Li et al., 2009; Du et al., 2011) and looming sound perception (Lu et al., 2001; Maier and Ghazanfar, 2007). Compared with receding prepulse, the approaching or looming prepulse contained an intrinsic salient value, which could draw attention to it, therefore enhanced PPI.

Besides primary auditory cortex, prefrontal cortex (PFC) is involved in bias perception by looming sounds. In a human electroencephalography study, neural signals in PFC differentiated looming and receding stimuli at an early stage, indicating a faster top–down control on prioritizing approaching events *via* PFC (Bidelman and Myers, 2020). Its involvement in regulating PPI and attentional modulation of PPI has also been well documented. In rats, a behavioral positron emission tomography study reported that the left frontal cortex (area 3) and the left and right prelimbic cortex were related to PPI modulation (Rohleder et al., 2016). Furthermore, a relationship between decreased function and volume of the medial PFC and low PPI values was found (Tapias-Espinosa et al., 2019). It has been further suggested that the PFC mediates the attentional enhancement of PPI of the acoustic startle reflex (Meng et al., 2020). Additionally, prefrontal dysfunction caused by isolation stress during adolescence resulted in PPI deficits (Drzewiecki et al., 2021; Du et al., 2021). What other brain areas are involved in and how the brain network subserving attention is involved in this process need further investigation.

Isolation rearing abolished the looming sound–induced PPI enhancements to some extent. We found that in socially reared rats, looming sounds caused higher PPI than stationary and receding sounds with the same duration. Although isolation rearing did not abolish their discrimination ability for looming and stationary sounds, it impaired their discrimination between looming and receding sounds as no difference of PPI was found. The failure of differentiating looming and receding sounds in isolation-reared rats may be related to their flawed attention (Li et al., 2009; Wu et al., 2018). Previously we found that isolation rearing abolished both fear conditioning–induced PPI enhancement (Du et al., 2010; Wu et al., 2016) and perceptual spatial separation–induced PPI enhancement (Du et al., 2009; Wu et al., 2016), which indicates an impairment of attending to the prepulse with acquired salient valence. Our study further extends it to show that isolation rearing abolished their attention to intrinsic salient value of prepulse (i.e., looming sounds) without acquisition. From evolutionary perspective, the failure of attending to potential biological salience of looming cues and lack of adaptive bias for an approaching object may leave their life at risk in the real world. PPI is a cross-species paradigm, which has been commonly used in animals (Wilkinson et al., 1994; Li et al., 1998a,b, 2009; Weiss et al., 2000; Swerdlow et al., 2001; van den Buuse et al., 2003; Zou et al., 2007; Du et al., 2009, 2011; Li, 2009; Uehara et al., 2009; Zhao et al., 2013; Lei et al., 2014; Wu et al., 2016, 2019; Ding et al., 2019) and humans (Li et al., 2009; Lei et al., 2018). The looming effect–induced attentional modulation of PPI can be used as a behavioral paradigm to probe attentional deficits in animal models and patients with mental disorders,

such as schizophrenia (Dawson et al., 1993; Chen and Faraone, 2000; Dondé et al., 2020) and ADHD (Sharma and Couture, 2014; Tarver et al., 2014; Luo et al., 2019).

There are some limitations to our current study. First, the chosen frequency and intensity of looming and receding prepulse stimuli may make rats not to detect all the three components of the harmonic tone complex. To make a comparison to external attentional modulation of PPI paradigms (Zou et al., 2007; Du et al., 2011; Wu et al., 2016), we chose the same broadband noise at 60 dB SPL as background noise masker and the three-harmonic tone complex (1.3, 2.6, 3.9 kHz) as prepulse stimuli in our study. This was not an issue for stationary prepulse stimuli (67 dB SPL) but caused the intensity of the looming and receding prepulse (17–67 dB SPL) below background noise for some time (approximately 40 ms when the intensity was below 25 dB SPL, see **Figure 1**) and thus it is hard for rats to detect 1.3 and 2.6 kHz of the three-harmonic tone complex. Additionally, our rats may not be able to detect the 1.3 and 2.6 kHz tone of the three-harmonic tone complex at 17 dB SPL, since their hearing threshold for 1–2 kHz tones is approximately 25 dB SPL (Kelly and Masterton, 1976; Borg, 1982; Heffner et al., 1994). As the hearing threshold for 3–4 kHz tones is approximately 15 dB SPL (Kelly and Masterton, 1976; Borg, 1982; Heffner et al., 1994), our rats could detect 3.9 kHz of the three-harmonic tone complex for both looming and receding prepulse stimuli (17–67 dB SPL). Second, the lower detection of 1.3 and 2.6 kHz tone of the three-harmonic tone complex may cause the interval between prepulse and startling stimuli differed for looming and receding prepulse conditions. As stated above, the two tones were most likely detected for approximately 80 ms in a 120 ms prepulse. Consequently, the time between the recognized prepulse and the startle stimulus was longer in the receding prepulse condition than in the looming prepulse condition. This effect was additionally increased as the background noise was 60 dB SPL during each session and may affect the effectiveness of the prepulse on the PPI of the acoustic startle response. Third, the rats tested in Experiment 1 were involved in another attentional modulation of PPI paradigms, in which they have encountered a 50-ms three-harmonic tone complex (1.3, 2.6, 3.9 kHz) and another three-harmonic tone complex (2.3, 4.6, and 6.9 kHz). As the previous prepulse had the same frequency range of the one used in current study, their early exposure to the same frequency prepulse may cause less sensitivity to the looming prepulse used in our study and therefore their PPI was lower compared with the PPI in socially reared rats in Experiment 2.

In conclusion, our study showed a looming effect–induced PPI enhancement, which is that rising-intensity sound at an adequate time duration could effectively dampen the startle response elicited by a sudden stimulus, therefore it helps organisms to adapt to the complex environment by recruiting attentional and physiological resources. Isolation rearing impaired the abilities of the animals to discriminate between the looming and receding sounds, which makes them vulnerable to the approaching predators in the wild.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The animal study was reviewed and approved by the Beijing Laboratory Animal Center, in the School of Psychological and Cognitive Sciences at Peking University.

AUTHOR CONTRIBUTIONS

ZW, XB, LeL, and LiL conceived the project. LiL supervised this research project. ZW and XB conducted the experiment. ZW performed the data analysis. ZW and LiL interpreted the data and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 31771252 and 32071057) and the High-Performance Computing Platform of Peking University.

REFERENCES

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., and Kievit, R. A. (2019). Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Res.* 4:63. doi: 10.12688/wellcomeopenres.15191.1

Atmore, K. H., Stein, D. J., Harvey, B. H., Russell, V. A., and Howells, F. M. (2020). Differential effects of social isolation rearing on glutamate- and GABA-stimulated noradrenaline release in the rat prefrontal cortex and hippocampus. *Eur. Neuropsychopharmacol.* 36, 111–120. doi: 10.1016/j.euroneuro.2020.05.007

Aumond, P., Can, A., de Coensel, B., Ribeiro, C., Botteldooren, D., and Lavandier, C. (2017). Global and continuous pleasantness estimation of the soundscape perceived during walking trips through urban environments. *Appl. Sci.* 7:144. doi: 10.3390/app7020144

Bach, D. R., Furl, N., Barnes, G., and Dolan, R. J. (2015). Sustained magnetic responses in temporal cortex reflect instantaneous significance of approaching and receding sounds. *PLoS One* 10:e0134060. doi: 10.1371/journal.pone.0134060

Baumgartner, R., Reed, D. K., Tóth, B., Best, V., Majdak, P., Colburn, H. S., et al. (2017). Asymmetries in behavioral and neural responses to spectral cues demonstrate the generality of auditory looming bias. *Proc. Natl. Acad. Sci. U.S.A.* 114, 9743–9748. doi: 10.1073/pnas.1703247114

Bidelman, G. M., and Myers, M. H. (2020). Frontal cortex selectively overrides auditory processing to bias perception for looming sonic motion. *Brain Res.* 1726:146507. doi: 10.1016/j.brainres.2019.146507

Borg, E. (1982). Auditory thresholds in rats of different age and strain. A behavioral and electrophysiological study. *Hear. Res.* 8, 101–115. doi: 10.1016/0378-5955(82)90069-7

Chen, W. J., and Faraone, S. V. (2000). Sustained attention deficits as markers of genetic susceptibility to schizophrenia. *Am. J. Med. Genet. Semin. Med. Genet.* 97, 52–57.

Dawson, M. E., Hazlett, E. A., Filion, D. L., Nuechterlein, K. H., and Schell, A. M. (1993). Attention and schizophrenia: impaired modulation of the startle reflex. *J. Abnorm. Psychol.* 102, 633–641. doi: 10.1037/0021-843X.102.4.633

Deneux, T., Kempf, A., Daret, A., Ponsot, E., and Bathellier, B. (2016). Temporal asymmetries in auditory coding and perception reflect multi-layered nonlinearities. *Nat. Commun.* 7:12682. doi: 10.1038/ncomms12682

DiGiovanni, J. J., and Schlauch, R. S. (2007). Mechanisms responsible for differences in perceived duration for rising-intensity and falling-intensity sounds. *Ecol. Psychol.* 19, 239–264. doi: 10.1080/10407410701432329

Ding, Y., Xu, N., Gao, Y., Wu, Z., and Li, L. (2019). The role of the deeper layers of the superior colliculus in attentional modulations of prepulse inhibition. *Behav. Brain Res.* 364, 106–113. doi: 10.1016/j.bbr.2019.01.052

Dondé, C., Brunelin, J., Mondino, M., Cellard, C., Rolland, B., and Haesebaert, F. (2020). The effects of acute nicotine administration on cognitive and early sensory processes in schizophrenia: a systematic review. *Neurosci. Biobehav. Rev.* 118, 121–133. doi: 10.1016/j.neubiorev.2020.07.035

Drzewiecki, C. M., Willing, J., Cortes, L. R., and Juraska, J. M. (2021). Adolescent stress during, but not after, pubertal onset impairs indices of prepulse inhibition in adult rats. *Dev. Psychobiol.* 63, 837–850. doi: 10.1002/dev.22111

Du, W., Li, M., Zhou, H., Shao, F., and Wang, W. (2021). Alteration of the PKA-CREB cascade in the mPFC accompanying prepulse inhibition deficits: evidence from adolescent social isolation and chronic SKF38393 injection during early adolescence. *Behav. Pharmacol.* 32, 487–496. doi: 10.1097/FBP.0000000000000643

Du, Y., Li, J., Wu, X., and Li, L. (2009). Precedence-effect-induced enhancement of prepulse inhibition in socially reared but not isolation-reared rats. *Cogn. Affect. Behav. Neurosci.* 9, 44–58. doi: 10.3758/CABN.9.1.44

Du, Y., Wu, X., and Li, L. (2010). Emotional learning enhances stimulus-specific top-down modulation of sensorimotor gating in socially reared rats but not isolation-reared rats. *Behav. Brain Res.* 206, 192–201. doi: 10.1016/j.bbr.2009.09.012

Du, Y., Wu, X., and Li, L. (2011). Differentially organized top-down modulation of prepulse inhibition of startle. *J. Neurosci.* 31, 13644–13653. doi: 10.1523/JNEUROSCI.1292-11.2011

Fendt, M., Li, L., and Yeomans, J. S. (2001). Brain stem circuits mediating prepulse inhibition of the startle reflex. *Psychopharmacology* 156, 216–224. doi: 10.1007/s002130100794

Ghazanfar, A. A., Neuhoff, J. G., and Logothetis, N. K. (2002). Auditory looming perception in rhesus monkeys. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15755–15757. doi: 10.1073/pnas.242469699

Glatz, C., and Chuang, L. L. (2019). The time course of auditory looming cues in redirecting visuo-spatial attention. *Sci. Rep.* 9:743. doi: 10.1038/s41598-018-36033-8

Graham, F. K. (1975). The more or less startling effects of weak prestimulation. *Psychophysiology* 12, 238–248. doi: 10.1111/j.1469-8986.1975.tb01284.x

Grassi, M. (2010). Sex difference in subjective duration of looming and receding sounds. *Perception* 39, 1424–1426. doi: 10.1068/p6810

Grassi, M., and Darwin, C. J. (2006). The subjective duration of ramped and damped sounds. *Percept. Psychophys.* 68, 1382–1392. doi: 10.3758/BF03193737

Hall, D. A., and Moore, D. R. (2003). Auditory neuroscience: the salience of looming sounds. *Curr. Biol.* 13, 91–93. doi: 10.1016/S0960-9822(03)00034-4

Harvey, B. H., Regenass, W., Dreyer, W., and Möller, M. (2019). Social isolation rearing-induced anxiety and response to agomelatine in male and female rats: role of corticosterone, oxytocin, and vasopressin. *J. Psychopharmacol.* 33, 640–646. doi: 10.1177/0269881119826783

Heffner, H. E., Heffner, R. S., Contos, C., and Ott, T. (1994). Audiogram of the hooded Norway rat. *Hear. Res.* 73, 244–247. doi: 10.1016/0378-5955(94)90240-2

Kelly, J. B., and Masterton, R. B. (1976). Auditory sensitivity of the albino rat. *J. Acoust. Soc. Am.* 60:S88. doi: 10.1121/1.2003583

Lei, M., Luo, L., Qu, T., Jia, H., and Li, L. (2014). Perceived location specificity in perceptual separation-induced but not fear conditioning-induced enhancement of prepulse inhibition in rats. *Behav. Brain Res.* 269, 87–94. doi: 10.1016/j.bbr.2014.04.030

Lei, M., Zhang, C., and Li, L. (2018). Neural correlates of perceptual separation-induced enhancement of prepulse inhibition of startle in humans. *Sci. Rep.* 8:472. doi: 10.1038/s41598-017-18793-x

Li, L. (2009). Precedence-effect-induced enhancement of prepulse inhibition in socially reared but not isolation-reared rats. *Cogn. Affect. Behav. Neurosci.* 9, 44–58.

Li, L., Du, Y., Li, N., Wu, X., and Wu, Y. (2009). Top-down modulation of prepulse inhibition of the startle reflex in humans and rats. *Neurosci. Biobehav. Rev.* 33, 1157–1167. doi: 10.1016/j.neubiorev.2009.02.001

Li, L., Korngut, L. M., Frost, B. J., and Beninger, R. J. (1998a). Prepulse inhibition following lesions of the inferior colliculus: prepulse intensity functions. *Physiol. Behav.* 65, 133–139. doi: 10.1016/S0031-9384(98)00143-7

Li, L., Priebe, R. P. M., and Yeomans, J. S. (1998b). Prepulse inhibition of acoustic or trigeminal startle of rats by unilateral electrical stimulation of the inferior colliculus. *Behav. Neurosci.* 112, 1187–1198. doi: 10.1037/0735-7044.112.5.1187

Lim, A. L., Taylor, D. A., and Malone, D. T. (2012). A two-hit model: behavioural investigation of the effect of combined neonatal MK-801 administration and isolation rearing in the rat. *J. Psychopharmacol.* 26, 1252–1264. doi: 10.1177/0269881111430751

Lu, T., Liang, L., and Wang, X. (2001). Neural representations of temporally asymmetric stimuli in the auditory cortex of awake primates. *J. Neurophysiol.* 85, 2364–2380. doi: 10.1152/jn.2001.85.6.2364

Luo, Y., Weibman, D., Halperin, J. M., and Li, X. (2019). A review of heterogeneity in attention deficit/hyperactivity disorder (ADHD). *Front. Hum. Neurosci.* 13:42. doi: 10.3389/fnhum.2019.00042

Ma, J., Wang, F., Yang, J., Dong, Y., Su, G., Zhang, K., et al. (2017). Xiaochaihutang attenuates depressive/anxiety-like behaviors of social isolation-reared mice by regulating monoaminergic system, neurogenesis and BDNF expression. *J. Ethnopharmacol.* 208, 94–104. doi: 10.1016/j.jep.2017.07.005

Maier, J. X., and Ghazanfar, A. A. (2007). Looming biases in monkey auditory cortex. *J. Neurosci.* 27, 4093–4100. doi: 10.1523/JNEUROSCI.0330-07.2007

Maier, J. X., Neuhoff, J. G., Logothetis, N. K., and Ghazanfar, A. A. (2004). Multisensory integration of looming signals by rhesus monkeys. *Neuron* 43, 177–181. doi: 10.1016/j.neuron.2004.06.027

McLean, S. L., Grayson, B., Harris, M., Protheroe, C., Bate, S., Woolley, M. L., et al. (2010). Isolation rearing impairs novel object recognition and attentional set shifting performance in female rats. *J. Psychopharmacol.* 24, 57–63. doi: 10.1177/0269881108093842

Meng, Q., Ding, Y., Chen, L., and Li, L. (2020). The medial agranular cortex mediates attentional enhancement of prepulse inhibition of the startle reflex. *Behav. Brain Res.* 383, 112511. doi: 10.1016/j.bbr.2020.112511

Mosaferi, B., Babri, S., Ebrahimi, H., and Mohaddes, G. (2015). Enduring effects of post-weaning rearing condition on depressive- and anxiety-like behaviors and motor activity in male rats. *Physiol. Behav.* 142, 131–136. doi: 10.1016/j.physbeh.2015.02.015

Neuhoff, J. G. (1998). Perceptual bias for rising tones [4]. *Nature* 395, 123–124. doi: 10.1038/25862

Neuhoff, J. G. (2001). An adaptive bias in the perception of looming auditory motion. *Ecol. Psychol.* 13, 87–110. doi: 10.1207/S15326969ECO1302_2

Oberfeld, D., and Plank, T. (2011). The temporal weighting of loudness: effects of the level profile. *Attent. Percept. Psychophys.* 73, 189–208. doi: 10.3758/s13414-010-0011-8

Ouchi, H., Ono, K., Murakami, Y., and Matsumoto, K. (2013). Social isolation induces deficit of latent learning performance in mice: a putative animal model of attention deficit/hyperactivity disorder. *Behav. Brain Res.* 238, 146–153. doi: 10.1016/j.bbr.2012.10.029

Rohleder, C., Wiedermann, D., Neumaier, B., Drzezga, A., Timmermann, L., Graf, R., et al. (2016). The functional networks of prepulse inhibition: neuronal connectivity analysis based on FDG-PET in awake and unrestrained rats. *Front. Behav. Neurosci.* 10:148. doi: 10.3389/fnbeh.2016.00148

Schlauch, R. S., Ries, D. T., and DiGiovanni, J. J. (2001). Duration discrimination and subjective duration for ramped and damped sounds. *J. Acoust. Soc. Am.* 109, 2880–2887. doi: 10.1121/1.1372913

Seifritz, E., Neuhoff, J. G., Bilecen, D., Scheffler, K., Mustovic, H., Schächinger, H., et al. (2002). Neural processing of auditory looming in the human brain. *Curr. Biol.* 12, 2147–2151. doi: 10.1016/S0960-9822(02)01356-8

Sharma, A., and Couture, J. (2014). A review of the pathophysiology, etiology, and treatment of attention-deficit hyperactivity disorder (ADHD). *Ann. Pharmacother.* 48, 209–225. doi: 10.1177/1060028013510699

Susini, P., Meunier, S., Trapeau, R., and Chatron, J. (2010). End level bias on direct loudness ratings of increasing sounds. *J. Acoust. Soc. Am.* 128, EL163–EL168. doi: 10.1121/1.3484233

Swerdlow, N. R., Geyer, M. A., and Braff, D. L. (2001). Neural circuit regulation of prepulse inhibition of startle in the rat: current knowledge and future challenges. *Psychopharmacology* 156, 194–215. doi: 10.1007/s002130100799

Swerdlow, N. R., Light, G. A., Cadenhead, K. S., Sprock, J., Hsieh, M. H., and Braff, D. L. (2006). Startle gating deficits in a large cohort of patients with schizophrenia. *Arch. Gen. Psychiatry* 63, 1325–1335. doi: 10.1001/archpsyc.63.12.1325

Tapias-Espinosa, C., Río-Álamos, C., Sánchez-González, A., Oliveras, I., Sampedro-Viana, D., Castillo-Ruiz, M., et al. (2019). Schizophrenia-like reduced sensorimotor gating in intact inbred and outbred rats is associated with decreased medial prefrontal cortex activity and volume. *Neuropsychopharmacology* 44, 1975–1984. doi: 10.1038/s41386-019-0392-x

Tarver, J., Daley, D., and Sayal, K. (2014). Attention-deficit hyperactivity disorder (ADHD): an updated review of the essential facts. *Child Care Health Dev.* 40, 762–774. doi: 10.1111/cch.12139

Thorsell, A., Slawecki, C. J., el Khoury, A., Mathe, A. A., and Ehlers, C. L. (2006). The effects of social isolation on neuropeptide Y levels, exploratory and anxiety-related behaviors in rats. *Pharmacol. Biochem. Behav.* 83, 28–34. doi: 10.1016/j.pbb.2005.12.005

Uehara, T., Sumiyoshi, T., Seo, T., Itoh, H., Matsuoka, T., Suzuki, M., et al. (2009). Long-term effects of neonatal MK-801 treatment on prepulse inhibition in young adult rats. *Psychopharmacology* 206, 623–630. doi: 10.1007/s00213-009-1527-2

van den Buuse, M., Garner, B., and Koch, M. (2003). Neurodevelopmental animal models of schizophrenia: effects on prepulse inhibition. *Curr. Mol. Med.* 3, 459–471. doi: 10.2174/1566524033479627

Weiss, I. C., di Iorio, L., Feldon, J., and Domeney, A. M. (2000). Strain differences in the isolation-induced effects on prepulse inhibition of the acoustic startle response and on locomotor activity. *Behav. Neurosci.* 114, 364–373. doi: 10.1037/0735-7044.114.2.364

Weiss, I. C., Pryce, C. R., Jongen-Rêlo, A. L., Nanz-Bahr, N. I., and Feldon, J. (2004). Effect of social isolation on stress-related behavioural and neuroendocrine state in the rat. *Behav. Brain Res.* 152, 279–295. doi: 10.1016/j.bbr.2003.10.015

Wilkinson, L. S., Killcross, S. S., Humby, T., Hall, F. S., Geyer, M. A., and Robbins, T. W. (1994). Social isolation in the rat produces developmentally specific deficits in prepulse inhibition of the acoustic startle response without disrupting latent inhibition. *Neuropsychopharmacology* 10, 61–72. doi: 10.1038/npp.1994.8

Wu, C., Ding, Y., Chen, B., Gao, Y., Wang, Q., Wu, Z., et al. (2019). Both Val158Met polymorphism of Catechol-O-Methyltransferase gene and menstrual cycle affect prepulse inhibition but not attentional modulation of prepulse inhibition in younger-adult females. *Neuroscience* 404, 396–406. doi: 10.1016/j.neuroscience.2019.02.001

Wu, Z., Yang, Z., Zhang, M., Bao, X., Han, F., and Li, L. (2018). The role of N-methyl-D-aspartate receptors and metabotropic glutamate receptor 5 in the prepulse inhibition paradigms for studying schizophrenia: pharmacology, neurodevelopment and genetics. *Behav. Pharmacol.* 29, 13–27. doi: 10.1097/FBP.0000000000000352

Wu, Z. M., Ding, Y., Jia, H. X., and Li, L. (2016). Different effects of isolation-rearing and neonatal MK-801 treatment on attentional modulations of prepulse inhibition of startle in rats. *Psychopharmacology* 233, 3089–3102. doi: 10.1007/s00213-016-4351-5

Yates, J. R., Darna, M., Gipson, C. D., Dwoskin, L. P., and Bardo, M. T. (2012). Isolation rearing as a preclinical model of attention/deficit-hyperactivity disorder. *Behav. Brain Res.* 234, 292–298. doi: 10.1016/j.bbr.2012.04.043

Zhao, Y.-Y., Li, J.-T., Wang, X.-D., Li, Y.-H., Huang, R.-H., Su, Y.-A., et al. (2013). Neonatal MK-801 treatment differentially alters the effect of adolescent or adult MK-801 challenge on locomotion and PPI in male and female rats. *J. Psychopharmacol.* 27, 845–853. doi: 10.1177/0269881113497613

Zou, D., Huang, J., Wu, X., and Li, L. (2007). Metabotropic glutamate subtype 5 receptors modulate fear-conditioning induced enhancement of prepulse inhibition in rats. *Neuropharmacology* 52, 476–486. doi: 10.1016/j.neuropharm.2006.08.016

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

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

frontiers in Psychology

REVIEW
published: 16 November 2021
doi: 10.3389/fpsyg.2021.763432

Why Did the Earwitnesses to the John F. Kennedy Assassination Not Agree About the Location of the Gunman?

*Dennis McFadden**

Department of Psychology, Center for Perceptual Systems, University of Texas, Austin, TX, United States

Earwitnesses to the 1963 assassination of President John F. Kennedy (JFK) did not agree about the location of the gunman even though their judgments about the number and timing of the gunshots were reasonably consistent. Even earwitnesses at the same general location disagreed. An examination of the acoustics of supersonic bullets and the characteristics of human sound localization help explain the general disagreement about the origin of the gunshots. The key fact is that a shock wave produced by the supersonic bullet arrived prior to the muzzle blast for many earwitnesses, and the shock wave provides erroneous information about the origin of the gunshot. During the government's official re-enactment of the JFK assassination in 1978, expert observers were highly accurate in localizing the origin of gunshots taken from either of two locations, but their supplementary observations help explain the absence of a consensus among the earwitnesses to the assassination itself.

Keywords: JFK assassination, US House Select Committee on Assassinations (HSCA), re-enactment of JFK assassination, acoustics of rifle bullets, sound localization, earwitnesses, shock wave, muzzle blast

OPEN ACCESS

Edited by:
Howard Charles Nusbaum, University of Chicago, United States

Reviewed by:
Brian Gygi, Veterans Affairs Northern California Health Care System (VANCHCS), United States
Sverre Holm, University of Oslo, Norway

***Correspondence:**
Dennis McFadden
mcfadden@utexas.edu

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

Received: *23 August 2021*
Accepted: *14 October 2021*
Published: *16 November 2021*

Citation:
McFadden D (2021) Why Did the Earwitnesses to the John F. Kennedy Assassination Not Agree About the Location of the Gunman? Front. Psychol. 12:763432. doi: 10.3389/fpsyg.2021.763432

INTRODUCTION

President John F. Kennedy (JFK) was assassinated in Dallas, Texas, on 22 November 1963. Within days, the new president, Lyndon B. Johnson, appointed a blue-ribbon panel of seven legislators and statesmen to investigate the assassination. The commission was headed by the chief justice of the US Supreme Court, Earl Warren, and consisted of about 400 staff and a budget of about $10 million. The commission held public hearings and officially interviewed over 500 people. About 10 months after the assassination, the commission published a report of nearly 900 pages plus 26 volumes of interviews, depositions, and exhibits, all of which came to be called the Warren Report (Warren Commission, 1964). This was arguably the most thoroughly investigated murder in the history of the world.

