

Article

Familiarity of Background Music Modulates the Cortical Tracking of Target Speech at the “Cocktail Party”

Jane A. Brown^{1,2}  and Gavin M. Bidelman^{1,2,3,4,*} ¹ School of Communication Sciences and Disorders, University of Memphis, Memphis, TN 38152, USA² Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152, USA³ Department of Speech, Language and Hearing Sciences, Indiana University, Bloomington, IN 47408, USA⁴ Program in Neuroscience, Indiana University, Bloomington, IN 47405, USA

* Correspondence: gbidel@indiana.edu

Abstract: The “cocktail party” problem—how a listener perceives speech in noisy environments—is typically studied using speech (multi-talker babble) or noise maskers. However, realistic cocktail party scenarios often include background music (e.g., coffee shops, concerts). Studies investigating music’s effects on concurrent speech perception have predominantly used highly controlled synthetic music or shaped noise, which do not reflect naturalistic listening environments. Behaviorally, familiar background music and songs with vocals/lyrics inhibit concurrent speech recognition. Here, we investigated the neural bases of these effects. While recording multichannel EEG, participants listened to an audiobook while popular songs (or silence) played in the background at a 0 dB signal-to-noise ratio. Songs were either familiar or unfamiliar to listeners and featured either vocals or isolated instrumentals from the original audio recordings. Comprehension questions probed task engagement. We used temporal response functions (TRFs) to isolate cortical tracking to the target speech envelope and analyzed neural responses around 100 ms (i.e., auditory N1 wave). We found that speech comprehension was, expectedly, impaired during background music compared to silence. Target speech tracking was further hindered by the presence of