The primary conclusions of the Warren Commission were: there was only one assassin, Lee H. Oswald, who shot from a corner window on the sixth floor of the Texas School Book Depository (TSBD) building, a location behind the president. Oswald fired three shots using the Mannlicher-Carcano rifle found in the TSBD, one shot missed and two struck the president. The second bullet struck the president in the upper back, exited from the front of his throat, and then struck Texas Governor John Connally, who was seated in front of, below, and inboard of the president (the single-bullet account). The third shot struck the president in the head, killing him within minutes. The commission noted that an 8-mm film taken by Abraham Zapruder greatly aided their investigation.

Even before the Warren Report was released, questions emerged about various people and events surrounding the assassination, and these led to a number of conspiracy theories about aspects of the JFK assassination. The Warren Commission emphasized that they found no evidence of any conspiracy between Oswald, organized crime, Cuba, the FBI, the CIA, the military-industrial complex, Jack Ruby (Oswald's assassin), or any other entities. Even so, conspiracy theories continued to thrive. In April 1968, Martin Luther King, Jr. (MLK) was assassinated in Memphis, Tennessee, and again there was widespread disbelief about several details of the official account of that event.

The persistence of public disbelief in the official accounts of the JFK and MLK assassinations led the US House of Representatives to create a House Select Committee on Assassinations (HSCA) in 1976 (Bugliosi, 2007a, p. 370+). The hope was that scientific advances made since the Warren Report and the MLK investigation would allow additional information to emerge that could help resolve some of the prevalent questions. The HSCA employed about 250 people, spent about $5.8 million, and issued its report in 1979. Their activities included a partial re-enactment of the assassination in Dealey Plaza in August 1978 using live ammunition. A primary motivation for the partial re-enactment was a dictabelt recording inadvertently transmitted from a motorcycle of the Dallas Police Department on the day of the assassination that might have contained the sounds of the gunshots (explained in McFadden, 2021).

The HSCA report (House Select Committee on Assassinations, 1979) received world-wide attention because at the last minute it was rewritten around a conclusion of there having been four gunshots, not the three mentioned in the Warren Report (Bugliosi, 2007a, p. 380; Bugliosi, 2007b, p. 153+). For various reasons, four shots meant two shooters, exactly what many conspiracy theorists had been arguing for years. As explained in **Supplementary Material** (McFadden, 2021), that conclusion was shown to be unquestionably erroneous within months of the publication of the HSCA report, but unfortunately the bell could not be unrung, and the HSCA report has contributed to the widespread belief that the Warren Report was wrong.

The HSCA's partial re-enactment of the assassination was primarily concerned with the physical acoustics of gunshots in Dealey Plaza (Barger et al., 1979), but also present was a psychoacoustics team, consisting of David M. Green, Frederic L. Wightman, and your author. The activities and the findings of the psychoacoustics team are primary components of this article (see below).

The assassination of President John Kennedy was unquestionably one of the major historical events of the 20th century, and unfortunately there are many erroneous claims and theories about the event scattered throughout the public record. Among the more serious errors is the conclusion by the HSCA that there were four gunshots taken at the President, not three. There are multiple goals for this article: to discuss the acoustics of rifle bullets and the psychoacoustics of localizing the source of rifle bullets in highly reverberant locations in order to help explain why earwitnesses to the JFK assassination were so uncertain about the location of the gunman; to retell the story of the HSCA partial re-enactment of the JFK assassination and to explain the mistake made by the physical acousticians; to provide some personal perspective about the partial re-enactment; and to provide an example of applying knowledge obtained in laboratory studies to real-world situations, in this case an important historical event. Some of the content of this article repeats publicly available information that is not generally known to recent generations of acousticians, psychoacousticians, or the citizenry. Specifically, the report from the psychoacoustics team to the HSCA (Green, 1978, p. 111–130; Green, 1979) contained a section describing the difficulties humans have localizing the origin of a supersonic rifle bullet in a highly reverberant space, along with the observations made by the psychoacoustics team during the re-enactment, discussions that are repeated and elaborated here and in **Supplementary Material** (McFadden, 2021). My belief is that these stories deserve retelling so that citizens are more knowledgeable about the facts of their history.

EARWITNESS REPORTS

Figure 1 illustrates the scene of the assassination, Dealey Plaza. The president's motorcade began at the Dallas airport (Love Field), and consisted of over a dozen vehicles carrying local, state, and national dignitaries, the press, and Secret Service agents. The JFK limousine was second in line, and it carried the President and his wife, Texas Governor Connally and his wife, and two Secret Service agents. The motorcade left downtown Dallas on Main Street, turned right on Houston Street for one block, and then turned left onto Elm Street in front of the TSBD. The assassin's first shot came soon after the turn onto Elm Street, at ∼12:30 p.m. Upon hearing the shots, the Secret Service agents accelerated, and the presidential limousine took Stemmons freeway to Parkland Hospital where the President was pronounced dead about 1:00 p.m. Vice-President Johnson was sworn in as president just prior to takeoff on Air Force 1, which then returned the presidential party to Washington, D.C.

Estimates are that somewhere between 500 and 700 people were in Dealey Plaza at the time of the assassination. Many were employed nearby and came out during their lunch hour to watch the motorcade pass by. Most of the spectators were arrayed along Houston Street, with fewer on the sidewalk and grass on either side of Elm Street.

Soon after the assassination, subsets of these witnesses were interviewed by various investigators. This was not a well-planned research study, so the interviews generally lacked consistency. However, they do provide the best subjective evidence available about the assassination. One of the better compilations of these interviews comes from the work of Thompson (1967, p. 25). He summarized the responses of 190 witnesses who either testified to the Warren Commission or spoke with a law-enforcement officer whose report reached the Commission's files. Fortunately, these interviews contained information about the general location of each witness, which is important because of the acoustic complexity of Dealey Plaza. Unfortunately, the interviews occurred at different times after the assassination, meaning that

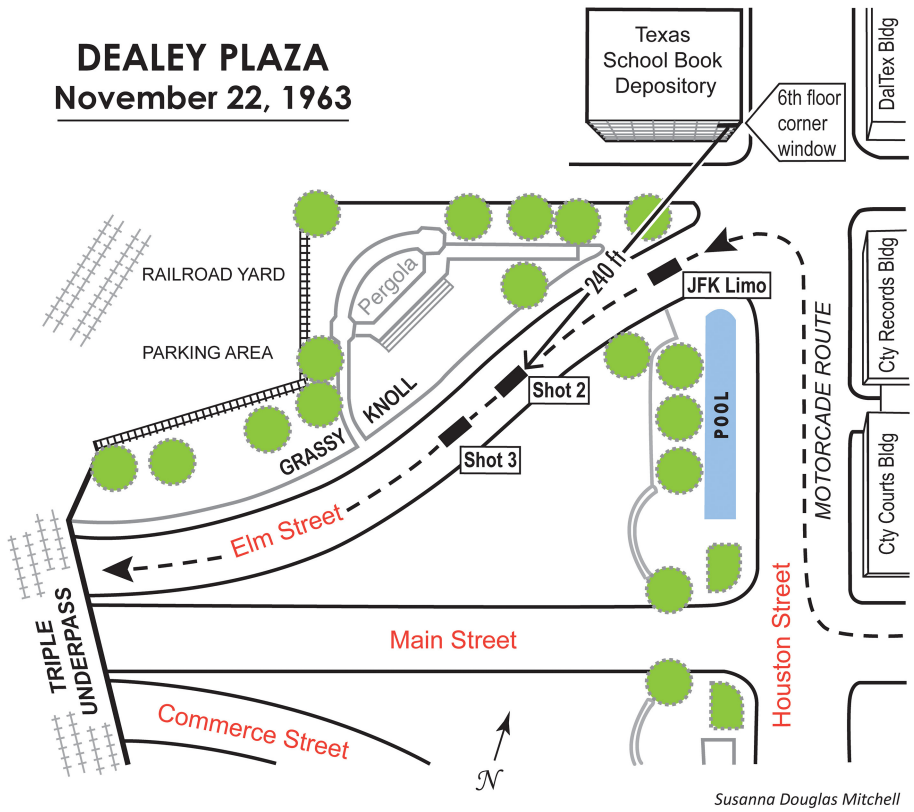

FIGURE 1 | Diagram of Dealey Plaza, showing the path of the motorcade (dashed line); the location of the assassin on the 6th floor of the TSBD; three locations of the presidential limousine corresponding roughly to its locations when the three shots were fired from the TSBD; the location of the pergola where A. Zapruder stood while filming the assassination; and other relevant details. Figure adapted from Green (1979) by Susanna Douglas Mitchell.

the earwitnesses had been exposed to differing numbers of newspaper and television reports about the investigation to date. Thus, their accounts may have been different from what they would have been if obtained immediately after the event. Also, all relevant questions about the gunshots apparently were not asked of every earwitness, or at least all the responses were not noted. Accordingly, caution is required not to over-interpret the results. It is important to understand that, unlike what would happen today, no commercial or personal audio equipment was operating in Dealey Plaza at the time of the assassination (and the Zapruder film was silent). All we have are earwitness reports plus the physical evidence.

The obvious needs to be said here, and acknowledged as important. Each of the earwitnesses to the JFK assassination was startled, surprised, confused, disbelieving, excited, and fearful, to varying degrees. Further, they brought to the event different previous experiences with gunshots. Finally, memories are known to change with time. All of these factors unquestionably contributed to the perceptions and memories reported, and that needs to be recognized when evaluating the verbal descriptions of the auditory (and visual) events provided by the earwitnesses. There should be no surprise that the recollections about details during those critical few seconds differed across individuals. As an extreme example of the differences across individual earwitnesses, Thompson reported that the descriptions of the time interval between the first and last shot varied from seconds to minutes! By contrast, there was good agreement about the number of shots heard, with 79% of the 172 people who answered that question saying 3 shots. Similarly, about 61% of the 65 people answering agreed that the final two shots were closer in time than were the first two shots (Thompson, 1967, p. 25, Appendix).

Although the earwitnesses were in reasonable agreement about the number and spacing of the gunshots, there was less consistency in their comments about arguably *the* most crucial question about the assassination: the origin of the gunshots. Only 64 of Thompson's 190 earwitnesses gave any opinion at all on that issue. That is, fully *two-thirds* of the earwitnesses apparently were too uncertain of the source of the gunshots to offer a location for the gunman (or maybe the answer was so obvious that it did not require comment?). This was a truly peculiar fact that deserved serious attention from the HSCA psychoacoustics team. What factors might have led so few earwitnesses to express an opinion about the origin of the gunshots when most of them did have opinions about the number and spacing of the shots?

By the way, of Thompson's earwitnesses who did mention a location, about 52% identified the grassy knoll or triple underpass region and about 39% mentioned the TSBD building. When evaluating various conspiracy theories, it is important to know

that, by one count, *only four* earwitnesses mentioned more than one location for gunmen (Green, 1979, p. 10; Bugliosi, 2007b, p. 174).

Having acknowledged the potential contribution of rapidly aroused emotion to the earwitness reports, are there other factors that could have contributed to the general uncertainty and disagreement the earwitnesses exhibited about the origin of the gunshots? Yes, but first some background.

LOCALIZING SOUND SOURCES

In less emotion-filled settings, humans are extraordinarily skilled at localizing the origins of sounds in three-dimensional space. As readers of this journal are aware, humans use three basic cues for sound localization. The direction of the source along the left/right axis is determined by the differences at the two ears in both the time of arrival of the sound and the sound-pressure level (SPL) of the sound, with spectral cues indicating source location in the vertical and front/back directions (Wightman and Kistler, 1992, 1999). The first cue (timing) exists because the speed of sound in air is relatively slow compared to the distance between our two ears; it takes more time for a sound to reach the further ear than the nearer ear. These time differences can be as large as 600–800 μsec, depending upon the size of the observer's head, but most humans are sensitive to interaural time differences as small as 10 μsec. The second cue (SPL) exists because the head throws a "shadow" when its width becomes large compared to the wavelength of the incident sound. At high frequencies, the magnitude of the interaural level difference can be tens of decibels, but most humans are sensitive to interaural differences of $<$1 dB. In the everyday world, interaural time and level cues typically operate in unison; the nearer ear typically also receives the stronger sound. The third cue of spectral differences originates in the acoustics of the pinnae; these cues are specific to the individual observer, and are essential for resolving confusions of front and rear source locations (Wightman and Kistler, 1999).

THE ACOUSTICS OF GUNSHOTS

Considerable insight into the earwitnesses' uncertainty about the location of the gunman is gained by examining the acoustics of rifle bullets. Gunshots are not like the stimuli commonly used in laboratories to measure human abilities to localize sounds. The bullets commonly fired from rifles travel at supersonic speed, meaning that there are two sources of sound associated with every rifle shot. First there is the *muzzle blast* that originates from the explosion inside the barrel and the exit of the bullet from the barrel (Beck et al., 2011). This sound behaves in a familiar manner. It propagates away from the rifle spherically, traveling at the speed of sound in air (∼1,125 feet per second or 343 meters per second), while losing about 6 dB of level for each doubling of the distance from the source.

The second source of sound for a supersonic bullet is a *shock wave* that is generated as the bullet moves through the air (Sapozhnikov et al., 2019). At the front of the bullet is a substantial overpressure produced because the surrounding air cannot move fast enough to "get out of the way." At the rear of the bullet is a substantial underpressure because the rapid flight of the bullet leaves a partial vacuum that takes time for the surrounding air to equalize. The overpressure at the front of the bullet and the underpressure at the rear of the bullet produce a rapid N-shaped pressure change that propagates away from the bullet. As we will see, the existence of these N waves has the potential to explain much of the uncertainty the earwitnesses had about the origin of the gunshots during the JFK assassination, so it is worthwhile to provide more details about N waves (also see Green, 1979; Sapozhnikov et al., 2019; McFadden, 2021).

For small objects like bullets, the time between the peak overpressure and the peak underpressure is so small that the two disturbances can be conceptualized as a single impulse. Once that N wave reaches a human observer, the perceptual experience commonly is described as a boom, snap, or crack. For large objects like supersonic aircraft, the peak overpressure and the peak underpressure of the N wave both can produce impulses, milliseconds apart (the double boom familiar to people living near military airbases). In general, the N waves produced by large objects have larger peak overpressures and larger peak underpressures than N waves from small objects.

Some additional details: (1) Technically, both the positive and negative peaks of the N wave are called shock waves, but for the purposes of this article "shock wave" typically will be used to denote the initial overpressure of the N wave. (2) The shock wave associated with the overpressure at the front of the bullet travels slightly faster than the speed of sound in the surrounding air, and the shock wave associated with the underpressure at the rear of the bullet travels slightly slower. Thus, the duration of the N-shaped wave increases with distance traveled from the source; the shock wave associated with the overpressure increasingly outraces the shock wave from the underpressure. (3) Unlike the familiar spherical pattern of propagation for the muzzle blast, the N wave propagates away from the supersonic bullet (or airplane) in a pattern having the shape of a cone (the Mach cone), with the tip of the cone at the front of the bullet. (4) In general, an N wave that spreads with the shape of a Mach cone loses about 4.5 dB of level for each doubling of distance from the source, as compared to the familiar −6 dB for the spherically spreading muzzle blast. This means that the *relative* strength of the N wave compared with the muzzle blast becomes greater with increasing distance from the rifle. **Figure 2** illustrates both the muzzle blast (curved lines) and the shock wave (cone shapes) from a supersonic bullet at two instants in time as the bullet travels from right to left across the figure (For an animation of an N wave, see https://en.wikipedia.org/wiki/Sonic_boom#/media/File:Sonicboom_animation.gif).

One important point to note from **Figure 2** (adapted from Green, 1979) is that the curved and conical wavefronts provide localization cues for different points of origin of the sounds. The curved, muzzle-blast waves did originate at the location of the gunman, and they carry values of interaural time difference and interaural level difference appropriate for that location. By contrast, the conical patterns originated from *the path of the bullet*, and they carry values of interaural time difference and

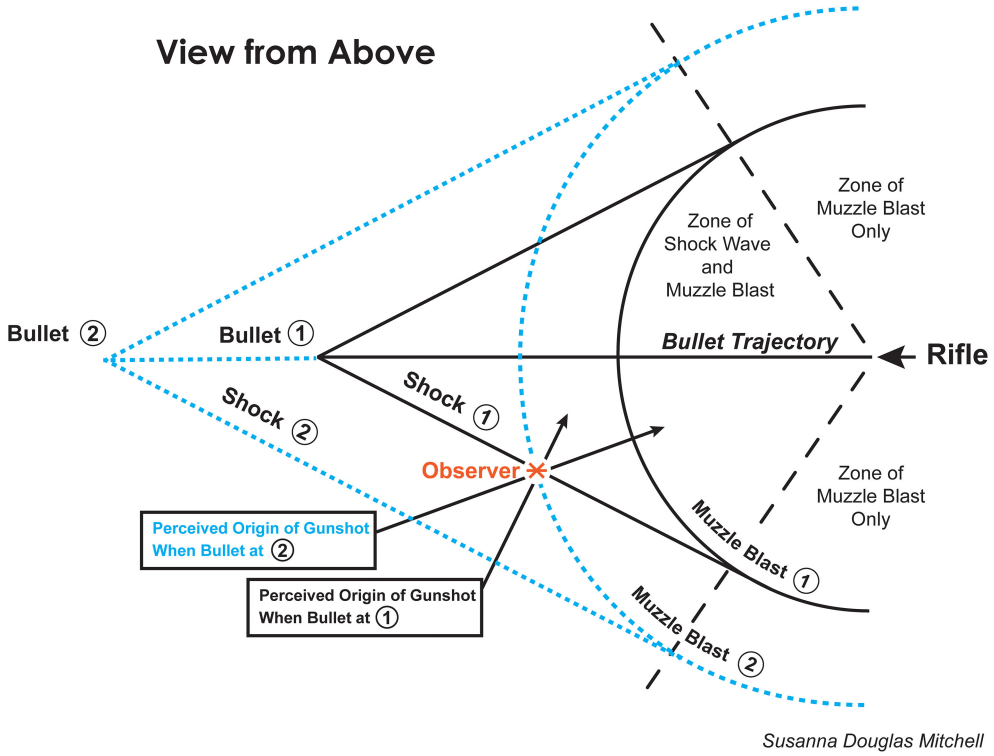

FIGURE 2 | A diagram of the muzzle blast and the shock wave associated with a gunshot from a rifle firing supersonic bullets. The rifle is located at the right of the figure and the bullet proceeds toward the left. The curved lines in the figure illustrate the muzzle blast propagating spherically away from the rifle, and the cone-shapes illustrate the path traversed by the shock wave propagating away from the path of the bullet. Because the bullet travels faster than the speed of sound, the sound produced by the shock wave arrives prior to the sound produced by the muzzle blast at most locations in front of the gunman. When the bullet reaches position 1, an ideal observer located at the red asterisk would hear the sound generated by the shock wave and would localize the source as the path of the bullet. This judgment is correct based on the stimulus, but incorrect about the origin of the bullet. When the bullet reaches position 2, an ideal observer located at the red asterisk now hears the muzzle blast and localizes that source as the location of the gunman. Of course, humans are not ideal at separating the two sources of sound, so the potential for confusion is high (The duration of the N wave increases moving from left to right along the edges of a cone; see text). Figure adapted from Green (1979) by Susanna Douglas Mitchell.

interaural level difference appropriate for the location of that path, not the location of the rifle (see **Figure 2**).

The second important point to note from **Figure 2** is that, at all locations more than a few meters in front of the rifle, the sound from the shock wave arrives *before* the sound from the muzzle blast. Also, the time difference between the two increases with increasing distance from the source (the ratio of the time differences is constant). At a distance of about 240 feet (about 73 meters; the approximate distance between the TSBD and the president at the time of the shot that hit him in the neck), the sound of the shock wave from a bullet from a Mannlicher-Carcano rifle will precede the muzzle blast by about 100 ms. The N wave and muzzle blast for a Mannlicher-Carcano rifle are shown in McFadden (2021) (**Supplementary Figure 1**).

ACOUSTICAL REASONS FOR CONFUSION

To repeat, the muzzle blast, like all other everyday sounds, gives rise to values of interaural time and level difference and spectral cues that an observer can use to accurately localize the source of that sound. However, the location cues extracted from the earlier-arriving shock wave are informative only about *the path of the bullet*, not the origin of the bullet. Thus, observers presuming that the interaural information extracted from the first-arriving sound at their location was informative about the location of the rifle would be led to compute an erroneous location for the gunman. In the case of Dealey Plaza, that computed location could even be logically absurd.

Consider an observer standing on the sidewalk along Elm Street between the TSBD and the grassy knoll, in front of the pergola and facing Elm Street (see **Figures 1**, **2**). The location cues extracted only from the first-arriving shock wave would lead to a conclusion that the rifle was slightly overhead, somewhere in the sky above the buildings in the distance to the southeast. The lifetime of knowledge a human observer brings to such a situation runs contrary to such an impossible conclusion, and confusion would be the result. Also, soon after the shock wave, the muzzle blast would arrive from the left carrying quite contradictory information about the source of the sound. The potential for confusion was high [Abraham Zapruder was filming from in front of the pergola, and at one time he said that the shots seemed to come from behind him (which the TSBD was not), although later he said he could not be certain about their origin (Bugliosi, 2007b, p. 169).].

In an acoustic environment like Dealey Plaza, there also are reflections of both the muzzle blast and shock wave off the various hard surfaces in the vicinity (Barger et al., 1979, Figure 2; Beck et al., 2011). These reflections reach each observer in rapid succession, with a temporal pattern that will depend upon the location of the observer. Note that the location cues associated with each reflection also are not informative about the initial source of the sounds: the rifle. Rather, they are informative only about the *origin of that particular reflection*. The shock wave and muzzle blast are spectrally and temporally different sounds; by contrast, reflections are repetitions of the same sound after it has bounced off nearby surfaces.

The time difference between the shock wave and the muzzle blast, and the timing of the various reflections depend upon the distance and location of the observer in relation to the gunman, meaning that the auditory experience of each observer in Dealey Plaza was different. With each gunshot, every observer was exposed to a complex pattern of successive sounds, all carrying location information, but only one of those incident wavefronts, the direct path from the muzzle blast, carried *correct* information about the initial source of all those sounds: the rifle.

This analysis of the acoustics of the events in Dealey Plaza helps us understand why the earwitnesses to the JFK assassination were uncertain about the origin of the gunshots that day.

PSYCHOACOUSTICAL REASONS FOR CONFUSION

Auditory science knows much about how humans react to presentations of multiple sounds in rapid succession. One relevant phenomenon is known as the *precedence effect*. This is a binaural effect that operates over the course of *microseconds*. When two sounds from different locations in space occur in rapid succession, the location cues extracted from the first sound dominate the perception of the location of the fused sounds (Wallach et al., 1949; Litovsky et al., 1999). Also, architectural acousticians know that an auditorium having strong reflections that occur more than about 30 ms after the arrival of the direct sound will be perceived as "hollow," and with increasing time delays, echoes will be reported.

Another phenomenon known to every architectural acoustician operates over the course of tens of *milliseconds,* and does not depend upon binaural cues (McFadden, 1973). Consider a public-address system in a large room with a soloist performing before a microphone at the front of the room. If a loudspeaker at the rear of the room were energized immediately with the sound collected by the microphone, listeners at the rear of the room would receive the sound from that loudspeaker before receiving the direct sound from the soloist, and they would perceive the sound as originating from the loudspeaker, not from the front of the room. To eliminate this unnatural experience, public-address systems introduce a time delay between the sound picked up by the microphone and the electrical waveform delivered to the distant loudspeakers. When the delay is a few tens of milliseconds, the listeners close to the distant loudspeakers perceive the sound as originating from the (distant) soloist rather than from the loudspeaker located nearby. This illusion exists even when the direct sound is tens of decibels weaker than the amplified sound from the loudspeaker; also, this illusion exists for monaural listeners as well as binaural listeners. This effect could lead earwitnesses to emphasize the shock wave over the muzzle blast and thus reach an erroneous conclusion about the origin of the gunshots.

Another factor relevant to understanding the varied perceptions of the earwitnesses about the origin of the gunman is "*front/back*" *confusions*. The geometry of sound localization using interaural time differences is such that there are inherent ambiguities about the source of the sound. For every value of interaural time difference, there always are a large number of locations in three-dimensional space from which the sound could have originated. For example (to the extent that the head is a sphere), one sound source located at 30 degrees to the right of the listener and another located at 150 degrees to the right both give rise to the same value of interaural time difference. Indeed, the locus of all locations possibly giving rise to the value of time difference in this example is described by a cone having its apex located at the right ear canal and its base off to the right of the head. These loci often are called cones of confusion. In everyday environments almost all relevant sounds originate from sources on the ground (essentially at "ear level"), so most of that cone can be ignored, and people typically do. However, the front/back ambiguities still exist.

When a sound is ongoing, small head movements can resolve the front/back confusion (Wightman and Kistler, 1999), but rifles make impulsive sounds. It is believable that, at the time of the JFK assassination, some of the earwitnesses who localized the path of the bullet using the shock-wave information and then realized the gunman could not be suspended in the sky or in some other impossible location simply concluded instead that the sound came from behind them, a front/back reversal based on the shock wave (Green, 1979). These would not have been conscious decisions; each earwitness had a lifetime of subconscious experience with cones of confusion. Garinther (cited in Green, 1979, p. 5) observed front/back reversals about 25% of the time in listeners trying to localize gunshots whose muzzle blasts had been attenuated. Yost and Pastore (2019) reported similar percentages using much longer stimuli. As noted, A. Zapruder initially reported that the gunshots seemed to originate from behind him (Bugliosi, 2007b, p. 169).

All of these various psychoacoustical effects reveal that a first-arriving sound is given more emphasis by the human auditory brain than are any later-arriving sounds. This makes sense from the perspective of evolution because in most real-world settings the direct sound arrives prior to its reflections, which carry erroneous localization information. However, this characteristic of human auditory perception can lead to errors of localization in certain unusual situations, and gunshots in reverberant environments are one of those situations.

Taken together, then, a consideration of the physical acoustics of rifle bullets and the psychoacoustics of sound localization helps explain the uncertainties and inconsistencies among the earwitnesses to the JFK assassination (Green, 1979).

THE PARTIAL RE-ENACTMENT

Two teams of investigators were involved in the partial re-enactment of the JFK assassination in Dealey Plaza: physical acousticians and psychoacousticians. The physical acousticians were there to make high-quality recordings using an array of microphones while gunshots were taken at sandbags located at positions known to be relevant. The goal was to determine whether any of the sounds recorded during the re-enactment could be matched to any of the impulse patterns on the police dictabelt, thereby demonstrating that the motorcycle with the stuck Transmit button was in Dealey Plaza at the time of the assassination. If it was, there was a possibility that some additional knowledge could be obtained about the number, origin, and path of the bullets fired during the assassination.

The psychoacousticians were at the re-enactment because the staff of the HSCA believed that having some "expert listeners" present during the re-enactment might provide explanations for some of the contradictory reports of the earwitnesses present at the assassination. As noted, some earwitnesses were adamant that all the gunshots came from the TSBD, some were convinced that shots came from the vicinity of the grassy knoll, and some gave other reports. These discrepancies contributed to the breadth and persistence of the conspiracy theories that had emerged since the assassination. The hope was that some of these discrepancies would be better understood after the re-enactment. The activities and findings of the psychoacoustics team were described to the HSCA by Green (1978, 1979); additional details are in **Supplementary Material** (McFadden, 2021).

During the partial re-enactment in 1978, marksmen shot either from the sixth-floor window of the TSBD or from behind a fence on the so-called grassy knoll commonly viewed as a possible location for a second gunman (see **Figure 1**). There were four sandbag targets, and several sequences of shots were made as the 12-microphone array was positioned in three different locations. The positions of the microphones are shown in McFadden (2021) (**Supplementary Figure 2**). Mannlicher-Carcano rifles like the one Oswald left in the TSBD were used along with a pistol fired only from behind the fence on the grassy knoll. During those sequences of shots, the psychoacousticians were located together or separately at different positions in Dealey Plaza. Our tasks were to indicate the perceived origin of each shot (forced choice: TSBD or grassy knoll) and to note additional aspects of each perceptual experience (the nature, number, direction, and duration of the echoes, etc.). The primary observers were Drs. Wightman (FLW) and McFadden (DM) (Dr. Green has a unilateral hearing loss.). Although observations were made from multiple locations, this was not a carefully controlled, counter-balanced study. As is true for most applied-science projects, we had to get a feel for the situation during the early gunshots, and our experiences led to some last-minute changes in the initial plan.

OBSERVATIONS DURING THE PARTIAL RE-ENACTMENT

The plan was for three sequences of 12 gunshots each (see Supplementary Table 1 in McFadden, 2021), but in the event, repeat shots were taken when the physical acousticians needed them. Observers FLW and DM were ignorant of the planned sequences, and we did not know the exact moment of each shot, but we were told when a shot was about to be taken. The individual shots were separated by differing intervals, but the average was 1–2 min. Including the repeats required by the physical acousticians, we observed 57 gunshots. Some locations gave rise to greater uncertainty about the origin of the shots than others, but overall, the two observers did quite well at localizing the source of the gunshots (see McFadden, 2021; **Supplementary Table 2**). Averaged over the sequences and locations, observer FLW was 86% correct at identifying the origin of the shots and DM was 93% correct. For most of the test shots we were not located together (Green, 1979, p. 14), and there was no comparison of responses and perceptions until the end of the re-enactment. To be sure, we faced a simpler task than the earwitnesses because we had far less uncertainty about the possible source of the shots, and there was no factor of surprise. The shot-by-shot field notes for both observers are in Appendix A in Green (1979).

One primary observation was that the gunshots were extremely loud from all our observer positions. They were so loud that we were frankly mystified by the reports from several earwitnesses of initially having thought they heard firecrackers. Those people must have had experience with firecrackers far larger than any the psychoacoustics team was allowed to play with. Admittedly, Dealey Plaza was much quieter during the re-enactment than on the day of the assassination, so there was less masking. HSCA staff intentionally did have three idling motorcycles to add some masking noise, but they contributed little.

Almost all of the gunshots during the re-enactment gave rise to sounds whose origins were diffuse, not narrowly focused or precise. No matter what our observer positions or the marksmen's target, our perceptions were of general *areas* for the origin of the gunshots, never anything as precise as the corner window on the 6th floor of the TSBD or the corner of the fence behind the grassy knoll. Indeed, from some observer positions the origin might appear off to the east of the TSBD or from the underpass down Elm Street, but our forced-choice decisions in those situations were TSBD and grassy knoll, respectively. [Some earwitnesses to the assassination also gave vague responses that did not specifically mention the TSBD or the grassy knoll, but for expediency Thompson (1967) encoded them as TSBD or knoll].

Although observers FLW and DM did quite well at identifying the source of the gunshots during the re-enactment, we also were impressed by the complexity of the acoustic perceptions aroused by the different shots. Multiple reflections from multiple directions often could be heard. In accord with expectation, the re-enactment rifle shots taken with the muzzle inside the window frame typically were noticeably less loud than those taken with the muzzle outside the window. Because this manipulation would affect the muzzle blast but have no effect on the shock wave, it reveals that much of the loudness percept was attributable to the later-arriving muzzle blast. Not surprisingly, at some observer positions, the perceptions changed noticeably when the marksmen changed targets, thereby changing the path of the bullet.

Perhaps most importantly, both primary observers were overwhelmed by those rifle shots originating from the grassy knoll. Those were *very* loud and unambiguous. We are convinced that had any rifle shots actually originated from the knoll area on the day of the assassination, the earwitnesses from that vicinity would have shown high confidence and high agreement about that fact. The fence on the grassy knoll was only a few meters to the right of the amateur photographer, A. Zapruder. Had a rifle shot actually originated from the grassy knoll, his startle response might well have knocked him sideways, off his perch on the pergola.

The pistol shots from the grassy knoll during the re-enactment were noticeably different from the rifle shots, being much less loud and much more narrowly focused than the rifle shots. This was evident to both observers from every observer position. We believe that earwitnesses would have been far more accurate and consistent in their reports had any subsonic pistol shots been fired at the motorcade.

In McFadden (2021) are some additional details about the methods of the psychoacoustics team, some detailed analyses of the earwitness reports from 1963 (**Supplementary Table 3**), and some comparisons between the perceptions of the psychoacoustics team and the earwitnesses.

SUMMARY

Localizing the origin of a supersonic gunshot is not easy under optimal conditions. On the day of the JFK assassination, the earwitnesses present were startled, surprised, confused, disbelieving, excited, and likely scared, so there is little wonder that their perceptions were inconsistent, and with the passage of time, fluid. Once the confusing acoustics of supersonic bullets and the vagaries of human sound localization are taken into account, the widespread uncertainty amongst the earwitnesses to the assassination becomes more understandable. The key point is that for many earwitnesses, the N wave arrived first, and it carried erroneous information about the location of the gunman. It is truly unfortunate that the physical acoustics measured during the partial re-enactment of the JFK assassination led to an official government report that was incorrect, but in the end, history received another sterling example of self-correction in science. Also, I believe that the psychoacoustics work done during the re-enactment was helpful in clarifying the earwitness testimony from the day of the assassination. Nominally "expert" listeners knowing that the gunshots could originate from only one of two locations were quite accurate in identifying the source. Furthermore, we identified some of the misperceptions that could arise for observers at various locations in Dealey Plaza. For those of us on the psychoacoustics team, the partial re-enactment was a welcome opportunity to apply our lab-based knowledge to a real-world problem of enduring international interest.

AUTHOR'S NOTE

The account presented here is the author's retelling of the report by Green (1978, 1979) plus historical information from before and after the HSCA partial re-enactment; no current entity of the US Government contributed to this material.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

The author did receive *per diem* from the US government for the time spent at the 1978 partial re-enactment.

ACKNOWLEDGMENTS

Many thanks to David Blackstock and Mark Hamilton for encouraging me to prepare a written version of a talk i gave on this topic to the UT Acoustics group. Dr. Hamilton also kindly helped get my description of shock waves correct. David Green and Frederic Wightman provided helpful comments about early versions of this retelling of events.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://doi.org/10.15781/tg6v-h395

REFERENCES

Barger, J.E., Robinson, S.P., Schmidt, E.C., and Wolf, J.J. (1979). *Analysis of recorded sounds relating to the assassination of President John F. Kennedy. BBN report 3947*. Cambridge, MA: Bolt Beranek and Newman, Inc. Available online at: http://aarclibrary.org/publib/jfk/hsca/reportvols/vol8/pdf/HSCA_Vol8_AS_2_BBN.pdf (accessed October 29, 2021).

Beck, S.D., Nakasone, H., and Marr, K.W. (2011). Variations in recorded acoustic gunshot waveforms generated by small firearms. *J. Acoust. Soc. Am.* 129, 1748–1759. doi: 10.1121/1.3557045

Bugliosi, V. (2007a). *Reclaiming History: The Assassination of President John F. Kennedy*. New York, NY: Norton.

Bugliosi, V. (2007b). *Endnotes to Reclaiming History: The Assassination of President John F. Kennedy*. New York, NY: Norton. Available online at: https://archive.org/stream/ReclaimingHistoryEndnotes/Reclaiming%20History%20Endnotes_djvu.txt (accessed October 29, 2021).

Green, D.M. (1978). *Testimony before House Select Committee on Assassinations*. Available online at: https://aarclibrary.org/publib/jfk/hsca/reportvols/vol2/pdf/HSCA_Vol2_0911_7_Green.pdf (accessed October 29, 2021).

Green, D.M. (1979). *Analysis of earwitness reports relating to the assassination of President John F. Kennedy. BBN report 4034*. Cambridge, MA: Bolt Beranek and Newman, Inc. Available online at: https://historymatters.com/archive/jfk/hsca/reportvols/vol8/pdf/HSCA_Vol8_AS_3_Earwitness.pdf (accessed October 29, 2021).

House Select Committee on Assassinations (1979). *Final Report of the Select Committee on Assassinations, U.S. House of Representatives, Ninety-fifth Congress, Second Session, Summary of Findings and Recommendations. House Report 95-1828*. Washington, D.C.: Government Printing Office. Available online at: archives.gov/research/jfk/select-committee-report

Litovsky, R.Y., Colburn, H.S., Yost, W.A., and Guzman, S.J. (1999). The precedence effect. *J. Acoust. Soc. Am.* 106, 1633–1654. doi: 10.1121/1.427914

McFadden, D. (1973). Precedence effects and auditory cells with long characteristic delays. *J. Acoust. Soc. Am.* 54, 528–530. doi: 10.1121/1.1913611

McFadden, D. (2021). *Online Supplement to Why Did the Earwitnesses to the John F. Kennedy Assassination Not Agree About the Location of the Gunman?* doi: 10.15781/tg6v-h395

Sapozhnikov, O.A., Khokhlova, V.A., Cleveland, R.O., Blanc-Benon, P., and Hamilton, M.F. (2019). Nonlinear acoustics today. *Acoust. Today* 15, 55–64. doi: 10.1121/AT.2019.15.3.55

Thompson, J. (1967). *Six Seconds in Dallas: A Micro-Study of the Kennedy Assassination.* New York, NY: B. Geis.

Wallach, H., Newman, E. B., and Rosenzweig, M. R. (1949). The precedence effect in sound localization. *Am. J. Physiol.* 52, 315–336. doi: 10.2307/1418275

Warren Commission (1964). *Report of the President's Commission on the Assassination of President Kennedy*. Washington, D.C.: Government Printing Office. Available online at: https://www.archives.gov/research/jfk/warren-commission-report (accessed October 29, 2021).

Wightman, F.L., and Kistler, D.J. (1992). The dominant role of low-frequency interaural time differences in sound localization. *J. Acoust. Soc. Am.* 91, 1648–1661. doi: 10.1121/1.402445

Wightman, F. L., and Kistler, D. J. (1999). Resolution of front-back ambiguity in spatial hearing by istener and source movement. *J. Acoust. Soc. Am.* 105, 2841–2853. doi: 10.1121/1.426899

Yost, W.A., and Pastore, M.T. (2019). Individual listener differences in azimuthal front-back reversals. *J. Acoust. Soc. Am.* 146, 2709–2715. doi: 10.1121/1.5129555

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

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 28 January 2022
doi: 10.3389/fnins.2021.730744

Effects of Spatial Speech Presentation on Listener Response Strategy for Talker-Identification

Stefan Uhrig[1,2]*, *Andrew Perkis*[1], *Sebastian Möller*[2,3], *U. Peter Svensson*[1] and *Dawn M. Behne*[4]

[1] Department of Electronic Systems, Norwegian University of Science and Technology, Trondheim, Norway, [2] Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany, [3] Speech and Language Technology, German Research Center for Artificial Intelligence, Berlin, Germany, [4] Department of Psychology, Norwegian University of Science and Technology, Trondheim, Norway

This study investigates effects of spatial auditory cues on human listeners' response strategy for identifying two alternately active talkers ("turn-taking" listening scenario). Previous research has demonstrated subjective benefits of audio spatialization with regard to speech intelligibility and talker-identification effort. So far, the deliberate activation of specific perceptual and cognitive processes by listeners to optimize their task performance remained largely unexamined. Spoken sentences selected as stimuli were either clean or degraded due to background noise or bandpass filtering. Stimuli were presented via three horizontally positioned loudspeakers: In a non-spatial mode, both talkers were presented through a central loudspeaker; in a spatial mode, each talker was presented through the central or a talker-specific lateral loudspeaker. Participants identified talkers via speeded keypresses and afterwards provided subjective ratings (speech quality, speech intelligibility, voice similarity, talker-identification effort). In the spatial mode, presentations at lateral loudspeaker locations entailed quicker behavioral responses, which were significantly slower in comparison to a talker-localization task. Under clean speech, response times globally increased in the spatial vs. non-spatial mode (across all locations); these "response time switch costs," presumably being caused by repeated switching of spatial auditory attention between different locations, diminished under degraded speech. No significant effects of spatialization on subjective ratings were found. The results suggested that when listeners could utilize task-relevant auditory cues about talker location, they continued to rely on voice recognition instead of localization of talker sound sources as primary response strategy. Besides, the presence of speech degradations may have led to increased cognitive control, which in turn compensated for incurring response time switch costs.

Keywords: speech perception, spatial auditory cues, talker-identification, voice recognition, sound localization, response strategy, spatial auditory attention, switch costs

OPEN ACCESS

Edited by:
Howard Charles Nusbaum, University of Chicago, United States

Reviewed by:
Jaakko Kauramäki, University of Helsinki, Finland
Moïra-Phoebé Huet, Johns Hopkins University, United States

***Correspondence:**
Stefan Uhrig
stefan.uhrig@protonmail.com

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

Received: *25 June 2021*
Accepted: *13 December 2021*
Published: *28 January 2022*

Citation:
Uhrig S, Perkis A, Möller S, Svensson UP and Behne DM (2022) Effects of Spatial Speech Presentation on Listener Response Strategy for Talker-Identification. Front. Neurosci. 15:730744. doi: 10.3389/fnins.2021.730744

1. INTRODUCTION

The ability of the human auditory system for rapid extraction of spatial auditory cues is thought to facilitate perceptual and cognitive speech processing, especially under adverse and dynamic listening conditions (Zekveld et al., 2014; Koelewijn et al., 2015). Several practical attempts have been made to incorporate advantages of binaural hearing (Blauert, 1997) into the design of spatialized speech displays to counteract degradation factors in speech signal transmission, encompassing application domains like air traffic control (Brungart et al., 2002; Ericson et al., 2004) and audio teleconferencing (Kilgore et al., 2003; Blum et al., 2010; Raake et al., 2010).

Past research usually centered around listening situations involving multiple, simultaneously active talkers. Consequent challenges for auditory information processing are discussed as "cocktail party" problems (Bronkhorst, 2000, 2015), for instance, segregation and streaming of target and masker speech sounds, as well as segmentation and binding to form distinct objects within the auditory scene (Bregman, 1990; Ihlefeld and Shinn-Cunningham, 2008b; Shinn-Cunningham, 2008). The present study addresses another kind of listening situation in dyadic human–human conversation, namely when two talkers take turns in active speaking time (with silence gaps in between). Does auditory information cuing talker location affect behavioral talker-identification (TI) performance in this "turn-taking" listening scenario (Lin and Carlile, 2015, 2019)? If significant effects exist, what are their underlying perceptual and cognitive processes? To what extent are such effects dependent on speech degradations as well as impacting various attributes of subjective listening experience like perceived speech quality, speech intelligibility, or talker-identification effort?[1]

Audio spatialization techniques have been implemented to produce speech signals originating from physical (e.g., presentation through loudspeakers placed in the room) or virtual space (e.g., presentation through stereo headphones, based on head-related transfer functions), which proved advantageous in terms of speech (e.g., word, phrase/sentence) identification performance for simultaneous talkers, speech intelligibility, and listening effort (Yost et al., 1996; Ericson and McKinley, 1997; Nelson et al., 1999; Drullman and Bronkhorst, 2000; Ericson et al., 2004; Kidd et al., 2005b; McAnally and Martin, 2007; Allen et al., 2008; Ihlefeld and Shinn-Cunningham, 2008a,b; Koelewijn et al., 2015). Key influencing factors included the perceived location and relative sound level of talkers as well as listeners' prior knowledge about the task, that is, who will talk, when and where (Brungart et al., 2002; Ericson et al., 2004; Singh et al., 2008; Kitterick et al., 2010; Koelewijn et al., 2015). An in-depth study by Brungart et al. (2005) explored several configurations of auditory and visual cues in a simultaneous two-talker listening situation: presenting each talker through a different, spatially separate loudspeaker had by far the highest impact on word identification performance.

In applied settings, audio spatialization has been recommended as an important design feature of multi-talker speech displays that optimizes its effectiveness without having to attenuate or exclude non-target channels and losing potentially relevant information (Ericson et al., 2004). Related research work suggests strongest effects of spatial auditory information on TI performance if the number of talkers is high, talkers' voices are perceptually similar (e.g., due to same gender of talkers), and quality of transmitted speech is perceived as being low (e.g., due to limited transmission bandwidth) (Blum et al., 2010; Raake et al., 2010; Skowronek and Raake, 2015). Moreover, listening-only test scenarios have proven to be more sensitive to experimental manipulations of audio spatialization (and speech degradation) than conversational test scenarios (Skowronek and Raake, 2015).

Not sufficiently investigated, to date, is the question of how speech stimuli are internally processed by human listeners to achieve fast and accurate TI. Depending on available auditory cues, listeners might develop, combine, or switch between different strategies based on different perceptual and cognitive processes (Allen et al., 2008): During non-spatial speech presentation, TI would have to rely on the recognition of talkers' individual voice characteristics (Best et al., 2018); in the following sections, this cognitive process will be referred to as "voice recognition" (Latinus and Belin, 2011). During spatial speech presentation, TI could instead be based on the localization of active talker sound sources, given that associations between talkers and locations in auditory space are kept unique (Ihlefeld and Shinn-Cunningham, 2008b). As mentioned above, talker-specific spatial cues should become even more relevant when the voices of different talkers are unfamiliar and/or perceptually similar—the latter may also be a consequence of low speech transmission quality in technologically mediated listening situations (Wältermann et al., 2010). Furthermore, TI performance might be differentially affected by different types of speech degradation, like background noise or bandwidth limitation, that impose varying load on human perceptual and cognitive processing (Wickens, 2008) [as can be indicated via methods for continuous physiological recording like pupillometry (Zekveld et al., 2014; Koelewijn et al., 2015) or electroencephalography, EEG (Uhrig et al., 2019a,b)].

The present study deployed a loudspeaker-based test layout to examine listeners' response strategies for behavioral TI during non-spatial and spatial speech presentation modes. A simplified listening scenario with two talkers was realized, wherein only a single talker would be actively speaking at a time. It presumed undisturbed "turn-taking" (i.e., without any instances of talk-over or barge-in) in order to avoid the higher-order acoustic complexity and auditory processing demands of a "cocktail party" scenario involving simultaneous utterances by multiple talkers. Participants were instructed to quickly identify talkers by pushing associated response keys. Task conditions were devised to enable experimental isolation of different kinds of perceptual and cognitive processes (sound source localization, voice recognition).

[1]The authors of this article have introduced a detailed Quality of Experience (QoE) model of loudspeaker-based spatial speech presentation in Uhrig et al. (2020a). This QoE model specifies relations between objective degradation factors (quality elements) and subjective attributes of overall listening experience (quality features). It has been validated by analyzing subjective ratings collected during the listening test described in the present study.

2. MATERIALS AND METHODS

2.1. Participants

The 34 participants recruited for this study were native Norwegian listeners who reported normal hearing as well as normal or corrected-to-normal vision. Data collection, storage, and handling complied with guidelines through the Norwegian Centre for Research Data and with recommendations from the International Committee of Medical Journal Editors (ICMJE). All participants gave their informed consent and received an honorarium at the end of their test sessions, which lasted around one hour.

Due to technical problems during data collection, two participants had to be excluded, leaving a sample size of $N = 32$ (age: $M = 26.8$, $SD = 5.9$, $R = 19 - 44$ years; 11 female, 21 male; 5 left-handed, 27 right-handed) for further analysis.

2.2. Stimuli

The "NB Tale – Speech Database for Norwegian" provided the source stimulus material for the present study. Created by Lingit AS and made publicly available by the National Library of Norway[2], this database contains a module with audio recordings of sentences, manuscript-read by native talkers from several dialect areas in Norway. Two anonymous male talkers from the Oslo dialect area were chosen for the present study. Twenty sentences had been recorded per talker, which consisted of statements about various neutral topics (based on preceding subjective evaluation by the authors of this paper): Three sentences had the same semantic content for both talkers, whereas the other 17 sentences had different content for each talker. The sentences were of relatively long and varying duration ($M = 4.9$ s, $SD = 1.5$ s, $R = 2.1 - 8.0$ s). Presentation of longer, more complex, and variable speech stimuli was deemed a necessary precondition for establishing a more realistic listening situation.

To manipulate speech degradation, all stimuli were presented either as clean, noisy, or filtered versions. Hence, manipulations of two degradation factors, signal-to-noise ratio (SNR) and transmission bandwidth, should entail variation in perceptual quality dimensions of "noisiness" and "coloration," respectively (Wältermann et al., 2010). The influence of different types of background noise of varying stationarity and informational content on speech quality perception has been investigated before, including pink noise (Leman et al., 2008). Despite traditional telephone bandwidth ranging from 300 to 3,400 Hz (Fernández Gallardo et al., 2012, 2013), a narrower bandpass was chosen in order to provoke a strong enough perceived degradation intensity, similar to recent studies on perceptual discrimination of clean and filtered spoken words (Uhrig et al., 2019a,b). Using the "P.TCA toolbox" for MATLAB software (v. R2018a) (Köster et al., 2015), the clean sentence recordings were impaired along two perceptual dimensions of speech transmission quality (Wältermann et al., 2010) to create degraded stimuli:

- Impairment along the perceptual dimension of "noisiness" was induced by addition of pink noise, aiming at a target SNR of -5 dB, to create noisy stimuli.
- Impairment along the "coloration" dimension was induced by applying a bandpass Butterworth filter, with a low-cutoff frequency of 400 Hz and a high-cutoff frequency of 800 Hz, to create filtered stimuli.

As a final processing step, all 120 generated stimuli (40 clean, 40 noisy, 40 filtered) were normalized to an active speech level of −26 dBov (dBov: decibel relative to the overload point of the digital system) according to ITU-T Recommendation P.56 (2011).

To manipulate audio spatialization, stimuli were presented through different loudspeakers: In the non-spatial mode, stimuli for both talkers were presented through only a single central loudspeaker placed in front of the listener; in spatial modes, stimuli for each talker were presented through either the central or a talker-specific lateral (left, right) loudspeaker, thereby keeping mappings between lateral locations and talkers unique.

TABLE 1 | Experiment specifications.

Task	Presentation mode	Loudspeaker location	Cue	Strategy
Talker-identification	Non-spatial_id	Central	Vocal	Voice recognition
	Spatial_id	Central	Vocal	Voice recognition
		Left, Right	Vocal	Voice recognition
			Spatial	Sound localization
Talker-localization	Spatial_loc	Left, Right	Spatial	Sound localization

Participants carry out behavioral listening tasks (talker-identification, talker-localization) where speech presentation modes (non-spatial_id[entification], spatial_id[entification], spatial_loc[alization]) activate loudspeakers at three spatial locations (left, central, right). The availability of task-relevant auditory cues (vocal, spatial) constraints possible response strategies based on different perceptual and cognitive processes (voice recognition, sound localization).

2.3. Experimental Procedure

Test sessions comprised a talker-identification (TI) task and a talker-localization task, both of which were performed in a quiet, sound-attenuated laboratory room. The order of tasks (identification-localization, localization-identification) was randomized across participants. **Table 1** lists the experiment specifications (behavioral tasks, presentation modes, loudspeaker locations, available task-relevant auditory cues, and possible response strategies adopted by listeners).

At the beginning of a test session, participants gave their informed consent. They received a print-out information before each task (TI, talker-localization), which instructed them about the two different talkers, the task goals and proper usage of the subjective rating scales (see below). Afterwards, demographic data (age, gender, handedness, vision correction, known hearing problems) were collected.

As illustrated in **Figure 1**, participants were seated at a small table with a Cedrus RB-740 response pad and a standard computer mouse placed on it. They were facing an array of three Dynaudio BM6A loudspeakers, which were

[2]https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-31/

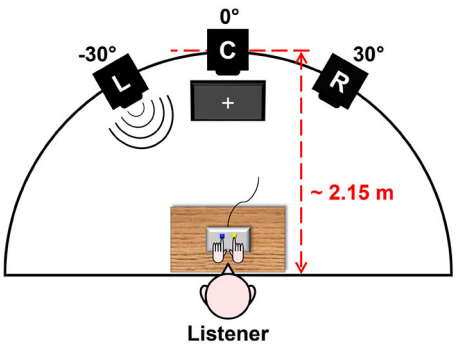

FIGURE 1 | Test layout deployed in the present study. The listener sits at a table, facing an array of three left (L), central (C), and right (R) loudspeakers (L = −30°, C = 0°, R = 30° azimuth) at a distance of approximately 2.15 m. The listener responds to stimuli by pressing keys on a response pad, while fixating a white cross displayed on a monitor screen below C. This figure originally appeared in Uhrig et al. (2020a), copyright 2020, with permission from IEEE.

equiangularly separated along the azimuthal direction and elevated approximately at the height of seated listeners' heads. On the floor below the central loudspeaker, a standard computer monitor was positioned. Participants put their left and right index fingers on the left, blue-colored and right, yellow-colored keys of the response pad. By pressing these two response keys, they were able to navigate through task instructions displayed on the monitor screen and respond to presented stimuli. In the task instructions, keys were always referred to by their arbitrarily assigned color ("blue" vs. "yellow") instead of their direction ("left" vs. "right") in order to avoid explicit associations of talker locations with keypress responses (Lu and Proctor, 1995). Over the ongoing stimulus presentation phase, participants were asked to fixate a white cross on the monitor to reduce contributions of orienting head movements to binaural hearing (Blauert, 1997).

The TI task was subdivided into six test blocks, one for each experimental condition (see Section 2.4). During a test block, all 40 stimuli (2 talkers × 20 sentences) of a certain *speech degradation* level were presented in series, with an inter-stimulus interval of 1,500 ms, adding a random jitter of either 0, +500, or −500 ms. Silence gaps between spoken sentences were longer than in usual conversational turn-taking [estimated to be around 300–350 ms for conversations in English, see (Lin and Carlile, 2015, 2019)] to ensure that participants would clearly notice when a trial had ended and be ready to respond in the upcoming trial (especially in same-talker trial transitions). To control for general learning and time-on-task effects, the order of conditions was randomized across participants. The order of stimuli was pseudo-randomized across blocks and participants such that only sentences with different semantic content were following each other within the trial sequence. Effects of varying acoustic form of degraded speech stimuli were distinguished from differences in semantic meaning by presenting all sentences in each experimental condition. This counterbalancing of sentence content across conditions is crucial given that variation in speech content has been shown to influence perceived speech quality (Raake, 2002).

In the TI task, participants were instructed to *identify* the current talker after each new stimulus as fast and accurately as possible, by pressing one of the two response keys[3]. In the non-spatial_id mode, stimuli for both talkers were serially presented through the central loudspeaker. In the spatial_id mode, stimuli for a particular talker were presented through the central or one talker-specific lateral (either the left or the right) loudspeaker, with equal probability (i.e., 50% each). Generally, probability of stimulus occurrence was 50% in the center, 25% at the left position, and 25% at the right position (i.e., in sum 50% at any lateral position). The reason for presenting talkers both at lateral positions and at the central position was to ensure that participants had to retain a task set of talker-*identification* throughout the spatial blocks of the TI task (see **Table 1**). All speech stimuli presented within a block were of the same *speech degradation* level, that is, either clean, noisy, or filtered.

Mappings between talkers and keypress responses for TI (blue vs. yellow) were randomized across blocks. However, in the spatial_id mode, the left and right loudspeaker location was always mapped onto the left (blue) and right (yellow) response key, respectively. Which talker would be presented at which lateral location (i.e., lateral location-talker mappings) was again randomized across blocks. Before starting a new block, participants needed to learn the current talker-response mappings. To accomplish this, they first had to listen (at least once) to a demo stimulus uttered by the particular talker to whom they would respond to with the blue key, and secondly had to listen (at least once) to a demo stimulus uttered by the other talker to respond to with the yellow key. Participants had the option to repeat each demo stimulus as often as they wanted to, for better talker voice memorization before continuing. The two demo stimuli were randomly selected from the current experimental condition and later presented again during the block, yet never occurred as the first stimulus for either talker in the trial sequence.

The talker-localization task consisted of a single test block involving a randomized sequence of clean stimuli, with each talker being randomly presented through the left or right loudspeaker (spatial_loc presentation mode, see **Table 1**). Participants were asked to indicate the perceived *location* of any active talker as left or right after each new stimulus, as fast and accurately as possible, via left or right keypresses. Thus, here they explicitly engaged in talker-*localization*, contrary to instructions for talker-*identification* in the TI task (see **Table 1**).

At the end of each block continuous rating scales were presented, one after the other, on the monitor screen (for scale

[3]A talker-identification task requires discrimination of talker-specific auditory cues, usually pertaining to individual talkers' voice characteristics, as well as retrieval of the (voice) identity category for the current talker from long-term memory (Latinus and Belin, 2011; Best et al., 2018); a talker-verification task requires detection of auditory cues signifying the presence/absence of a particular target talker, for example, as employed in a multi-talker scenario by Drullman and Bronkhorst (2000). It might be argued that identification of the two different talkers in the present study (TI task) is exchangeable for verification of only one of the two talkers, since either way the binary decision outcome would be the same. However, to achieve "true" TI, the identity of *each new talker* must be fully determined, whereas talker-verification is already completed after determining *the particular target talker* as present or absent. Thus, on a functional level, additional internal processes would be involved in the former case.

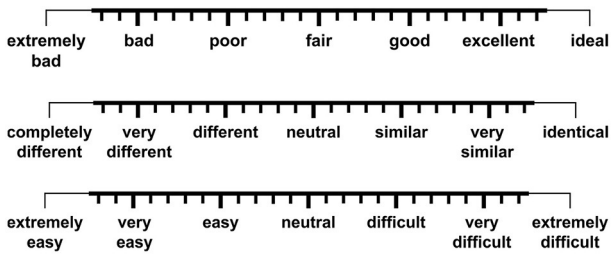

FIGURE 2 | Three continuous rating scales were employed for subjective assessment of perceived speech quality (top scale), speech intelligibility (top scale), voice similarity (middle scale), and talker-identification effort (bottom scale). A seven-point "extended continuous scale" design was implemented in accordance with ITU-T Recommendation P.851 (2003). Another version of this figure originally appeared in Uhrig et al. (2020a), copyright 2020, with permission from IEEE.

design and labelling, see **Figure 2**). These scales operationalized subjective constructs related to evaluative and task-related attributes of subjective listening experience (speech quality, speech intelligibility, voice similarity, TI effort) (Uhrig et al., 2020a). By using the computer mouse, participants could move a cursor along the full scale range and click at the scale position which in their opinion best described their subjective judgment. The order of scales was randomized across blocks and participants.

Stimulus presentation and data acquisition were executed by means of Psychophysics Toolbox Version 3 (PTB-3)[4] for MATLAB, running on a Windows-based computer in a control booth adjacent to the laboratory room. Speech files were played back using a high-quality audio interface (Roland UA-1610 Studio-Capture). A sound pressure level meter (Norsonic Nor150) was used to confirm approximately equal loudspeaker levels. The sound pressure level during stimulus presentation was adjusted to be around 65 dB at the listener position, ensuring a comfortable listening level.

2.4. Data Analysis

The present study followed a repeated-measures design with two fixed effects, *presentation mode* (non-spatial_id[entification], spatial_id[entification]), and *speech degradation* (clean, noisy, filtered), the full crossing of which resulted in six experimental conditions of the TI task (non-spatial_id/clean, non-spatial_id/noisy, non-spatial_id/filtered, spatial_id/clean, spatial_id/noisy, spatial_id/filtered). One additional *presentation mode* in the talker-localization task (spatial_loc[alization]) served as a talker-localization performance baseline for the ensuing behavioral data analyses (see Section 2.4.2). To further account for variability between participants and speech stimuli, two fully crossed random effects, *subject* (32 levels) and *stimulus* (120 levels), were included in the statistical models.

2.4.1. Rating

For statistical analysis of collected rating data, four repeated-measures analyses of variance (ANOVAs) were computed in R (v. 3.6.1) using the "ez"[5] package, with *presentation mode* (non-spatial_id, spatial_id) and *speech degradation* (clean, noisy, filtered) as within-subject factors, and rating for each subjective construct (speech quality, speech intelligibility, voice similarity, TI effort) as a dependent variable.

A statistical significance level of $\alpha = 0.05$ was chosen and Šidák-adjusted for four ANOVAs ($\alpha_{SID} = 0.013$). Generalized eta squared (η^2_G) was computed as an effect size measure. For *post-hoc* comparisons, paired *t*-tests with Holm correction were calculated.

2.4.2. Correct Response Time

To analyze correct response times, only trials with correct keypress responses faster than the 0.95-quantile of the raw response time data series (TI task: $RT_{0.95} = 1709$ ms; talker-localization task: $RT_{0.95} = 864$ ms) were included.

Using "lme4"[6] and "lmerTest"[7] packages, linear mixed-effects models (LMEMs) were fitted in R, similar to a previous study (Pals et al., 2015). To mitigate non-normality of their heavily right-skewed distributions, correct response times were log-transformed before entering the LMEMs. The "lmerTest" package provides *p*-values for statistical tests of fixed effects based on Satterthwaite's approximation method (which may give non-integer degrees of freedom) (Kuznetsova et al., 2017).

A statistical significance level of $\alpha = 0.05$, Šidák-adjusted for five LMEMs ($\alpha_{SID} = 0.010$), was assumed. For *post-hoc* analyses, general linear hypotheses with Holm correction were calculated using the "multcomp"[8] package.

2.4.2.1. Global Analyses of Presentation Mode

An initial LMEM was computed with *presentation mode* (non-spatial_id, spatial_id) and *speech degradation* (clean, noisy, filtered) as crossed fixed effects, *subject* (32 levels) and *stimulus* (40 levels) as crossed random effects (random intercepts), and correct response time as a dependent variable.

Including only trials from the spatial_id mode, another LMEM was calculated with *speech degradation* and *loudspeaker location* (left, central, right) as crossed fixed effects, *subject* and *stimulus* as crossed random effects (random intercepts), and correct response time as a dependent variable.

2.4.2.2. Local Analyses of Presentation Mode (Lateral/Central Loudspeaker Location)

Two follow-up analysis steps compared behavioral performance at either the center or lateral loudspeaker locations during the spatial_id mode, with spatial_loc and non-spatial_id modes as baselines:

A first analysis contrasted lateral locations (left, right) between the spatial_id mode (TI task) and the spatial_loc mode (talker-localization task). For this purpose, an LMEM was computed with *loudspeaker location* (left, right) and *presentation mode* (spatial_id, spatial_loc) as fixed effects, *subject* and *stimulus* as crossed random effects (random intercepts), and correct response time as a dependent variable.

[4]http://psychtoolbox.org/

[5]https://cran.r-project.org/web/packages/ez/

[6]https://cran.r-project.org/web/packages/lme4/

[7]https://cran.r-project.org/web/packages/lmerTest/

[8]https://cran.r-project.org/web/packages/multcomp/

A second analysis compared average response times between spatial_id and non-spatial_id modes at the central location. Thus, an LMEM was calculated with *presentation mode* (non-spatial_id, spatial_id) and *speech degradation* as fixed effects, *subject* and *stimulus* as crossed random effects (random intercepts), and correct response time as a dependent variable.

2.4.2.3. Analysis of Learning Effects (Within/Across Spatial Blocks)

In a final analysis step, learning effects speeding up behavioral responses within and across the three test blocks employing the spatial_id mode ("spatial blocks") were analyzed. Prior to trial selection, each spatial block was split into a first and a second half of trials that had presented stimuli through lateral loudspeakers ("lateral trial halfs").

An LMEM was computed with *loudspeaker location* (left, central, right), *spatial block* (spatial block 1, spatial block 2, spatial block 3) and *lateral trial half* (first trial half, second trial half) as crossed fixed effects, *subject* and *stimulus* as crossed random effects (random intercepts), and correct response time as a dependent variable.

2.4.3. Correct Response Rate

Statistical analyses of correct response rates were carried out using the same R packages described in Section 2.4.2 above. *Correct response rate* was defined as the number of correct keypress responses divided by the total number of keypress responses in a given test block.

A statistical significance level of $\alpha = 0.05$ was set and Šidák-adjusted for two LMEMs ($\alpha_{SID} = 0.025$). *Post-hoc* analyses involved general linear hypotheses with Holm correction.

2.4.3.1. Global Analysis of Presentation Mode

Selecting only trials from the spatial_id mode, an LMEM was computed with *speech degradation* (clean, noisy, filtered) and *loudspeaker location* (left, central, right) as crossed fixed effects, *subject* (32 levels) as a random effect (random intercept), and correct response rate as a dependent variable.

2.4.3.2. Local Analysis of Presentation Mode (Central Loudspeaker Location)

Focusing solely on the central loudspeaker location, spatial_id and non-spatial_id modes were compared against each other. An LMEM was calculated with *presentation mode* (non-spatial_id, spatial_id), and *speech degradation* as fixed effects, *subject* as a random effect (random intercept), and correct response rate as a dependent variable.

3. EXPECTATIONS

It was anticipated that the two manipulated factors, *presentation mode* (non-spatial_id, spatial_id) and *speech degradation* (clean, noisy, filtered), would influence evaluative and task-related attributes of overall listening experience (speech quality, speech intelligibility, voice similarity, TI effort) (Uhrig et al., 2020a), as well as exert effects on auditory information processing.

Participants should adapt their response strategy to the availability of auditory cues (Kidd et al., 2005a) that are relevant to effectively and efficiently solve their behavioral task goals of quick and accurate TI and talker-localization.

3.1. Presentation Mode

During the non-spatial_id mode (TI task), participants would be left only with vocal cues for TI based on voice recognition. However, during the spatial_id mode, performance improvements were expected to depend on *loudspeaker location* (left, central, right): At the central location, again participants would have to rely entirely on voice recognition for identifying different talkers. At lateral (left, right) locations, they would be able to base their response strategy on sound source localization (due to unique lateral location-talker mappings) and ignore the vocal cues, in order to generate quicker and more accurate behavioral responses. Subjectively, this effect should manifest as reduced TI effort for the spatial_id vs. non-spatial_id mode.

Previous studies demonstrated influences of specific target sentence content on word identification performance in multi-talker listening situations (Kidd et al., 2005a; Brungart and Simpson, 2007). In principle, initial portions of talker-specific sentences (see Section 2.2) could offer prosodic ["supra-segmental or phrase-prosody level"; (Fernández Gallardo et al., 2012)] and/or semantic cues (Darwin and Hukin, 2000) to facilitate decision-making about talker identity. Despite this possibility, prosodic/semantic evaluation of each newly occurring stimulus would involve later cognitive processing than rapid sound source localization or voice recognition, thus providing a less efficient alternative/additional response strategy.

Dynamic change in perceived talker location introduces uncertainty in listeners' spatial expectations, ultimately affecting behavioral task performance (Shinn-Cunningham, 2008; Zuanazzi and Noppeney, 2018, 2019). Research on simultaneous multi-talker listening situations demonstrated that prior knowledge about the probability of (change in) talker location leads to improvements in behavioral speech identification (Ericson et al., 2004; Kidd et al., 2005a; Brungart and Simpson, 2007; Singh et al., 2008; Koelewijn et al., 2015). Such improvements are explainable by an anticipatory shift of spatial auditory attention toward the most probable talker location, which in turn enhances information processing within this selected part of the auditory scene. Participants in the present study were not informed in advance about different talker locations or unique lateral location-talker mappings, nor did they receive any feedback on the correctness of their behavioral responses (Kidd et al., 2005a; Best et al., 2018). In spite of this, they would very likely acquire explicit or implicit knowledge about these regularities over the course of the experiment due to incidental statistical/covariation learning (Schuck et al., 2015; Gaschler et al., 2019): It was predicted that within and across spatial blocks participants would utilize talker location cues more and more often, leading to gradually faster behavioral responses at lateral locations.

3.2. Speech Degradation

In general, presence of speech degradations should impede TI due to obscuring of individual talkers' voice characteristics. Thus, presentations of degraded (noisy, filtered) speech stimuli would be anticipated to reduce perceived speech quality and speech

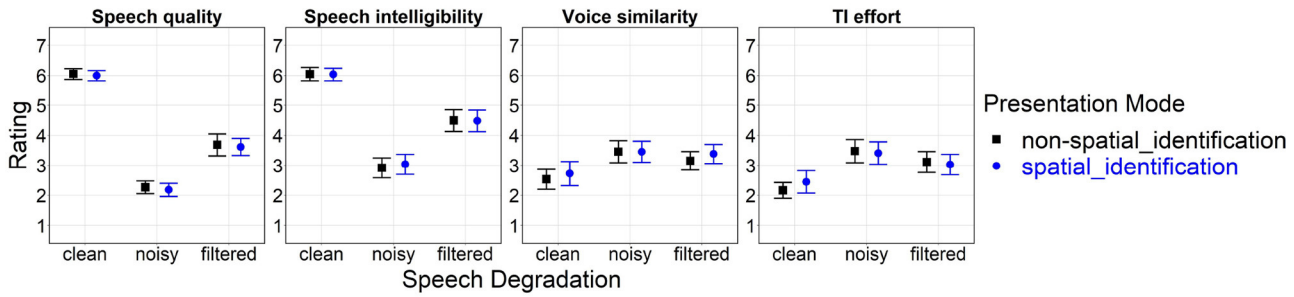

FIGURE 3 | Effects of *presentation mode* and *speech degradation* on rating for evaluative (speech quality, speech intelligibility) and task-related [voice similarity, talker-identification (TI) effort] attributes of overall listening experience. The numeric range of the y-axis (1–7) corresponds to scale labels shown in **Figure 2**. Error bars represent 95% confidence intervals. Another version of this figure originally appeared in Uhrig et al. (2020a), copyright 2020, with permission from IEEE.

intelligibility as well as increase voice similarity and TI effort relative to clean stimuli (Leman et al., 2008; Raake et al., 2010; Skowronek and Raake, 2015).

Fernández Gallardo et al. (2012, 2013, 2015) examined human TI performance for spoken words, sentences and paragraph-long speech when being transmitted through wideband and narrowband communication channels. The authors reported notably slower behavioral responses and reduced TI accuracies for narrowband vs. wideband. Studies on speech-in-noise perception established positive relationships between SNR and accuracy of identifying spoken syllables (Kaplan-Neeman et al., 2006), words (Sarampalis et al., 2009) and sentences (Pals et al., 2015), as well as negative relationships between SNR and behavioral response time for identification/recognition of syllables (Kaplan-Neeman et al., 2006), words (Mackersie et al., 1999; Sarampalis et al., 2009) and sentences (Gatehouse and Gordon, 1990; Baer et al., 1993; Houben et al., 2013; Pals et al., 2015); the same relationships were obtained for identification of words within target sentences, being presented concurrently with noise maskers at different SNRs (Ericson and McKinley, 1997; Brungart, 2001; Brungart et al., 2001, 2002; Ericson et al., 2004). Based on this previous evidence, it was expected that behavioral responses to identify single talkers would be generally delayed under degraded (noisy, filtered) vs. clean speech.

Furthermore, this effect of *speech degradation* should probably be more pronounced in the non-spatial_id vs. spatial_id mode, during which participants should rely exclusively on voice characteristics to decide about talker identity.

4. RESULTS

4.1. Rating

Statistically significant main effects of *speech degradation* resulted for all four subjective constructs: Speech quality [$F_{(2, 62)} = 357.69$, $p < 0.001$, $\eta^2_G = 0.83$], speech intelligibility [$F_{(2,62)} = 114.89$, $p < 0.001$, $\eta^2_G = 0.67$], voice similarity [$F_{(2, 62)} = 21.39$, $p < 0.001$, $\eta^2_G = 0.11$], and TI effort [$F_{(2, 62)} = 32.54$, $p < 0.001$, $\eta^2_G = 0.18$]. *Post-hoc* pairwise comparisons for clean vs. noisy and clean vs. filtered speech were significant for all constructs ($p < 0.001$); noisy vs. filtered speech showed lower speech quality ($p < 0.001$) and speech intelligibility ($p < 0.001$), and increased TI effort ($p < 0.01$). Neither the main effect of *presentation mode* nor the interaction between the two factors turned out to be statistically significant.

Figure 3 shows arithmetic mean values for effects of *presentation mode* (non-spatial_id, spatial_id) and *speech degradation* (clean, noisy, filtered) on rating for each subjective construct.

4.2. Correct Response Time

4.2.1. Global Analyses of Presentation Mode

The initial analysis delivered a significant main effect of *speech degradation* [$F_{(2, 7141.80)} = 288.71$, $p < 0.001$] and a significant interaction between *presentation mode* and *speech degradation* [$F_{(2, 7141.80)} = 12.87$, $p < 0.001$] on correct response time. *Post-hoc* comparisons of the spatial_id mode with the non-spatial_id mode revealed delayed responses for clean speech ($p < 0.001$), faster responses for noisy speech ($p < 0.001$), but no significant difference for filtered speech ($p = 0.22$).

Figure 4 shows arithmetic mean values for effects of *presentation mode* (non-spatial_id, spatial_id) and *speech degradation* (clean, noisy, filtered) on correct response time.

A follow-up analysis yielded significant main effects of *loudspeaker location* [$F_{(2, 3520.40)} = 90.56$, $p < 0.001$] and *speech degradation* [$F_{(2, 3499.90)} = 87.16$, $p < 0.001$] on correct response time. *Post-hoc* comparisons indicated slower responses at the center vs. lateral (left, right) locations (both $p < 0.001$), but no difference between the left and right location. Significant differences further occurred among all levels of *speech degradation* (all pairs $p < 0.001$): Generally, responses were fastest under clean speech, slower under filtered speech and slowest under noisy speech.

Figure 5 depicts effects of *speech degradation* and *loudspeaker location* (left, central, right) on correct response time for behavioral TI in the spatial_id mode; the figure also contains confidence ranges of the non-spatial_id mode (bars) and mean values of the spatial_loc mode (diamonds) to serve as baselines, in the latter case for behavioral talker-localization performance.

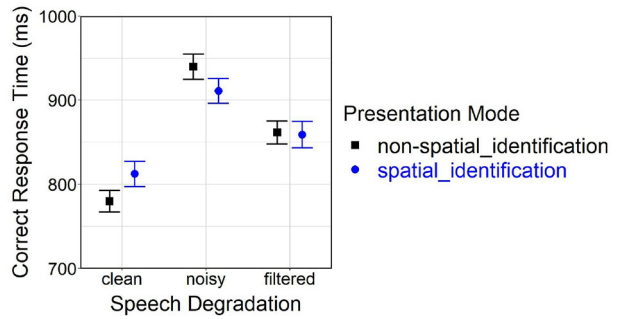

FIGURE 4 | Effects of *presentation mode* and *speech degradation* on correct response time. Error bars represent 95% confidence intervals. Another version of this figure originally appeared in Uhrig (2022), copyright 2021, with permission from Springer.

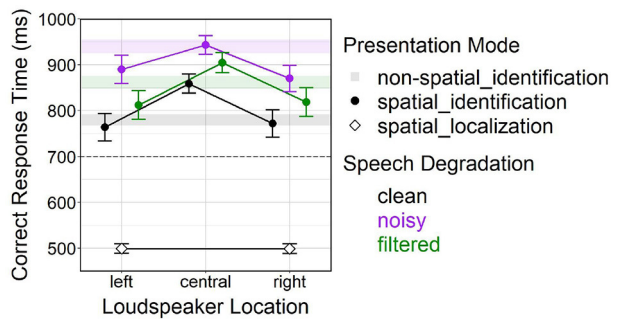

FIGURE 5 | Effects of *speech degradation* and *loudspeaker location* on correct response time in the spatial_id mode. Color-shaded bars represent 95% confidence ranges for the non-spatial_id mode under different *speech degradation* levels (i.e., black bar = non-spatial_id/clean, purple bar = non-spatial_id/noisy, green bar = non-spatial_id/filtered), as depicted in **Figure 4**. Diamond-shaped points represent the spatial_loc mode (talker-localization task). Error bars represent 95% confidence intervals. The dashed horizontal line at 700 ms marks the lower y-axis limit in **Figure 4**, for better comparability. Another version of this figure originally appeared in Uhrig (2022), copyright 2021, with permission from Springer.

4.2.2. Local Analyses of Presentation Mode (Lateral/Central Loudspeaker Location)

Two follow-up analyses examined behavioral performance in the spatial_id mode, with spatial_loc and non-spatial_id modes as baselines:

The first analysis confirmed a main effect of *presentation mode* [$F_{(1, 2964.60)} = 2780.25$, $p < 0.001$] on correct response time. Being plainly visible in **Figure 5**, much faster responses resulted in the spatial_loc vs. spatial_id mode at lateral (left, right) locations (averaged across all levels of *speech degradation*).

The second analysis found significant main effects of *presentation mode* [$F_{(1, 5324.80)} = 60.75$, $p < 0.001$] and *speech degradation* [$F_{(2, 5327.20)} = 182.01$, $p < 0.001$] on correct response time, as well as a significant interaction [$F_{(2, 5326.20)} = 16.29$, $p < 0.001$]. *Post-hoc* comparisons revealed significant differences between all speech degradation levels (all pairwise comparisons $p < 0.001$). Moreover, at the central location, responses were slower in the spatial_id vs. non-spatial_id mode under clean and filtered speech (both $p < 0.001$), but not under noisy speech ($p = 0.56$).

4.2.3. Analysis of Learning Effects (Within/Across Spatial Blocks)

A final analysis showed significant main effects of *loudspeaker location* [$F_{(2, 3511.80)} = 90.26$, $p < 0.001$], *lateral trial half* [$F_{(1, 3499.50)} = 12.64$, $p < 0.001$], and *spatial block* [$F_{(2, 3491.50)} = 55.18$, $p < 0.001$] on correct response time. *Post-hoc* comparisons were again significant between the central location and the lateral (left, right) locations (both $p < 0.001$). A gradual reduction in correct response time was observable both within spatial blocks (first lateral trial half: $M = 872.92$ ms, $SD = 279.24$ ms vs. second: $M = 848.40$ ms, $SD = 260.62$ ms) and across spatial blocks (spatial block 1: $M = 898.57$ ms, $SD = 280.14$ ms vs. 2: $M = 871.14$ ms, $SD = 269.30$ ms vs. 3: $M = 812.74$ ms, $SD = 253.69$ ms; all pairs $p < 0.001$).

Figure 6 depicts correct response time as a function of *loudspeaker location*, being partitioned into subplots by factor level combinations of *spatial block* and *lateral trial half* to illustrate the temporal development over the succession of spatial blocks.

4.3. Correct Response Rate

4.3.1. Global Analyses of Presentation Mode

The initial analysis resulted in a significant main effect of *loudspeaker location* [$F_{(2, 248)} = 11.76$, $p < 0.001$] for correct response rate, also revealing a significant interaction with *speech degradation* [$F_{(4, 248)} = 3.87$, $p < 0.01$]. *Post-hoc* comparisons suggested reduced correct response rates at the central location vs. the lateral (left, right) locations (both $p < 0.001$). The difference between the central location and lateral locations was emerging under clean speech (both $p < 0.001$), but not under degraded (noisy, filtered) speech. This pattern of results is supported by pairwise contrasts between clean and degraded speech that were only significant at the center (clean vs. noisy, $p < 0.001$; clean vs. filtered, $p = 0.013$).

Figure 7 depicts effects of *speech degradation* and *loudspeaker location* (left, central, right) on correct response rate in the spatial_id mode. The figure further contains mean values in the non-spatial_id mode at the central location (open squares), to aid visual inspection.

4.3.2. Local Analysis of Presentation Mode (Central Loudspeaker Location)

Analyzing only the central location trials demonstrated a significant main effect of *presentation mode* [$F_{(1, 155)} = 8.13$, $p < 0.01$] on correct response rate, and a significant interaction with *speech degradation* [$F_{(2, 155)} = 6.00$, $p < 0.01$]. *Post-hoc* comparisons suggested that average correct response rate was lower for spatial_id vs. non-spatial_id solely for clean speech ($p < 0.001$), as can be seen in **Figure 7**.

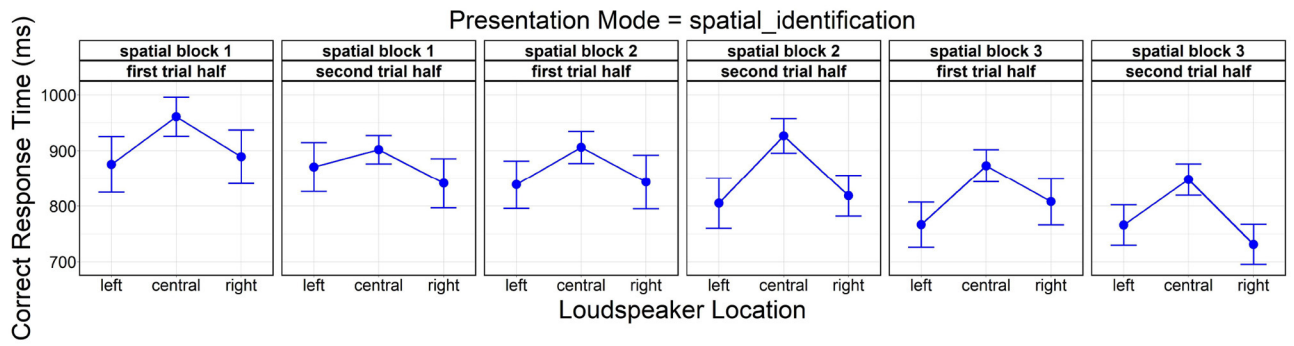

FIGURE 6 | Effects of *loudspeaker location*, *spatial block* (spatial block 1, spatial block 2, spatial block 3), and *lateral trial half* (first trial half, second trial half) on correct response time. Error bars represent 95% confidence intervals. Another version of this figure originally appeared in Uhrig (2022), copyright 2021, with permission from Springer.

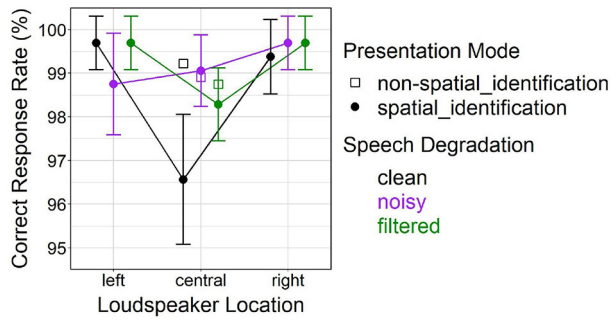

FIGURE 7 | Effects of *presentation mode* and *speech degradation* on correct response rate in the spatial_id mode. Open square-shaped points represent the non-spatial_id mode only (involving presentations only at the central loudspeaker location). Error bars represent 95% confidence intervals.

5. DISCUSSION

5.1. Rating

The rating analysis verified large-sized effects of *speech degradation* on perceived speech quality and speech intelligibility. Regarding the speech degradations induced in the present study, noisy speech impacted those subjective constructs more strongly than filtered speech, when compared to the clean speech reference; this could be attributed to variation in perceived degradation intensity due to different magnitude scaling of each degradation type. On average, the two talkers' voices were perceived to be "different"/"very different" for clean speech, yet slightly more similar for degraded (noisy, filtered) speech, presumably as individual talkers' voice characteristics were masked by added background noise and spectral details were removed by bandpass filtering.

The changes in voice similarity closely corresponded with experienced difficulty of TI, being "easy"/"very easy" under clean speech, but increasing to a small extent under degraded speech (Raake et al., 2010; Zekveld et al., 2014; Skowronek and Raake, 2015). Task difficulty should most probably depend on the amount of allocated information processing resources [i.e., the perceptual-cognitive load (Wickens, 2008); measurable, e.g., by pupillometry or EEG] to discriminate between the two talkers' voices, which was higher for degraded vs. clean speech; better talker voice discriminability would in turn ease TI based on voice recognition. Altogether, experimental manipulation of *speech degradation* could be considered successful, with a weak but nonetheless perceptible impact on TI effort.

No effects of *presentation mode* on any subjective construct turned out to be statistically significant. This stands in direct contrast to a number of previous studies examining "cocktail party" contexts (Raake et al., 2010; Koelewijn et al., 2015; Skowronek and Raake, 2015), which had reported higher perceived speech quality and speech intelligibility as well as reduced talker/speech identification effort following spatial speech presentation. The assessed subjective constructs appeared to be independent of spatialization, at least within the realized, probably less complex "turn-taking" listening scenario. Participants may not have been able to fully exploit talker location cues in the spatial_id mode of the TI task. Possible reasons for this include a lack of prior knowledge about the underlying spatial regularities and/or not having enough exposure time over each spatial block for learning these regularities well enough to noticeably speed up decisions about talker identity. In the future, gathering post-experiment feedback from participants might prove useful to gain better insight into whether they were actually aware of any spatial regularities and intentionally adopted (or refrained from adopting) an alternative localization-based response strategy.

Only half of the trials during the spatial_id mode involved stimulus presentations at talker-specific lateral loudspeaker locations. Participants might have been confused by the random talker location changes between the center and lateral locations, which could have counteracted any advantages of extracted spatial regularities (Uhrig et al., 2020a). Because the spatial_id mode would still demand relatively more effortful voice recognition in half of the trials—namely, when stimuli were presented at the central location—it remains unclear as to whether participants actually experienced a significant overall

reduction in task difficulty. Furthermore, with the TI task consistently being judged as "easy," even under degraded speech, behavioral performance improvements might not have been large enough to be reflected in the subjective ratings, hence constituting a ceiling effect (Best et al., 2018; Uhrig et al., 2020a).

5.2. Behavioral Measures (Correct Response Time, Correct Response Rate)

5.2.1. Global Analyses of Presentation Mode

In the initial analysis, significant main effects as well as a significant interaction between *speech degradation* and *presentation mode* on correct response time were observed. Behavioral TI took longer for degraded vs. clean speech, the response delay being more pronounced for noisy vs. filtered speech, which corresponded well with the subjective rating results (see Section 5.1) and past findings (Uhrig et al., 2019a,b). However, contradicting original predictions formulated in Section 3, behavioral responses in the spatial_id vs. non-spatial_id mode were slower for clean, faster for noisy and stayed the same for filtered speech (see **Figure 4**). This unexpected result pattern could not be easily interpreted without taking the additional factor *loudspeaker location* (left, center, right) into account.

Subsequent analysis consistently showed faster behavioral responses at lateral locations relative to the central location, across all levels of *speech degradation* (see approximately equal slopes of "reverse-V-shaped" lines connecting points within each level of *speech degradation* in **Figure 5**). In general, behavioral responses were delayed for degraded (noisy, filtered) vs. clean speech. Average correct response rates were very high (around 98–100 %, see **Figure 7**), remaining constant at lateral locations across the three *speech degradation* levels, only slightly dropping at the center (relative to lateral locations) under clean speech. Interestingly, although at lateral locations the presence of speech degradation had caused a temporal delay in behavioral responses, response accuracy at lateral locations remained unaffected by speech degradation; this fact might be attributable to some form of facilitation of early perceptual and/or late response-related processing, which will be further discussed in Section 5.2.2 below.

5.2.2. Local Analyses of Presentation Mode (Lateral/Central Loudspeaker Location)

The first analysis affirmed behavioral responses at lateral loudspeaker locations to be drastically slower for TI in the spatial_id mode than for talker-localization in the spatial_loc mode (see large deviations between circle- and diamond-shaped points for spatial_id and spatial_loc modes at lateral locations in **Figure 5**). Therefore, faster behavioral responses at lateral locations (relative to the central location) during the spatial_id mode could not be explained by a spontaneous, full strategy change from voice recognition to sound source localization over the course of a spatial block (Allen et al., 2008; Gaschler et al., 2019). Rather, TI based on voice recognition remained the primary response strategy, but was somehow improved by the additional spatial auditory information. This behavioral response facilitation might possibly originate from automatic, preattentive mechanisms at earlier stages of the auditory processing hierarchy—for instance, improved auditory streaming (Bregman, 1990; Ihlefeld and Shinn-Cunningham, 2008b; Shinn-Cunningham, 2008)—such that the total time needed to reach a behavioral decision on talker identity was slightly shortened.

During all spatial blocks, responses to the talker occurring at the left location corresponded with the left response key and responses to the talker occurring at the right location corresponded with the right response key (see Section 2.3). This constraint controlled for the so called *Simon effect*, describing the phenomenon that behavioral responses are executed faster when located on the same side as their associated stimuli (i.e., being *spatially compatible*; e.g., responding to a stimulus on the left side with the left key) than on the opposite side (i.e., being *spatially incompatible*; e.g., responding to a stimulus on the left side with the right key) (Simon, 1969; Lu and Proctor, 1995). The rationale behind realizing only spatially compatible location-response mappings in the TI and talker-localization tasks was to avoid participant confusion (by spatially incompatible location-response mappings) and instead enable natural directional response tendencies "toward the source of stimulation" (Simon, 1969), hereby improving ecological validity. However, this inevitably led to a confounding influence of *stimulus-response compatibility*: Faster behavioral responses at lateral locations, instead of reflecting facilitated early perceptual processing of spatial information, could also reflect facilitated late response selection. Such automatic response selection might have contributed to the observed "reverse-V-shaped" response time patterns in **Figure 5**, independently from any hypothetical (partial) adoption of a localization-based response strategy. Future studies might consider isolating possible contributions to TI performance at different perceptual, cognitive, and response-related processing stages (Wickens, 2008).

Surprisingly, differences in correct response time between spatial_id and non-spatial_id modes were also observable at the central loudspeaker location. The second analysis confirmed a response delay in the spatial_id vs. non-spatial_id mode that critically depended on *speech degradation*, being most prominent for clean, less prominent for filtered, and non-significant for noisy speech (see deviations of circle-shaped points from color-shaded bars at the central location in **Figure 5**). A similar pattern emerged for average correct response rate, where TI performance was reduced in the spatial_id vs. non-spatial_id mode under clean speech, but not under degraded speech (see deviations of circle-shaped points from open square-shaped points at the central location in **Figure 7**).

Investigations on top-down, intentional control of auditory selective attention (Shinn-Cunningham, 2008) have accrued evidence for a reduction in listening task performance under conditions where the location of a target talker changed vs. stayed constant across trials (Best et al., 2008, 2010; Koch et al., 2011; Lawo et al., 2014; Oberem et al., 2014). Such behavioral performance decrements, known as *switch-costs*, can be caused by repeated switching of selective attention between non-spatial stimulus features (e.g., after changes in target voice gender) (Best et al., 2008; Koch et al., 2011; Koch and Lawo, 2014; Lawo et al., 2014; Lin and Carlile, 2019) as well as between different locations within the auditory scene (Best et al., 2008; Ihlefeld and Shinn-Cunningham, 2008a; Lin and Carlile, 2015) or between ears [e.g., during dichotic listening (Lawo et al., 2014)]. Lin and

Carlile (2015) found that unpredictable location changes of target speech (embedded in simultaneous masker speech) decreased performance in memory recall and speech comprehension across successive turn-taking trials, which was attributed to costly switches in spatial attention, disrupted auditory streaming, and increased cognitive processing load. In addition, the authors reported corresponding effects for changes in target voice, which provoked switches in non-spatial selective attention (Lin and Carlile, 2019).

The global slowing of behavioral responses (across all loudspeaker locations) evident for clean speech in the spatial_id vs. non-spatial_id mode, during which talkers changed locations over trials, would support the notion of response time switch costs due to frequent switches in spatial auditory attention (Lawo et al., 2014; Oberem et al., 2014). Another type of response time switch-costs caused by frequent changes in talkers' voice characteristics over trials (Best et al., 2008) could not have systematically affected the results, since talker changes occurred randomly during both spatial_id and non-spatial_id modes. Interestingly, those presumably constant and additive switch-costs manifested most distinctly under clean speech, diminished to some degree under filtered speech and seemingly dissolved under noisy speech. As participants listened to degraded speech, they probably needed to concentrate more intensely on the TI task to ensure an optimal level of performance [i.e., heightened "compensatory effort" (Hockey, 1997)], hereby subjecting their internal information processing to proactive cognitive control mechanisms (Braver, 2012). The incurring switch-costs might therefore have been (partially) compensated, for example, by the preparatory and/or sustained mobilization of spare information processing resources (Chiew and Braver, 2013). On the contrary, while listening to clean speech, processing would either be determined by some degree of reactive cognitive control (Braver, 2012) or be (almost) automated toward the later course of the experiment (Wickens, 2008).

Taking a viewpoint opposite to switches in spatial attention, other studies have emphasized "spatial continuity" as a positive influencing factor on speech identification performance (Best et al., 2008, 2010). However, behavioral responses at the central location were not always faster in the non-spatial_id vs. spatial_id mode (see noisy speech in **Figure 5**), which would have been anticipated because of refined attentional selection of solely the central location. Thus, spatial continuity did not seem to play a major role in the present TI task.

Likewise, effects of spatial expectation [also: "spatial certainty" (Best et al., 2010)], being caused by varying probability of auditory stimuli occurring at certain locations, could be ruled out as an alternative explanation for the observed result pattern. With a stimulus probability distribution of 25:50:25% over left:central:right loudspeaker locations, the highest chance of stimuli occurring at the central location should have entailed fastest behavioral responses there compared to lateral (left/right) locations (Singh et al., 2008; Zuanazzi and Noppeney, 2018, 2019), which was never the case.

Repeated strategic switching between different perceptual and cognitive processes (i.e., voice recognition vs. sound localization, see **Table 1**), or between different task sets (TI vs. talker-localization), during a test block can produce another type of response time switch costs (Kiesel et al., 2010). As has already been detailed above, the result patterns shown in **Figure 5** imply that no systematic change in primary response strategy (from voice recognition to sound localization) happened during the spatial_id mode (lateral trials) of the TI task—otherwise average performance at lateral locations in the spatial_id mode should be much closer to average performance in the spatial_loc mode (i.e., the slopes of the "reverse-V-shaped" lines in **Figure 5** should be much steeper). Therefore, the inferred response time switch costs were not ascribable to strategy or task switching, but rather to frequent shifts in spatial attention focus between the three locations.

Neither the TI task nor the talker-localization task revealed any significant differences between the left and right location, suggesting that ear asymmetries [i.e., differences in processing of auditory information occurrent in the left vs. right hemifield, which arise from hemispheric lateralization (Bolia et al., 2001)] were negligible during the TI and talker-localization tasks.

5.2.3. Analysis of Learning Effects (Within/Across Spatial Blocks)

General learning effects speeding up behavioral responses within and across blocks because of increasing familiarity with the test layout, the current stimulus and task sets were controlled by counterbalancing experimental conditions, location-talker and talker-response mappings across blocks (see Section 2.3). Analysis of learning effects within and across *spatial blocks* implied gradually faster behavioral responses. It further showed the "reverse-V-shaped" response time pattern across *loudspeaker location* as clearly present already in the first lateral trial half of the first spatial block (see first plot from the left in **Figure 6**); this would substantiate the assumption stated above, namely that automatic perceptual and/or response selection processes may be (jointly) responsible for its emergence, facilitating behavioral responses irrespective of any to-be-acquired (explicit or implicit) knowledge about talker location. Also, since *loudspeaker location* neither interacted with *lateral trial half* nor *spatial block*, the response time benefit at lateral locations did not increase over the course of the TI task (as would be expected from more and more frequent strategic changes from voice recognition to sound localization), which is in line with the conclusion derived above that participants still relied primarily on voice recognition throughout the TI task.

6. CONCLUSION AND OUTLOOK

This study investigated the effects of spatial presentation of speech stimuli (spoken sentences) on human talker-identification (TI) performance in a "turn-taking"-like listening situation. In a behavioral TI task, participants extracted available spatial auditory cues during spatial speech presentation to achieve faster responses, which were still considerably slower than responses in a talker-localization task (see **Table 1** for experiment specifications, see Section 5.2 for an in-depth behavioral result discussion). Apparently, their primary response

strategy remained to be based on the cognitive process of voice recognition.

Moreover, repeated switching of spatial auditory attention between different locations during spatial speech presentation introduced a global delay in behavioral responses. These "response time switch-costs" diminished when speech was degraded by background noise or bandpass filtering, hereby obscuring individual talkers' voice characteristics and increasing subjectively judged talker-identification effort. Internally, this diminishing may have corresponded to greater cognitive control, which in turn activated processes to compensate the switch-costs. A systematic future investigation of the potential role of controlled allocation of information processing resources would necessitate the use of established physiological methods like pupillometry (Zekveld et al., 2014; Koelewijn et al., 2015) or EEG, that enable continuous assessment of the amount of allocated processing resources [i.e., the perceptual-cognitive load (Wickens, 2008)]. Regarding EEG, specifically the P1-N1-P2 complex, the N2, P3a, and P3b components of the event-related brain potential can be considered suitable candidates for neural indication of spatial attention shifts (Getzmann et al., 2015, 2020; Begau et al., 2021) and processing load influenced by varying speech transmission quality (Uhrig et al., 2020b; Uhrig, 2022).

The general lack of a (stronger) impact of spatialization on subjective listening experience and behavioral performance in the present study calls for further exploration. Apparently, constant switching of individual talkers between the central location and talker-specific lateral locations during the spatial presentation mode (TI task) resulted in a more dynamic, less predictable auditory scene, which might have prevented listeners from extracting or (fully) utilizing available spatial auditory cues. It seems intuitive that another spatial mode that *fixes* talkers at separate locations within the auditory scene could alleviate experienced mental effort of TI and speed up behavioral responses. These benefits might be especially pronounced for multimodal, *audio-visual* speech, as it forms the basis of most human-human communication situations. In a recent study by Begau et al. (2021) examining younger and older adults, the presence of visual lip movement in talking faces congruent with auditory speech stimuli was shown to improve perceptual and cognitive speech processing; such audio-visual facilitation effects (manifesting in behavioral and neurophysiological measures) were observable when the target talker remained fixed at the central location, but were cancelled out when the target talker dynamically changed between the central location and lateral locations.

Overall, self-report ratings of speech quality, speech intelligibility and listening effort [as employed, e.g., to subjectively assess speech transmission quality in telephony settings (ITU-T Recommendation P.800, 1996)] seemed to lack sensitivity for detecting more subtle effects of spatialization on TI performance and its interactions with speech degradation (Uhrig et al., 2020a). Follow-up studies might consider adopting a *multi-method* assessment approach, that combines subjective measures with behavioral and (neuro-)physiological measures, in order to achieve highest possible sensitivity as well as convergent validity across multiple levels of analysis. For instance, continuous measurement of behavioral and physiological responses might prove useful to trace performance changes over longer listening episodes (Borowiak et al., 2014).

Viewing the above conclusions from a practical perspective, audio spatialization implemented in modern speech communication systems may not be solely beneficial [as manifested subjectively, e.g., in higher perceived speech quality, improved speech intelligibility/comprehension and reduced mental effort (Baldis, 2001; Kilgore et al., 2003; Raake et al., 2010; Skowronek and Raake, 2015)]. Configurations involving redundant, inconsistent or uncertain mappings between talkers and locations in physical or virtual space could even expend additional information processing resources due to necessary switches in spatial auditory attention. At what scene complexity (e.g., number of spatially separated talkers) and under which contextual listening conditions (e.g., intensity of present speech degradations) such detrimental effects will be strong enough to reach listeners' awareness and/or noticeably reduce their task performance remain open research questions.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SU together with AP, SM, DMB, and UPS developed the concept and methodological approach of the study. SU implemented the experiments in Aura Lab (Dept. of Electronic Systems, NTNU), collected the data, analyzed the data, and drafted the article, which was further revised based on feedback from the coauthors. All authors have read and finally approved the article.

FUNDING

This work was supported by the strategic partnership program between the Norwegian University of Science and Technology in Trondheim, Norway, and Technische Universität (TU) Berlin in Berlin, Germany. The experiments reported in this paper have also been part of the corresponding author's (SU's) doctoral thesis, which has been published online in DepositOnce, the repository for research data and publications of TU Berlin (Uhrig, 2021), and in monograph form (Uhrig, 2022).

ACKNOWLEDGMENTS

We thank the reviewers for their valuable comments, which have helped to further improve the present article.

REFERENCES

Allen, K., Carlile, S., and Alais, D. (2008). Contributions of talker characteristics and spatial location to auditory streaming. *J. Acoust. Soc. Amer.* 123, 1562–1570. doi: 10.1121/1.2831774

Baer, T., Moore, B. C. J., and Gatehouse, S. (1993). Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times. *J. Rehabil. Res. Dev.* 30, 49–72.

Baldis, J. J. (2001). "Effects of spatial audio on memory, comprehension, and preference during desktop conferences," in *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '01* (Seattle, WA: ACM Press), 166–173. doi: 10.1145/365024.365092

Begau, A., Klatt, L.-I., Wascher, E., Schneider, D., and Getzmann, S. (2021). Do congruent lip movements facilitate speech processing in a dynamic audiovisual multi-talker scenario? An ERP study with older and younger adults. *Behav. Brain Res.* 412, 113436. doi: 10.1016/j.bbr.2021.113436

Best, V., Ahlstrom, J. B., Mason, C. R., Roverud, E., Perrachione, T. K., Kidd, G., et al. (2018). Talker identification: effects of masking, hearing loss, and age. *J. Acoust. Soc. Amer.* 143, 1085–1092. doi: 10.1121/1.5024333

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. Amer.* 127, EL258–EL264. doi: 10.1121/1.3431093

Blauert, J. (1997). *Spatial Hearing: The Psychophysics of Human Sound Localization, Rev. Edn*. Cambridge, MA: MIT Press. doi: 10.1518/001872001775900887

Blum, K., van Rooyen, G.-J., and Engelbrecht, H. (2010). "Spatial audio to assist speaker identification in telephony," in *Proc. IWSSIP 2010 - 17th International Conference on Systems, Signals and Image Processing* (Rio de Janeiro).

Bolia, R. S., Nelson, W. T., and Morley, R. M. (2001). Asymmetric performance in the cocktail party effect: implications for the design of spatial audio displays. *Hum. Fact. J. Hum. Fact. Ergon. Soc.* 43, 208–216.

Borowiak, A., Reiter, U., and Svensson, U. P. (2014). Momentary quality of experience: users' audio quality preferences measured under different presentation conditions. *J. Audio Eng. Soc.* 62, 235–243. doi: 10.17743/jaes.2014.0015

Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends Cognit. Sci.* 16, 106–113. doi: 10.1016/j.tics.2011.12.010

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

Bronkhorst, A. W. (2000). The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. *Acta Acust. Unit. Acust.* 86, 117–128. Available online at: https://www.ingentaconnect.com/content/dav/aaua/2000/00000086/00000001/art00016

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Atten. Percept. Psychophys.* 77, 1465–1487. doi: 10.3758/s13414-015-0882-9

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Amer.* 109, 1101–1109. doi: 10.1121/1.1345696

Brungart, D. S., Ericson, M. A., and Simpson, B. D. (2002). "Design considerations for improving the effectiveness of multitalker speech displays," in *Proceedings of the 2002 International Conference on Auditory Display* (Kyoto), 1–7.

Brungart, D. S., Kordik, A. J., and Simpson, B. D. (2005). Audio and visual cues in a two-talker divided attention speech-monitoring task. *Hum. Fact. J. Hum. Fact. Ergon. Soc.* 47, 562–573. doi: 10.1518/001872005774860023

Brungart, D. S., and Simpson, B. D. (2007). Cocktail party listening in a dynamic multitalker environment. *Percept. Psychophys.* 69, 79–91. doi: 10.3758/BF03194455

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. Amer.* 110, 2527–2538. doi: 10.1121/1.1408946

Chiew, K. S., and Braver, T. S. (2013). Temporal dynamics of motivation-cognitive control interactions revealed by high-resolution pupillometry. *Front. Psychol.* 4:15. doi: 10.3389/fpsyg.2013.00015

Darwin, C. J., and Hukin, R. W. (2000). Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *J. Acoust. Soc. Amer.* 107, 970–977. doi: 10.1121/1.428278

Drullman, R., and Bronkhorst, A. W. (2000). Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *J. Acoust. Soc. Amer.* 107, 2224–2235. doi: 10.1121/1.428503

Ericson, M. A., Brungart, D. S., and Simpson, B. D. (2004). Factors that influence intelligibility in multitalker speech displays. *Int. J. Aviat. Psychol.* 14, 313–334. doi: 10.1207/s15327108ijap14036

Ericson, M. A., and McKinley, R. L. (1997). "The intelligibility of multiple talkers separated spatially in noise," in *Binaural and Spatial Hearing in Real and Virtual Environments*, eds R. H. Gilkey and T. R. Anderson (Mahwah, NJ: Lawrence Erlbaum Associates), 701–724.

Fernández Gallardo, L., Möller, S., and Wagner, M. (2012). "Comparison of human speaker identification of known voices transmitted through narrowband and wideband communication systems," in *Proceedings of 10. ITG Symposium on Speech Communication* (Braunschweig), 1–4.

Fernández Gallardo, L., Möller, S., and Wagner, M. (2013). "Human speaker identification of known voices transmitted through different user interfaces and transmission channels," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (Vancouver, BC: IEEE), 7775–7779. doi: 10.1109/ICASSP.2013.6639177

Fernández Gallardo, L., Möller, S., and Wagner, M. (2015). "Importance of intelligible phonemes for human speaker recognition in different channel bandwidths," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)* (Dresden: ISCA), 1047–1051.

Gaschler, R., Schuck, N. W., Reverberi, C., Frensch, P. A., and Wenke, D. (2019). Incidental covariation learning leading to strategy change. *PLoS ONE* 14, e0210597. doi: 10.1371/journal.pone.0210597

Gatehouse, S., and Gordon, J. (1990). Response times to speech stimuli as measures of benefit from amplification. *Brit. J. Audiol.* 24, 63–68. doi: 10.3109/03005369009077843

Getzmann, S., Hanenberg, C., Lewald, J., Falkenstein, M., and Wascher, E. (2015). Effects of age on electrophysiological correlates of speech processing in a dynamic "cocktail-party" situation. *Front. Neurosci.* 9, 341. doi: 10.3389/fnins.2015.00341

Getzmann, S., Klatt, L.-I., Schneider, D., Begau, A., and Wascher, E. (2020). EEG correlates of spatial shifts of attention in a dynamic multi-talker speech perception scenario in younger and older adults. *Hear. Res.* 398, 108077. doi: 10.1016/j.heares.2020.108077

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: a cognitive-energetical framework. *Biol. Psychol.* 45, 73–93. doi: 10.1016/S0301-0511(96)05223-4

Houben, R., van Doorn-Bierman, M., and Dreschler, W. A. (2013). Using response time to speech as a measure for listening effort. *Int. J. Audiol.* 52, 753–761. doi: 10.3109/14992027

Ihlefeld, A., and Shinn-Cunningham, B. (2008a). Disentangling the effects of spatial cues on selection and formation of auditory objects. *J. Acoust. Soc. Amer.* 124, 2224–2235. doi: 10.1121/1.2973185

Ihlefeld, A., and Shinn-Cunningham, B. (2008b). Spatial release from energetic and informational masking in a selective speech identification task. *J. Acoust. Soc. Amer.* 123, 4369–4379. doi: 10.1121/1.2904826

ITU-T Recommendation P.56 (2011). *Objective Measurement of Active Speech Level*. Geneva: International Telecommunication Union.

ITU-T Recommendation P.800 (1996). *Methods for Subjective Determination of Transmission Quality*. Geneva: International Telecommunication Union.

ITU-T Recommendation P.851 (2003). *Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems*. Geneva: International Telecommunication Union.

Kaplan-Neeman, R., Kishon-Rabin, L., Henkin, Y., and Muchnik, C. (2006). Identification of syllables in noise: electrophysiological and behavioral correlates. *J. Acoust. Soc. Amer.* 120, 926–933. doi: 10.1121/1.2217567

Kidd, G., Arbogast, T. L., Mason, C. R., and Gallun, F. J. (2005a). The advantage of knowing where to listen. *J. Acoust. Soc. Amer.* 118, 3804–3815. doi: 10.1121/1.2109187

Kidd, G., Mason, C., Brughera, A., and Hartmann, W. (2005b). The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Acta Acust. Unit. Acust.* 91, 526–536. Available online at: https://www.ingentaconnect.com/content/dav/aaua/2005/00000091/00000003/art00014

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., et al. (2010). Control and interference in task switching a review. *Psychol. Bull.* 136, 849–874. doi: 10.1037/a0019842

Kilgore, R. M., Chignell, M., and Smith, P. (2003). "Spatialized audioconferencing: what are the benefits?," in *Proceedings of the 2003 Conference of the Centre for Advanced Studies on Collaborative Research* (Toronto, ON), 135–144. doi: 10.1145/961322.961345

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. Amer.* 127, 2498–2508. doi: 10.1121/1.3327507

Koch, I., and Lawo, V. (2014). Exploring temporal dissipation of attention settings in auditory task switching. *Atten. Percept. Psychophys.* 76, 73–80. doi: 10.3758/s13414-013-0571-5

Koch, I., Lawo, V., Fels, J., and Vorländer, M. (2011). Switching in the cocktail party: exploring intentional control of auditory selective attention. *J. Exp. Psychol. Hum. Percept. Perform.* 37, 1140–1147. doi: 10.1037/a0022189

Koelewijn, T., de Kluiver, H., Shinn-Cunningham, B. G., Zekveld, A. A., and Kramer, S. E. (2015). The pupil response reveals increased listening effort when it is difficult to focus attention. *Hear. Res.* 323, 81–90. doi: 10.1016/j.heares.2015.02.004

Köster, F., Schiffner, F., Guse, D., Ahrens, J., Skowronek, J., and Möller, S. (2015). "Towards a MATLAB toolbox for imposing speech signal impairments following the P.TCA schema," in *Audio Engineering Society Convention 139* (New York).

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest Package: tests in linear mixed effects models. *J. Statist. Softw.* 82, i13. doi: 10.18637/jss.v082.i13

Latinus, M., and Belin, P. (2011). Human voice perception. *Curr. Biol.* 21, R143–R145. doi: 10.1016/j.cub.2010.12.033

Lawo, V., Fels, J., Oberem, J., and Koch, I. (2014). Intentional attention switching in dichotic listening: exploring the efficiency of nonspatial and spatial selection. *Quart. J. Exp. Psychol.* 67, 2010–2024. doi: 10.1080/17470218.2014.898079

Leman, A., Faure, J., and Parizet, E. (2008). Influence of informational content of background noise on speech quality evaluation for VoIP application. *J. Acoust. Soc. Amer.* 123, 3066–3066. doi: 10.1121/1.2932822

Lin, G., and Carlile, S. (2015). Costs of switching auditory spatial attention in following conversational turn-taking. *Front. Neurosci.* 9, 124. doi: 10.3389/fnins.2015.00124

Lin, G., and Carlile, S. (2019). The effects of switching non-spatial attention during conversational turn taking. *Sci. Rep.* 9, 8057. doi: 10.1038/s41598-019-44560-1

Lu, C.-h., and Proctor, R. W. (1995). The influence of irrelevant location information on performance: a review of the Simon and spatial Stroop effects. *Psychon. Bull. Rev.* 2, 174–207. doi: 10.3758/BF03210959

Mackersie, C., Neuman, A. C., and Levitt, H. (1999). A comparison of response time and word recognition measures using a word-monitoring and closed-set identification task. *Ear Hear.* 20, 140–148. doi: 10.1097/00003446-199904000-00005

McAnally, K. I., and Martin, R. L. (2007). Spatial audio displays improve the detection of target messages in a continuous monitoring task. *Hum. Fact. J. Hum. Fact. Ergon. Soc.* 49, 688–695. doi: 10.1518/001872007X215764

Nelson, W. T., Bolia, R. S., Ericson, M. A., and McKinley, R. L. (1999). Spatial audio displays for speech communications: a comparison of free field and virtual acoustic environments. *Proc. Hum. Fact. Ergon. Soc. Annu. Meet.* 43(22):1202–1205. doi: 10.1177/154193129904302207

Oberem, J., Lawo, V., Koch, I., and Fels, J. (2014). Intentional switching in auditory selective attention: exploring different binaural reproduction methods in an anechoic chamber. *Acta Acust. Unit. Acust.* 100, 1139–1148. doi: 10.3813/AAA.918793

Pals, C., Sarampalis, A., van Rijn, H., and Başkent, D. (2015). Validation of a simple response-time measure of listening effort. *J. Acoust. Soc. Amer.* 138, EL187–EL192. doi: 10.1121/1.4929614

Raake, A. (2002). Does the content of speech influence its perceived sound quality? *Sign* 1, 1170–1176.

Raake, A., Schlegel, C., Hoeldtke, K., Geier, M., and Ahrens, J. (2010). "Listening and conversational quality of spatial audio conferencing," in *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space* (Tokyo).

Sarampalis, A., Kalluri, S., Edwards, B., and Hafter, E. (2009). Objective measures of listening effort: effects of background noise and noise reduction. *J. Speech Lang. Hear. Res.* 52, 1230–1240. doi: 10.1044/1092-4388(2009/08-0111)

Schuck, N., Gaschler, R., Wenke, D., Heinzle, J., Frensch, P., Haynes, J.-D., et al. (2015). Medial prefrontal cortex predicts internally driven strategy shifts. *Neuron* 86, 331–340. doi: 10.1016/j.neuron.2015.03.015

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends Cognit. Sci.* 12, 182–186. doi: 10.1016/j.tics.2008.02.003

Simon, J. R. (1969). Reactions toward the source of stimulation. *J. Exp. Psychol.* 81, 174–176. doi: 10.1037/h0027448

Singh, G., Pichora-Fuller, M. K., and Schneider, B. A. (2008). The effect of age on auditory spatial attention in conditions of real and simulated spatial separation. *J. Acoust. Soc. Amer.* 124, 1294–1305. doi: 10.1121/1.2949399

Skowronek, J., and Raake, A. (2015). Assessment of cognitive load, speech communication quality and quality of experience for spatial and non-spatial audio conferencing calls. *Speech Commun.* 66, 154–175. doi: 10.1016/j.specom.2014.10.003

Uhrig, S. (2021). *Human Processing of Transmitted Speech Varying in Perceived Quality*. Doctoral Thesis, Technische Universität Berlin, Berlin. doi: 10.14279/depositonce-11118

Uhrig, S. (2022). *Human Information Processing in Speech Quality Assessment*. T-Labs Series in Telecommunication Services. Cham: Springer International Publishing. doi: 10.1007/978-3-030-71389-8

Uhrig, S., Mittag, G., Möller, S., and Voigt-Antons, J.-N. (2019a). Neural correlates of speech quality dimensions analyzed using electroencephalography (EEG). *J. Neural Eng.* 16, 036009. doi: 10.1088/1741-2552/aaf122

Uhrig, S., Mittag, G., Möller, S., and Voigt-Antons, J.-N. (2019b). P300 indicates context-dependent change in speech quality beyond phonological change. *J. Neural Eng.* 16, 066008. doi: 10.1088/1741-2552/ab1673

Uhrig, S., Möller, S., Behne, D. M., Svensson, U. P., and Perkis, A. (2020a). "Testing a quality of experience (QoE) model of loudspeaker-based spatial speech reproduction," In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)* (Athlone: IEEE), 1–6. doi: 10.1109/QoMEX48832.2020.9123119

Uhrig, S., Perkis, A., and Behne, D. M. (2020b). Effects of speech transmission quality on sensory processing indicated by the cortical auditory evoked potential. *J. Neural Eng.* 17, 046021. doi: 10.1088/1741-2552/ab93e1

Wältermann, M., Raake, A., and Möller, S. (2010). Quality dimensions of narrowband and wideband speech transmission. *Acta Acust. Unit. Acust.* 96, 1090–1103. doi: 10.3813/AAA.918370

Wickens, C. D. (2008). Multiple resources and mental workload. *Hum. Fact. J.Hum. Fact. Ergon. Soc.* 50, 449–455. doi: 10.1518/001872008X288394

Yost, W. A., Dye, R. H., and Sheft, S. (1996). A simulated "cocktail party" with up to three sound sources. *Percept. Psychophys.* 58, 1026–1036. doi: 10.3758/BF03206830

Zekveld, A. A., Rudner, M., Kramer, S. E., Lyzenga, J., and Rönnberg, J. (2014). Cognitive processing load during listening is reduced more by decreasing voice similarity than by increasing spatial separation between target and masker speech. *Front. Neurosci.* 8, 88. doi: 10.3389/fnins.2014.00088

Zuanazzi, A., and Noppeney, U. (2018). Additive and interactive effects of spatial attention and expectation on perceptual decisions. *Sci. Rep.* 8, 6732. doi: 10.1038/s41598-018-24703-6

Zuanazzi, A., and Noppeney, U. (2019). Distinct neural mechanisms of spatial attention and expectation guide perceptual inference in a multisensory

world. *J. Neurosci.* 39, 2301–2312. doi: 10.1523/JNEUROSCI.2873-18.2019

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

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

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 03 March 2022
doi: 10.3389/fnins.2022.763324

ERP Evidences of Rapid Semantic Learning in Foreign Language Word Comprehension

*Akshara Soman, Prathibha Ramachandran and Sriram Ganapathy**

Learning and Extraction of Acoustic Patterns Lab, Indian Institute of Science, Bangalore, India

The event-related potential (ERP) of electroencephalography (EEG) signals has been well studied in the case of native language speech comprehension using semantically matched and mis-matched end-words. The presence of semantic incongruity in the audio stimulus elicits a N400 component in the ERP waveform. However, it is unclear whether the semantic dissimilarity effects in ERP also appear for foreign language words that were learned in a rapid language learning task. In this study, we introduced the semantics of Japanese words to subjects who had no prior exposure to Japanese language. Following this language learning task, we performed ERP analysis using English sentences of semantically matched and mis-matched nature where the end-words were replaced with their Japanese counterparts. The ERP analysis revealed that, even with a short learning cycle, the semantically matched and mis-matched end-words elicited different EEG patterns (similar to the native language case). However, the patterns seen for the newly learnt word stimuli showed the presence of P600 component (delayed and opposite in polarity to those seen in the known language). A topographical analysis revealed that P600 responses were pre-dominantly observed in the parietal region and in the left hemisphere. The absence of N400 component in this rapid learning task can be considered as evidence for its association with long-term memory processing. Further, the ERP waveform for the Japanese end-words, prior to semantic learning, showed a P3a component owing to the subject's reaction to a novel stimulus. These differences were more pronounced in the centro-parietal scalp electrodes.

Keywords: ERPs, language learning, semantics, word learning, speech perception

OPEN ACCESS

Edited by:
Howard Charles Nusbaum, University of Chicago, United States

Reviewed by:
Hema Murthy, Indian Institute of Technology Madras, India
Hynek Hermansky, Johns Hopkins University, United States

****Correspondence:***
Sriram Ganapathy
sriramg@iisc.ac.in

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

Received: *23 August 2021*
Accepted: *31 January 2022*
Published: *03 March 2022*

Citation:
Soman A, Ramachandran P and Ganapathy S (2022) ERP Evidences of Rapid Semantic Learning in Foreign Language Word Comprehension. Front. Neurosci. 16:763324. doi: 10.3389/fnins.2022.763324

1. INTRODUCTION

Humans have the unique ability to learn and process languages throughout ones life time. The main hypothesis for learning a new language states that the process begins by imitation of the sounds spoken by native speakers of the language (Speidel and Nelson, 2012). The association of semantics with speech sounds constitutes the next phase of learning (Clark, 1973), which further develops to sentence formation and syntax/grammar learning. These processes may not be sequential and may be interleaved with each other.

The brain activity related to the perception and cognition of language can be studied through event-related potentials (ERPs). The ERPs are computed by averaging the electroencephalogram (EEG) recordings evoked by the same event. The ERPs triggered by verbal stimuli have been

associated with different aspects of language learning (Kutas and Hillyard, 1984).

In this article, we present a study to analyze rapid language learning effects using ERP analysis where the end-words of English sentences were replaced with Japanese words. In the first phase of the experiment, the subjects who were proficient in English with no prior exposure to Japanese words were presented with English sentences containing Japanese end-words. A subsequent language learning phase introduces the semantics of the Japanese stimuli through its pictorial description. In the final phase, the subjects listen to English sentences again with Japanese end-words armed with the knowledge of the semantics. The Japanese end-words in the English sentences may be congruent or in-congruent with the sentence context. An ERP analysis on Japanese end-words before and after semantic learning illustrate significant changes that highlight the neural processes involved in learning. Further, the difference ERP of Japanese stimuli in congruent and incongruent condition (after language learning) elicits delayed and positive ERP components as opposed to the N400 effects observed in native language under semantic mis-match condition.

Contributions:

- The study contrasts the ERP effects of semantic incongruity in a proficient language vs. a newly acquired language.
- This study demonstrates that rapid semantic learning elicited ERP responses in the form of a delayed positive response at 500–700 ms from the onset of the end-word for the incongruent words.
- The scalp electrodes showing semantic effects from a newly learned word of a foreign language are located in the pareital and occipital regions.
- The amplitude of semantic incongruity (ERP component) in foreign language (Japanese) is more for pure foreign words (hiragana words in Japanese) vs. English loan words (katakana words in Japanese).

RELEVANT LITERATURE

The N400 component was first introduced by Kutas and Hillyard (1980), where the reading task composed of presenting the participant with a set of sentences that end with a congruent or incongruent word. These semantically incongruent end-words in a sentence elicited specific type of ERPs (Nobre and Mccarthy, 1995; Hoeks et al., 2004; Dikker and Pylkkanen, 2011), known as the N400, a negative-going deflection between 250 and 400 ms after the end-word onset. The presence of N400 in semantically incongruent stimuli was also observed in other stimuli presentations (Kutas et al., 1987) like reading (Szűcs et al., 2007) and visual forms (Nigam et al., 1992). The N400 was characterized as a reaction to an unexpected or inappropriate, but syntactically correct word at the end of a sentence. The N400 component was not observed for stimuli with syntactical and grammatical errors (Kutas and Hillyard, 1983). This result is seen as the evidence that the N400 wave is more closely related to the semantics than to the syntactical processing.

The N400 is not only a response to semantic improbability or anomaly, but also as an indicator of the access to semantic information associated with the stimuli (Kutas and Federmeier, 2011). When a word is congruent in its context, there is little new information to process and hence, this evokes lower N400 response than an incongruent word. The amplitude of the N400 is sensitive to a word's semantic expectancy (Brown and Hagoort, 1993) and found to be larger in response to more unexpected stimuli (Curran et al., 1993; Holcomb, 1993; DeLong et al., 2005). The N400 has also been used to show that language comprehension is incremental (Van Petten and Kutas, 1990) and involves prediction (Federmeier and Kutas, 1999b). Further, semantic information processing happens even without active awareness (Luck et al., 1996). It has also been pointed out that language mechanisms vary across the hemispheres (Federmeier and Kutas, 1999a) and can change over the course of normal aging (Wlotko et al., 2010).

Even though the N400 has contributed significantly to the understanding of language comprehension, the N400 response is not confined to language domain alone, and hence it is not a "language-component" (Leckey and Federmeier, 2020). The N400 is not only seen in word comprehension, but also for different kinds of pictorial stimuli (e.g., comics/cartoons, drawings, pictures of objects, natural scenes), faces, gestures, and environmental sounds. Thus, it can be elicited for any kind of stimulus linked to long-term memory representations (Kutas and Federmeier, 2011).

The P3a, or novelty P3 (Comerchero and Polich, 1999), is a positive-going component with peak latency in the range of 250–280 ms. It is topographic distribution that shows maximum amplitude over frontal and central electrode sites. P3a has been associated with cognitive tasks of involuntary attention and the processing of novelty (Polich, 2007).

1.1. N400 as an Index of Word Learning

The N400 has been established in numerous studies to be a useful index of new word learning. In a study by Perfetti et al. (2005), adults were taught the meanings of infrequent and unfamiliar words. The N400 component was seen for unrelated word pairings containing the trained words and not for those involving the unfamiliar words.

Mestres-Missé et al. (2007) investigated context-based learning of novel words using ERPs. The researchers specifically introduced novel word forms in the ending position of meaningful sentences that the participants read during the training phase. In a following relatedness judgment task on word pairs, consisting of a trained novel word (prime) and a real word (target), the study found a reduction in the N400 for targets words that were associated with the prime word compared to the unrelated target-prime pairs.

Batterink and Neville (2011) investigated contextual learning by embedding pseudo-words into meaningful short story contexts. In a subsequent relatedness judgment task, the study showed a reduction in the N400 amplitude for targets corresponding to the novel word.

The N400 has also been demonstrated to be sensitive to new word learning after just one exposure to a novel word

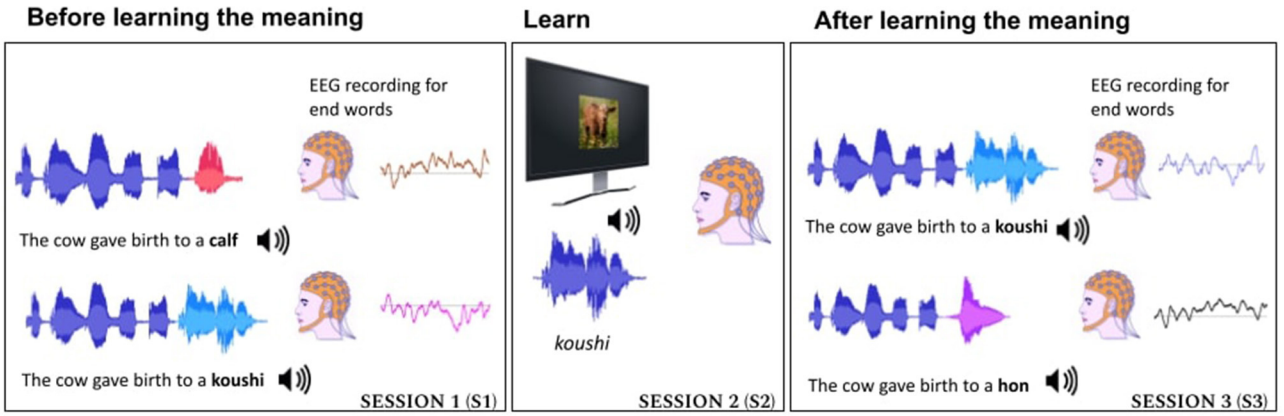

FIGURE 1 | Experiment pipeline consisted of three sessions. S1—where the subject listened to acoustics of Japanese words used in English context and English end-words in congruent and incongruent context. S2—where the semantics of Japanese words were introduced using images. S3—where the subject listened to Japanese words in English context (both congruent sentences and incongruent sentences).

in the context of a highly predictive sentence (Borovsky et al., 2010, 2012). The findings of such investigations provide neurophysiological evidence for understanding the semantic learning of the new words.

1.2. Intra-Sentential Code-Switching

The N400, left-lateralized anterior negativity (LAN), and the late positive component (LPC; also referred to as P600) are the three primary ERP components identified in research on intra-sentential code switching.

The LAN is a left-lateralized anterior negativity that occurs in the same time frame as the N400 (300–500 ms) but with a distinct topographic scalp distribution. Friederici (2002) observed LAN effects in morphosyntactic processing, as well as in the processing of code-switched sentences. The higher working memory load resulting from integrating morphological signals of the code-switched word into the wider sentence context was interpreted as the switch-related LAN component (Moreno et al., 2002).

The LPC (or P600) is a positive-going wave that arises 500–600 ms after the stimulus and lasts several hundred milliseconds (Osterhout and Holcomb, 1992; Hagoort et al., 1993). It has a wide posterior scalp distribution and is strongest in the centro-parietal areas. Friederici (1995), Kaan et al. (2000), and Tanner et al. (2017) infer that the LPC indexes sentence-level rearrangement or re-analysis. The LPC, according to this view, indicates sentence-level wrap-up or meaning revision process, which in the instance of intra-sentential code-switching, reflects the sentence-level reorganization of two languages into a cohesive utterance. The LPC has also been linked to the processing of unexpected or unlikely task-relevant events (McCallum et al., 1984; Coulson et al., 1998), as well as the reorganization of stimulus-response mapping (Moreno et al., 2008). A switch-related LPC represents bilinguals' perception of a language transition as an unexpected occurrence involving a shift in form rather than meaning (Moreno et al., 2002).

2. MATERIALS AND METHODS

The experimental paradigm used in this study is illustrated in **Figure 1**.

2.1. Stimuli

All the speech stimuli used in the study were recorded from speech provided by a single female speaker who was proficient in both English and Japanese languages. The speaker's first language was Tamil, a south-Indian language. The accent of the speaker was Indian English. Given that all the listeners were also from Indian origin, we found that the listeners had no issues in understanding the English content. Further, the work had employed a single speaker for all the stimuli. This helped to remove the effect of speaker variabilities or speaker switching in the stimuli used.

The speech stimuli used in the experiment consisted of isolated sentences and isolated words. They were recorded in a sound proof booth. The entire stimuli set used in the experiment is listed in **Supplementary Table 2**. The audio and image files used for the experiment are available in this project's GitHub repository[1]. The stimuli set consisted of 90 unique English sentences. The sentences were selected such that the end-word was highly predictable from the sentence context. Our stimuli set was taken from the high cloze probable sentences (cloze probability in the range of 67–100%) of the Block-Baldwin sentence set (Block and Baldwin, 2010). The audio duration of the sentence varied between 1.4 and 2.4 s with an average of 2.2 s.

The end-words of the sentences were either from English or Japanese language and they were designed to be either semantically congruent or incongruent to the preceding context in the sentence. All the stimuli conditions are listed in **Table 1**. The first condition (C1) consists of 90 English sentences from the original Block-Baldwin set without any modification. The stimuli

[1]https://github.com/iiscleap/Semantics-EEG-ERPStudy

TABLE 1 | Different conditions of the stimuli sentences used in our experiment with an example.

Stimuli condition	Example
C1—Eng. Congruent	*The cow gave birth to a **calf**.*
C2—Eng. Incongruent	*The cow gave birth to a **book**.*
C3—Jap. Congruent	*The cow gave birth to a **koushi**.*
C4—Jap. Incongruent	*The cow gave birth to a **hon**.*

Endword of the stimulus sentence is shown in bold. Different endwords give rise to different stimuli conditions.

for the other three conditions of the experiment were created by replacing the end-word with all the preceding words intact. In condition 2 (C2), the end-word was replaced by an English word which is unexpected in the sentence context (English incongruent condition); in condition 3 (C3), the end-word was replaced by a Japanese word of the congruent semantics (Japanese—congruent condition); and in condition 4 (C4), the end-word was replaced by a Japanese word of unexpected meaning (Japanese—incongruent condition). Thus, the experiment had a total of 360 sentences. The sentences were carefully chosen such that the end-word can be visualized as an image. Each of the stimuli condition used the same base set of English sentences. The conditions differed only in terms of the end-word of the sentences. Each stimulus was recorded as a full sentence for each condition separately.

We choose the Japanese language as the novel language as it does not belong to the same language family as English. At the same time, Japanese language contains a set of loan-words from English termed as katakana words. The katakana words are typically English words that have been adapted without translation into the Japanese language (Olah, 2007). The katakana words sound similar to their English counterparts (cognate words). By using a mix of non-cognate native Japanese words (referred to as hiragana words) and katakana words, we were able to study the effect of phonetic similarity in learning. The set of Japanese words used in our experiments (**Supplementary Table 2**) consisted of 38 katakana words and 52 hiragana Japanese words. This unbalanced distribution of katakana and hiragana words are in compliance with the word frequency distribution in the Japanese language (Chikamatsu et al., 2000). Thus, for stimuli conditions C3 and C4, the end-word can be either a katakana word or a hiragana word.

2.2. Experiment Flow

A schematic illustration of the experiment is shown in **Figure 1**. A short video depicting the experiment flow is available in this project's GitHub repository[2].The end-words were either congruent or incongruent to the context of the sentence. The participants learned the semantics of the unknown Japanese words using image supervision (shown in session 2 of **Figure 1**) and the learning was analyzed when the words were used in sentences both in congruent and incongruent condition (shown in session S3 of **Figure 1**). In each session, the sentences of different conditions were presented in random order using a loud speaker.

In session S1, the subject listened to English sentences with congruent or incongruent end-word (C1 and C2 from **Table 1**). This session also contained English sentences with Japanese end-word. The subject got the first exposure to these Japanese words in this session without the semantic information.

In the next session (S2), the participants were provided with the semantics of the Japanese words. A block of five new Japanese words was considered and word meanings were conveyed using the respective image. The image form of the word and its audio are presented simultaneously. The five words of each block were presented in random order. A retrieval task was also designed to ensure the learning ability of the subject. In the retrieval task, the subject was asked to speak the Japanese word for the image shown in the computer screen. After the subject provided the spoken response, the audio of the correct Japanese word was replayed. Here, the subject was affirming the learning or corrected their learning if the word recollection was inaccurate. We do not analyze the EEG data from S2 in this article.

In session 3 (S3), the subject listened to the sentence audio played through the loudspeaker. It was an English sentence with a Japanese end-word. Here, the end-word was either congruent or incongruent to the context of the sentence (C3 and C4 in **Table 1**). The end-word was one of the 5 Japanese words learned in the preceding session (S2). Hence, there were a total of 10 sentences (equal number of congruent and incongruent) played in session 3 for the current block. These 10 sentences were presented in random order. After the audio signal was played, a recognition task was carried out to ensure that the subject recollected the meaning of the Japanese end-word. In the recognition task, the subject was asked to pick the image corresponding to the Japanese end-word. The subjects recorded their choice of image by speaking the corresponding number index on the screen. The subject responses were later evaluated manually to assess their recognition accuracy. The behavioral results are discussed in Section 3.1. Only the EEG responses to the Japanese words, whose meaning was recollected correctly, were used in the subsequent ERP analysis.

In order to avoid the memory load of learning and recalling 90 new Japanese words, S2 and S3 were performed in 18 blocks of five words each. The session S3 for a set of five words was conducted immediately after the subject was trained on the semantics in session S2. We designed the experiment in this way to reduce the memory load on subject as we were more interested to analyze the semantic effects of the newly learned words in both congruent and incongruent condition. A particular Japanese word was presented five times to the subject: once in S1 (sentence end-word without semantic knowledge), twice in S2 (as isolated words in the learning phase) and twice in S3. In S3, it was used as the sentence end-word in congruent and incongruent context. A particular sentence and end-word pair was introduced only once in the whole experiment. For incongruent condition, sentence-end-word pairing was carried by random shuffling and incongruence was ensured by manual selection. The order of congruent and incongruent conditions in S3 was randomized.

[2]https://github.com/iiscleap/Semantics-EEG-ERPStudy

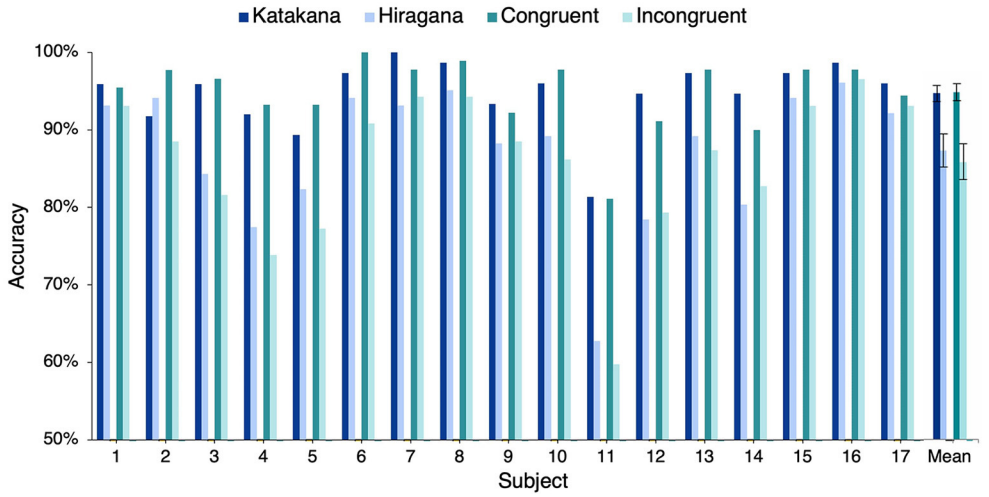

FIGURE 2 | Behavioral task performance: this plot shows the percentage of Japanese words correctly associated with the image description by the subjects.

Thus, the exposure to new Japanese words were balanced across conditions. The three sessions were recorded in an interleaved fashion in one recording setup for each of the subjects. The experiment design ensured that the subject does not get exposed to katakana words more than hiragana words.

2.3. Participants

The participants had self-reported normal hearing and no history of neurological disorders. Twenty-one subjects took part in the experiment. Two subjects were eliminated due to poor EEG data quality and another two were eliminated due to equipment failure. Seventeen adults participated in this study (mean age = 25.7, age span = 22–35, 7 female and 10 male) and they had an intermediate or higher level of English proficiency. This was verified with the Oxford Listening Level Test[3] before the commencement of the experiment. The native language of the subjects was one of the five Indian languages (Malayalam, Tamil, Kannada, Telugu, or Hindi). All subjects provided written informed consent to take part in the experiment and received a monetary compensation. The Indian Institute of Science Human Ethics Board approved all procedures of the experiment. The methods were carried out in accordance with the relevant guidelines and regulations.

We have performed the power analysis with an assumed effect size of d = 0.5 (Zwaan et al., 2018; Brysbaert, 2019) to check the sufficiency of the number of subjects and trials used in the experiments. We performed the power analysis on our experiment using the GPower software (Faul et al., 2007). This analysis revealed that our study is a properly powered experiment (with power value more than 95%). Hence, the effects reported in the study are very likely to be robust and reproducible.

2.4. EEG Recording Setup

The EEG signals were recorded employing a BESS F-64 amplifier with 64 passive gel-based Ag/AgCl electrodes placed according to the extended 10–20 positioning NeuroScan 4.5 system (Soman et al., 2019). It was recorded at a sampling rate of 1,024 Hz. An isolated frontal electrode was utilized as ground and the average of two earlobe electrodes was utilized as reference. The channel impedance was kept underneath 10 kω throughout the recording. The EEG recording took place in a sound-proof, electrically isolated room.

2.5. Data Preprocessing

Prior to ERP offline-averaging, the line noise was removed and the continuous EEG data were band-pass filtered between 1 and 8 Hz. The noisy trials were automatically removed using EEGLAB (Delorme and Makeig, 2004).

3. RESULTS

3.1. Behavioral Task

A behavioral task was conducted to ensure that the subject successfully recalled the meaning of the Japanese words while listening to the end-word of the sentence stimuli in session S3. From a set of five images, the subject was asked to identify the image corresponding to the Japanese end-word of the sentence. **Figure 2** shows the percentage of words' whose meaning was correctly identified by the subjects. The solid line shows the overall accuracy obtained by each subject. Eleven of 17 subjects correctly recalled more than 90% of the word semantics (chance accuracy was 20%). Thus, the subsequent ERP-based conclusions in S3 had a strong behavioral basis. As seen in **Figure 2**, the number of correct responses in the Japanese congruent condition was greater than number of correct responses in the incongruent condition for all the subjects. A right-tailed paired sample *t*-test showed that recognition accuracy of congruent end-words was significantly higher than incongruent end-words [congruent: 94.88, incongruent: 85.90, $t_{(16)} = 5.92$, $p = 1.06e-5$]. This indicated that it was easy to recollect the meaning of a word when it was used in the correct semantic context in a sentence. **Figure 2** shows that the recognition accuracy of katakana words

[3]https://www.oxfordonlineenglish.com/english-level-test/listening

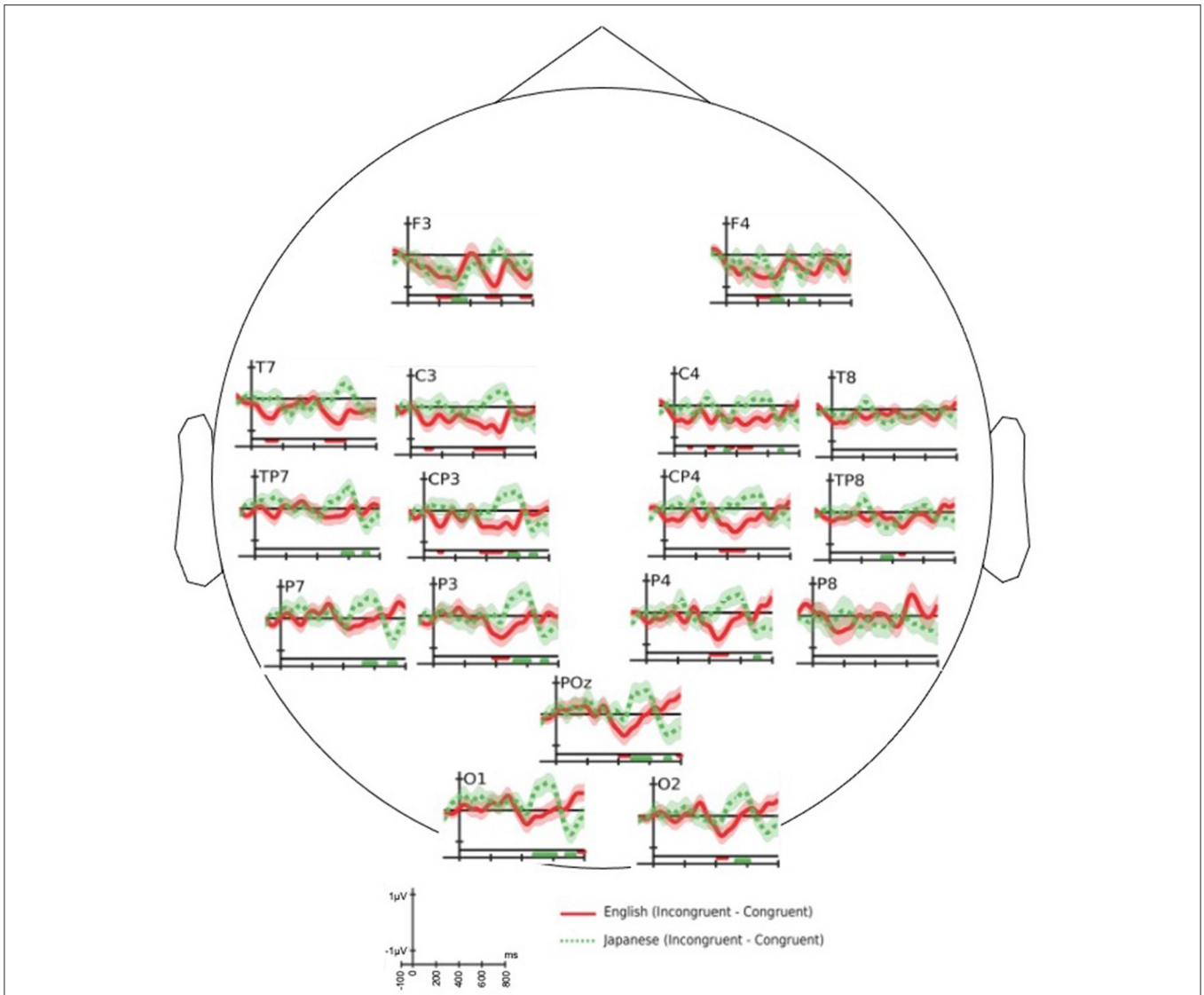

FIGURE 3 | The grand average of difference event-related potential (ERP) response of congruent end-word from in-congruent end-word: end-words in English (solid—red color) and Japanese: session 3 (dotted—green color). The horizontal bars drawn in the bottom of each plot identifies the time regions with significant (two-sample t-test with $p < 0.05$) difference in the ERP response. The bars have the same color of the associated difference waveform.

are better than that of hiragana words for all subjects except one (subject 2). A right-tailed paired sample t-test showed that recognition accuracy of katakana words was significantly higher than hiragana words [katakana: 94.72, hiragana: 87.31, $t_{(16)} = 5.32$, $p = 3.46e - 5$]. Hence, we conclude that the lexical association of katakana words with English words made it easier to recall katakana words. In the subsequent analysis, only words that were correctly recalled are used.

3.2. ERP Analysis

The ERPs are time-locked EEG responses averaged across multiple trials for the same stimulus condition. The ERPs are computed for epochs extending from 100 ms before the end-word onset to 800 ms after the end-word onset. The time $t = 0$ in the ERP plots corresponds to the onset of the end-word in the stimuli sentences. The difference ERP waves were calculated by subtracting an ERP wave of one condition from the other. All the grand average ERP plots shown in this paper are ERP responses averaged across 17 subjects. The two sample t-test was conducted at each time sample to validate the significance of the ERP responses. All difference ERP plots are marked with time regions of significance where the difference value is significantly above zero ($p < 0.05$). This is indicated by horizontal bars in the bottom of the plot.

3.2.1. Effect of Incongruity

The ERP response shown by solid line in **Figure 3** exhibits N400 effect [$t_{(16)} = -5.59$, $p = 2.05e - 05$] in 300–500 ms over centro-parietal and parietal electrodes) for the difference of English congruent response (*C1S1*) from English incongruent

response (*C2S1*). This result is aligned with the prior research on N400 for auditory tasks (Hagoort and Brown, 2000), which is elicited for semantically incongruent stimuli conditions. To the best of our knowledge, the ERP analysis for other conditions that follow are reported for the first time in literature.

The difference of grand average ERP response of Japanese congruent end-word from Japanese incongruent end-word is shown in **Figure 3** as dotted line. The semantic incongruity of newly learned Japanese end-words did not evoke an N400 response relative to the congruent Japanese end-words [$t_{(16)} = 0.32$, $p = 0.75$ over the time-window 300–500 ms] over the cluster of centro-parietal and parietal electrode locations.

As observed in **Figure 3**, the English and Japanese end-words had different responses around 400 and 600 ms. Both the English and Japanese difference ERP did not have significant peak in the early part of the time axis. The difference ERP response for Japanese end-words elicited a P600 like component [$t_{(16)} = 5.11$, $p = 5.24e-05$ for time-window: 500–700 ms] over a cluster of centro-parietal and parietal electrode locations. This figure also highlights that ERP effects of semantic incongruity are possible without a long-term learning process. A similar difference in ERP response for English end-words in the 500–700 ms time-window [$t_{(16)} = -1.05$, $p = 0.310$] was not observed. In other words, the semantic incongruency in the stimuli did not evoke P600 response for familiar language.

To confirm these differences statistically, we performed ANOVA on the mean ERP amplitudes across the scalp in two time-windows (300–500 and 500–700 ms) of interest. We considered congruity and scalp region (frontal/central/parietal) as the independent factors. In the N400 time-window (300–500

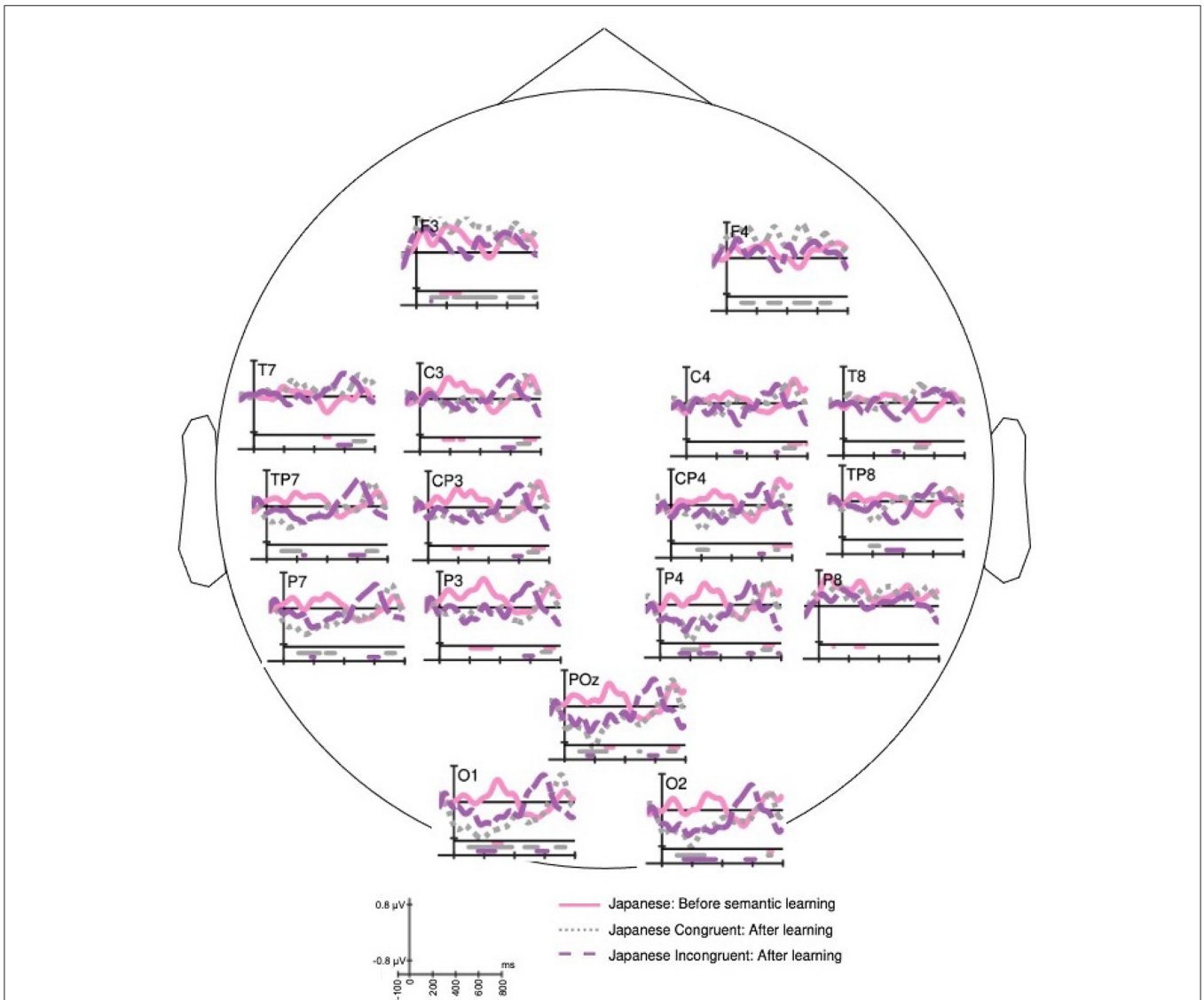

FIGURE 4 | The grand average of event-related potential (ERP) responses to Japanese end-word before learning its meaning (solid), Japanese congruent end-word (dotted), and Japanese in-congruent end-word (dashed) after learning the meaning. The horizontal bars drawn in the bottom of each plot signify the time regions with significant (*t*-test with $p < 0.05$) ERP amplitude from the value of 0. The bars have the same color of the associated waveform.

ms), we observed robust effects of congruity for English end-words. We observed similar effects for Japanese end-words in the P600 (500–700 ms) window. The ANOVA reveals significant interaction between language and congruity in both the time-windows [for 300–500 ms window: $F_{(1,68)} = 16.13$, $p = 6e - 5$ and for 500–700 ms window: $F_{(1,68)} = 37.85$, $p = 9e - 10$].

3.2.2. Effect of Word Learning

Figure 4 shows the ERP response for the same set of Japanese language words before and after learning the semantics. The Japanese words exposed without semantic knowledge (solid line) evoke early positive peaks around 100 and 300 ms. This P300 response is possibly an indicator of exposure to novel information. The P100 and P300 responses disappear in the exposures after the subject learned the meaning of the word. The three responses in **Figure 4** differ in the peak amplitude of N200 component. The ERP of Japanese end-word before semantic learning elicited a negative dip between P100 and P300 peaks. The ERP of Japanese congruent end-word after semantic learning elicited significant negative peak around 200 ms over parieto-occipital and occipital electrode sites. The incongruent end-word post semantic learning also elicited a negative peak, but with a lower peak amplitude than congruent case. The Japanese end-word response before semantic learning (solid line) and Japanese congruent response post semantic learning (dotted line) have a negative deflection before 600 ms. Both

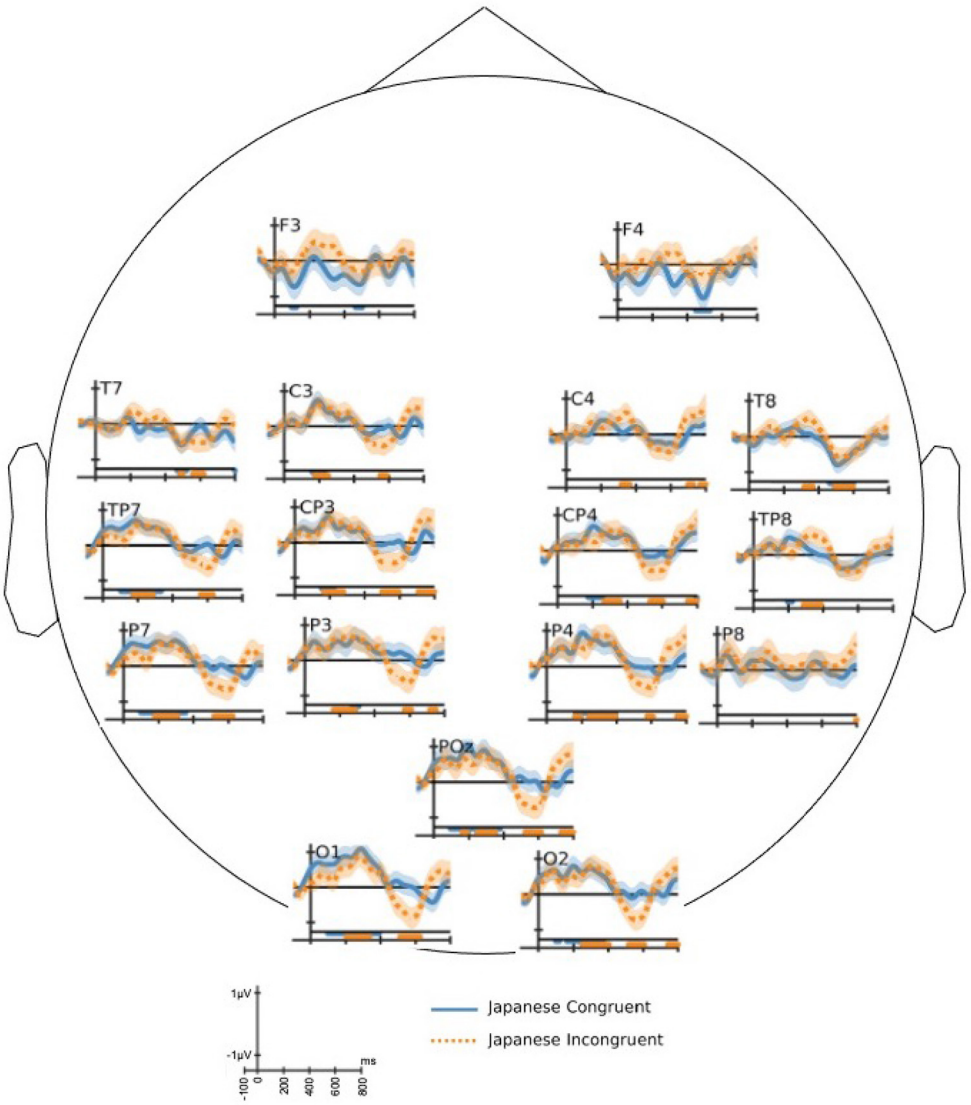

FIGURE 5 | Grand average difference event-related potential (ERP) response of Japanese Congruent end-word from Japanese end-word response before learning its meaning (solid), and Japanese in-congruent end-word from Japanese end-word response before learning its meaning (dotted). The horizontal bars drawn in the bottom of each plot signify the time regions with significant (two-sample *t*-test with $p < 0.05$) difference in the ERP response. The bars have the same color of the associated difference waveform.

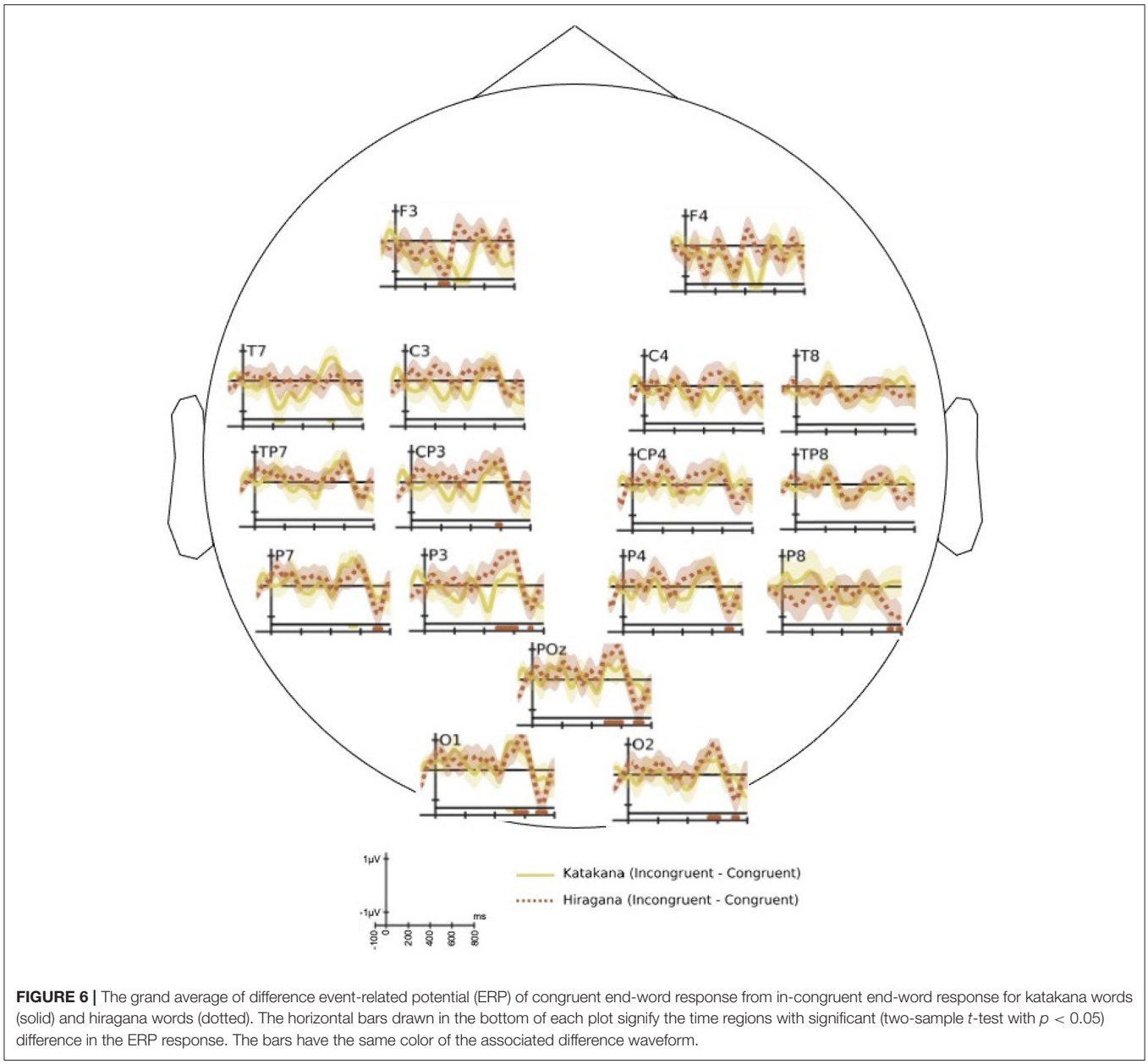

FIGURE 6 | The grand average of difference event-related potential (ERP) of congruent end-word response from in-congruent end-word response for katakana words (solid) and hiragana words (dotted). The horizontal bars drawn in the bottom of each plot signify the time regions with significant (two-sample *t*-test with $p < 0.05$) difference in the ERP response. The bars have the same color of the associated difference waveform.

these responses also have LPC after 600 ms, which possibly is an indicator of the recognition of code-switch. In summary, the ERPs for congruent [$t_{(16)} = 3.97$, $p = 5e-04$] and incongruent [$t_{(16)} = 5.89$, $p = 1.1e-05$] conditions post semantic learning elicit significantly different responses in 500–700 ms from word onset.

Figure 5 shows the differences in EEG responses to Japanese end-words before and after learning its meaning. The difference ERP wave forms of congruent (solid) and incongruent (dotted) conditions do not show any significant difference in 0–500 ms range. Both the conditions evoke significant negative peak around 200 ms. It is also noted that there is significant difference between before and after learning in the in-congruent condition than for congruent condition.

3.2.3. Effect of Phonetic Similarity to Known Words

Figure 6 shows the difference ERP response for katakana (loan words in Japanese) and hiragana words separately. It shows the difference ERP of congruent condition from incongruent condition. Both hiragana and katakana words show significant P600 response. We performed a paired *t*-test over a cluster of centro-parietal and parietal electrode locations in the time-window 500–700 ms to ascertain the statistical significance. The difference ERP of katakana words showed $t_{(16)} = 3.39$, $p = 1.87e-03$ and that of hiragana words showed $t_{(16)} = 4.18$, $p = 3.53e-04$ in the P600 window. Thus, the hiragana words show larger P600 amplitude than katakana words [$t_{(16)} = 3.91$, $p = 6.24e-04$]. The statistical significance is more established

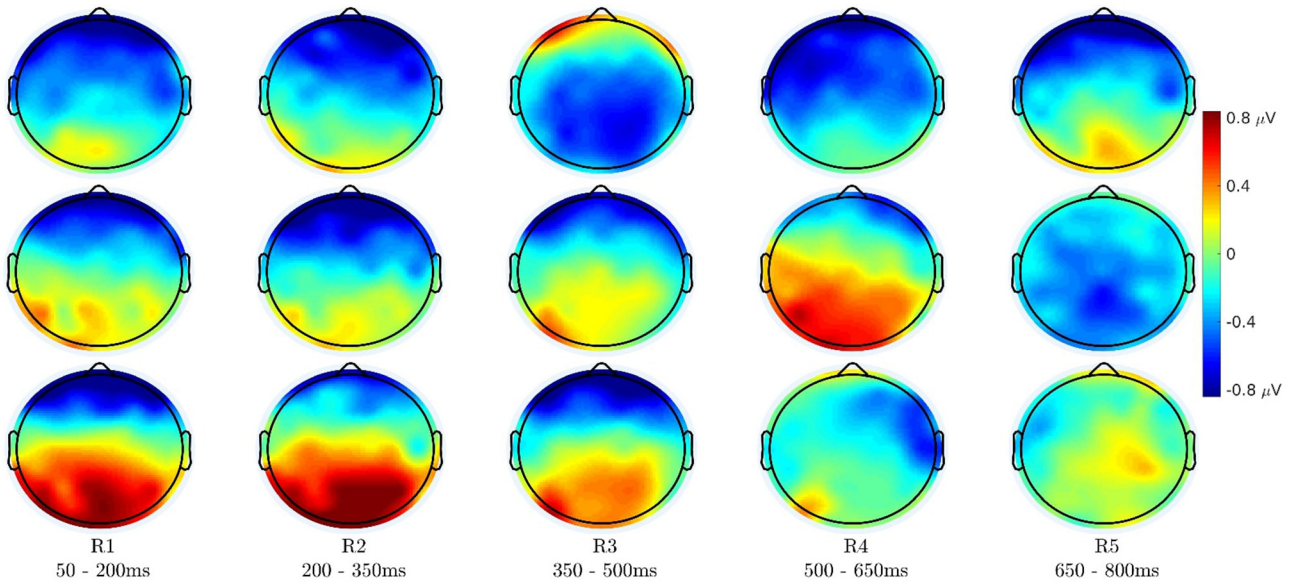

FIGURE 7 | Topography distribution of mean difference of event-related potential (ERP) (grand-average) amplitudes in different time-windows. (Top) Difference of English congruent end-word responses from English incongruent end-word responses (S1). (Middle) Difference of Japanese congruent end-word responses from Japanese incongruent end-word responses (S3). (Bottom) Difference of Japanese end-word responses before learning its meaning from the Japanese end-word responses after learning its meaning (both in congruent context).

for hiragana words in the occipital and left-parietal electrode locations.

3.3. Topographical Analysis

Figure 7 shows the topographical distribution of difference of ERP (grand-average) amplitudes in different time-windows. The mean value of ERP amplitudes in each time-window is plotted here. The top row shows the difference of English congruent end-word responses from English incongruent end-word responses, the middle row shows the difference of Japanese congruent end-word responses from Japanese incongruent end-word responses, and the bottom row shows the difference of Japanese end-word responses before learning its meaning from the Japanese end-word responses after learning its meaning.

As shown in previous works, the N400 response is significant over the centro-parietal region (see **Figure 7** top row). Similarly, the Japanese congruent vs. incongruent difference is significant over the centro-parietal region in 450–650 ms time-window as seen in **Figure 7** middle row. It is more evident in the left hemisphere than the right hemisphere of the scalp. The last row of **Figure 7** shows ERP differences between the EEG responses before and after semantic learning in frontal, parietal, and occipital regions in the initial time-windows after the end-word onset. The response over the rear part of the brain is low in magnitude from 350 ms onwards, while the response in the frontal part sustains longer. In the frontal electrodes, the ERP after semantic learning is more positive than before semantic learning. The P100, N200, and P300 components contribute to the higher positivity over the parietal and occipital regions. This is also more pronounced in the left hemisphere than the right.

The distance matrix shown in **Figure 8** is computed between the congruent and incongruent end-word ERP responses across all channel pairs (more details of the correlation analysis are given in **Supplementary Material**, Section 3). It is computed for different time-windows as shown in **Figure 8** to compare the correlation plot of the ERP waveforms for the two languages. The distance matrix in **Figure 8** shows that the English difference response in R3 (350–500 ms) has high similarity (least distance) with Japanese difference response in R4 (500–650 ms). Similarly, we observe a high similarity between the difference response of English in R4 (500–650 ms) window with difference response of Japanese in R5 (650–800 ms) window.

3.4. Statistical Analysis of ERP Effects

We have used the mean amplitudes extracted from five non-overlapping time-windows of 150 ms duration between 50 and 800 ms from the word onset in repeated-measures ANOVA. The ANOVA used four within-subject factors: language (English/Japanese), congruency (congruent/incongruent), learning (before/after), and scalp region (frontal/central/parietal). The ERP effects suggested group difference at particular topographic regions. The language and congruency had a significant interaction in all time-windows except at 200–350 ms. It should be noted that the interaction is highly significant with a larger F-ratio in the 350–500 ms window ($F = 31.09$, $p < 0.01$) and 500–650 ms window ($F = 73.08$, $p < 0.01$). The language and scalp region factors had significant interaction in two time-windows: at 50–200 and 500–650 ms. The factors of learning and scalp region had significant interactions in all time-windows except 500–650 ms. This implies that the process of learning had topographic selectivity. The

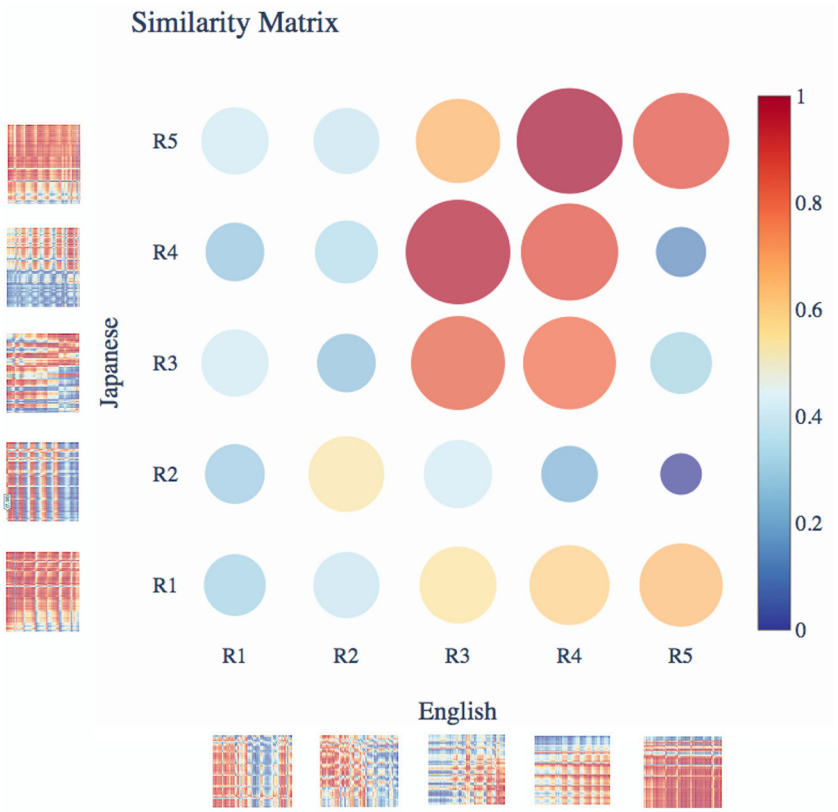

FIGURE 8 | Distance matrix between English and Japanese correlation matrices in different time-windows. The cross-channel correlation is computed between the congruent and incongruent end-word event-related potential (ERP) responses across all channels for English (S1) and Japanese (S3). The time regions in the figure are as follows R1: 50–200 ms, R2: 200–350 ms, R3: 350–500 ms, R4: 500–650 ms, and R5: 650–800 ms. The highest similarity was between English responses at R4 and Japanese responses at R5.

highest significance is observed in the 200–350 ms window ($F = 28.49$, $p < 0.01$). The congruency and scalp region also had significant interactions in the 50–200 and 500–650 ms windows.

DISCUSSION

The grand average difference of ERP response of Japanese congruent end-word from incongruent end-word showed a P600 component (**Figure 3**). This can be attributed to semantic P600 component. The work by Kuperberg et al. (2003) showed that the violation of semantic congruity in sentences with strong semantic relationship between its noun and verb can evoke a semantic P600 response. It is also worth noticing that the congruent and incongruent semantic conditions showed different brain responses for familiar language at around 400 ms and for newly acquired words at around 600 ms. The semantic differences in Japanese words evoked a later response than the English words. This may be due to the reason that the newly learned words required a reanalysis to integrate itself with the sentence. The ERP of incongruent words evoked a significant positive peak while the ERP of congruent words evoked a negative peak around 600 ms (**Figure 4**).

Figure 6 shows that both katakana and hiragana words evoked P600 component in the difference ERP. The hiragana P600 response has higher amplitude than the katakana response over the parietal and parieto-occipital electrodes. The katakana words are loan words from English, but they are pronounced with the Japanese adaptation. As shown in the behavioral responses, human subjects find it easier to recall the meaning of katakana words and hence, the involved reanalysis is not as strong as the hiragana words. The observation that hiragana words have higher P600 amplitude than katakana words is similar to the higher amplitude of N400 observed for highly unexpected end-word in the known language (Brown and Hagoort, 1993; Curran et al., 1993; Holcomb, 1993; DeLong et al., 2005).

Figures 4, **5** show the effects of semantic learning. The Japanese words in the first exposure elicited a significant P300 response owing to the novelty of the stimuli. After semantic learning, the ERP of congruent and incongruent conditions did not have significant differences in the early part of the response. The differences are significant after 500 ms from the onset of the end-word. This shows that the semantic processing of the newly acquired words may occur with a delay of more than 500 ms from the word onset.

The foreign word at the end of the sentence is comprehended as a form change, since it evokes a P600 potential instead of N400 potential. This is comparable to the semantic illusion condition shown by Brouwer et al. (2012). The code-switching to a newly learned word at the end of the sentence requires re-analysis to integrate it with the prior sentence context. As shown in other code-switching studies like Van Hell et al. (2018) and Fernandez et al. (2019), we observe that P600 is elicited for the congruent end-word in context. In this work, we show that P600 potential is evoked while perceiving newly acquired word from a foreign language employed in a code-switched manner. Further, we see a difference in the P600 latency in the response for the newly acquired end-word used in congruent and incongruent context. When the newly acquired word is used in congruent condition, we see the positive peak appearing at a slightly later time with a comparable amplitude. This difference in the response to the Japanese word used in congruent and incongruent condition shows the ERP effects of rapid semantic acquisition of foreign language words. This difference in responses for congruent and incongruent cases for the newly acquired word illustrates that the ambiguity is recognized by involving a higher cognitive load. For the newly acquired words, the word used in congruent condition evokes a lesser positive potential around 600 ms from the word onset and shows a more positive deflection in 700–800 ms (peaking around 750 ms). This can be observed in **Figure 4**. This implies that the semantic integration of newly learned words occurs much later in time compared to the similar process in a proficient language word. The underlying cognitive process may also be bi-phasic: recognition and then integration of the meaning with the sentence context.

The topographic plots show that P600 responses are also stronger over the centro-parietal and parietal regions like the N400 response. But the P600 response has a left hemisphere selectivity. The scalp distributions of difference ERP before and after semantic learning elicit significant responses in the early part of the ERP waveform. It has strong positive response over the parietal and parieto-occipital regions owing to the differences in N200 response and positive response in the frontal part owing to the P300 response. The correlation analysis in **Figure 8** shows that the highly negative correlation that exists between the EEG responses for congruent and incongruent conditions for English at about 400 ms also appear for Japanese stimuli but at a later time instant of 600 ms.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institute Human Ethics Committee (IHEC), Indian Institute of Science, Bangalore. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AS: conceptualization of this study, methodology, investigation, software, formal analysis, and writing-original draft and editing. PR: methodology and investigation. SG: conceptualization of this study, methodology, writing-review, supervision, and funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

This work was partly funded by the Ministry of Human Resource Development, Government of India, Department of Atomic Energy, Government of India (34/20/12/2018-BRNS/3408), the Pratiksha Trust and Department of Science and Technology, Government of India (DST-ECR/2017/01341).

ACKNOWLEDGMENTS

We would like to acknowledge Prof. Shihab Shamma of University of Maryland for the insightful discussions and Dr. Arun Sasidharan of Axxonet Technologies for the neuroscience discussions. We would like to thank Axxonet System Technologies, Bangalore for providing the facilities for the data collection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2022.763324/full#supplementary-material

REFERENCES

Batterink, L., and Neville, H. (2011). Implicit and explicit mechanisms of word learning in a narrative context: an event-related potential study. *J. Cogn. Neurosci.* 23, 3181–3196. doi: 10.1162/jocn_a_00013

Block, C. K., and Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: behavioral and neural validation using event-related potentials. *Behav. Res. Methods* 42, 665–670. doi: 10.3758/BRM.42.3.665

Borovsky, A., Elman, J. L., and Kutas, M. (2012). Once is enough: N400 indexes semantic integration of novel word meanings from a single exposure in context. *Lang. Learn. Dev.* 8, 278–302. doi: 10.1080/15475441.2011.614893

Borovsky, A., Kutas, M., and Elman, J. (2010). Learning to use words: Event-related potentials index single-shot contextual word learning. *Cognition* 116, 289–296. doi: 10.1016/j.cognition.2010.05.004

Brouwer, H., Fitz, H., and Hoeks, J. (2012). Getting real about semantic illusions: rethinking the functional role of the p600 in language comprehension. *Brain Res.* 1446, 127–143. doi: 10.1016/j.brainres.2012.01.055

Brown, C., and Hagoort, P. (1993). The processing nature of the N400: evidence from masked priming. *J. Cogn. Neurosci.* 5, 34–44. doi: 10.1162/jocn.1993.5.1.34

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *J. Cogn.* 2:16. doi: 10.5334/joc.72

Chikamatsu, N., Yokoyama, S., Nozaki, H., Long, E., and Fukuda, S. (2000). A Japanese logographic character frequency list for cognitive science research. *Behav. Res. Methods Instrum. Comput.* 32, 482–500. doi: 10.3758/BF03200819

Clark, E. V. (1973). "What's in a word? On the child's acquisition of semantics in his first language," in *Cognitive Development and Acquisition of Language* (Elsevier), 65–110. doi: 10.1016/B978-0-12-505850-6.50009-8

Comerchero, M. D., and Polich, J. (1999). P3A and P3B from typical auditory and visual stimuli. *Clin. Neurophysiol.* 110, 24–30. doi: 10.1016/S0168-5597(98)00033-1

Coulson, S., King, J. W., and Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Lang. Cogn. Process.* 13, 21–58. doi: 10.1080/016909698386582

Curran, T., Tucker, D. M., Kutas, M., and Posner, M. I. (1993). Topography of the N400: brain electrical activity reflecting semantic expectancy. *Electroencephalogr. Clin. Neurophysiol.* 88, 188–209. doi: 10.1016/0168-5597(93)90004-9

DeLong, K. A., Urbach, T. P., and Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* 8:1117. doi: 10.1038/nn1504

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Dikker, S., and Pylkkanen, L. (2011). Before the N400: effects of lexical-semantic violations in visual cortex. *Brain Lang.* 118, 23–28. doi: 10.1016/j.bandl.2011.02.006

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G* power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 175–191. doi: 10.3758/BF03193146

Federmeier, K. D., and Kutas, M. (1999a). Right words and left words: electrophysiological evidence for hemispheric differences in meaning processing. *Cogn. Brain Res.* 8, 373–392. doi: 10.1016/S0926-6410(99)00036-1

Federmeier, K. D., and Kutas, M. (1999b). A rose by any other name: long-term memory structure and sentence processing. *J. Mem. Lang.* 41, 469–495. doi: 10.1006/jmla.1999.2660

Fernandez, C. B., Litcofsky, K. A., and van Hell, J. G. (2019). Neural correlates of intra-sentential code-switching in the auditory modality. *J. Neurolinguist.* 51, 17–41. doi: 10.1016/j.jneuroling.2018.10.004

Friederici, A. D. (1995). The time course of syntactic activation during language processing: a model based on neuropsychological and neurophysiological data. *Brain Lang.* 50, 259–281. doi: 10.1006/brln.1995.1048

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends Cogn. Sci.* 6, 78–84. doi: 10.1016/S1364-6613(00)01839-8

Hagoort, P., Brown, C., and Groothusen, J. (1993). The syntactic positive shift (SPS) as an erp measure of syntactic processing. *Lang. Cogn. Process.* 8, 439–483. doi: 10.1080/01690969308407585

Hagoort, P., and Brown, C. M. (2000). ERP effects of listening to speech: semantic ERP effects. *Neuropsychologia* 38, 1518–1530. doi: 10.1016/S0028-3932(00)00052-X

Hoeks, J. C., Stowe, L. A., and Doedens, G. (2004). Seeing words in context: the interaction of lexical and sentence level information during reading. *Cogn. Brain Res.* 19, 59–73. doi: 10.1016/j.cogbrainres.2003.10.022

Holcomb, P. J. (1993). Semantic priming and stimulus degradation: implications for the role of the N400 in language processing. *Psychophysiology* 30, 47–61. doi: 10.1111/j.1469-8986.1993.tb03204.x

Kaan, E., Harris, A., Gibson, E., and Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Lang. Cogn. Process.* 15, 159–201. doi: 10.1080/016909600386084

Kuperberg, G. R., Sitnikova, T., Caplan, D., and Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cogn. Brain Res.* 17, 117–129. doi: 10.1016/S0926-6410(03)00086-7

Kutas, M., and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.* 62, 621–647. doi: 10.1146/annurev.psych.093008.131123

Kutas, M., and Hillyard, S. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* 207, 203–205. doi: 10.1126/science.7350657

Kutas, M., and Hillyard, S. A. (1983). Event-related brain potentials to grammatical errors and semantic anomalies. *Mem. Cogn.* 11, 539–550. doi: 10.3758/BF03196991

Kutas, M., and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307:161. doi: 10.1038/307161a0

Kutas, M., Neville, H. J., and Holcomb, P. J. (1987). A preliminary comparison of the N400 response to semantic anomalies during reading, listening and signing. *Electroencephalogr. Clin. Neurophysiol. Suppl.* 39, 325–330.

Leckey, M., and Federmeier, K. D. (2020). The P3b and P600 (s): positive contributions to language comprehension. *Psychophysiology* 57:e13351. doi: 10.1111/psyp.13351

Luck, S. J., Vogel, E. K., and Shapiro, K. L. (1996). Word meanings can be accessed but not reported during the attentional blink. *Nature* 383, 616–618. doi: 10.1038/383616a0

McCallum, W., Farmer, S., and Pocock, P. (1984). The effects of physical and semantic incongruites on auditory event-related potentials. *Electroencephalogr. Clin. Neurophysiol.* 59, 477–488. doi: 10.1016/0168-5597(84)90006-6

Mestres-Missé, A., Rodriguez-Fornells, A., and Münte, T. F. (2007). Watching the brain during meaning acquisition. *Cereb. Cortex* 17, 1858–1866. doi: 10.1093/cercor/bhl094

Moreno, E. M., Federmeier, K. D., and Kutas, M. (2002). Switching languages, switching palabras (words): an electrophysiological study of code switching. *Brain Lang.* 80, 188–207. doi: 10.1006/brln.2001.2588

Moreno, E. M., Rodríguez-Fornells, A., and Laine, M. (2008). Event-related potentials (ERPs) in the study of bilingual language processing. *J. Neurolinguist.* 21, 477–508. doi: 10.1016/j.jneuroling.2008.01.003

Nigam, A., Hoffman, J. E., and Simons, R. F. (1992). N400 to semantically anomalous pictures and words. *J. Cogn. Neurosci.* 4, 15–22. doi: 10.1162/jocn.1992.4.1.15

Nobre, A. C., and Mccarthy, G. (1995). Language-related field potentials in the anterior-medial temporal lobe: II. Effects of word type and semantic priming. *J. Neurosci.* 15, 1090–1098. doi: 10.1523/JNEUROSCI.15-02-01090.1995

Olah, B. (2007). English loanwords in Japanese: effects, attitudes and usage as a means of improving spoken english ability. *Bunkyo Gakuin Daigaku Ningen Gakubu Kenkyū Kiyo* 9, 177–188.

Osterhout, L., and Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *J. Mem. Lang.* 31, 785–806. doi: 10.1016/0749-596X(92)90039-Z

Perfetti, C. A., Wlotko, E. W., and Hart, L. A. (2005). Word learning and individual differences in word learning reflected in event-related potentials. *J. Exp. Psychol.* 31:1281. doi: 10.1037/0278-7393.31.6.1281

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* 118, 2128–2148. doi: 10.1016/j.clinph.2007.04.019

Soman, A., Madhavan, C., Sarkar, K., and Ganapathy, S. (2019). An EEG study on the brain representations in language learning. *Biomed. Phys. Eng. Exp.* 5, 25–41. doi: 10.1088/2057-1976/ab0243

Speidel, G. E., and Nelson, K. E. (2012). *The Many Faces of Imitation in Language Learning, Vol. 24.* Springer Science & Business Media.

Szűcs, D., Soltész, F., Czigler, I., and Csépe, V. (2007). Electroencephalography effects to semantic and non-semantic mismatch in properties of visually presented single-characters: the N2b and the N400. *Neurosci. Lett.* 412, 18–23. doi: 10.1016/j.neulet.2006.08.090

Tanner, D., Grey, S., and van Hell, J. G. (2017). Dissociating retrieval interference and reanalysis in the P600 during sentence comprehension. *Psychophysiology* 54, 248–259. doi: 10.1111/psyp.12788

Van Hell, J. G., Fernandez, C. B., Kootstra, G. J., Litcofsky, K. A., and Ting, C. Y. (2018). Electrophysiological and experimental-behavioral approaches to the study of intra-sentential code-switching. *Linguist. Approach. Biling.* 8, 134–161. doi: 10.1075/lab.16010.van

Van Petten, C., and Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Mem. Cogn.* 18, 380–393. doi: 10.3758/BF03197127

Wlotko, E. W., Lee, C.-L., and Federmeier, K. D. (2010). Language of the aging brain: event-related potential studies of comprehension in older adults. *Lang. Linguist. Compass* 4, 623–638. doi: 10.1111/j.1749-818X.2010.00224.x

Zwaan, R. A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., et al. (2018). Participant nonnaiveté and the reproducibility of cognitive psychology. *Psychon. Bull. Rev*. 25, 1968–1972. doi: 10.3758/s13423-017-1348-y

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

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

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership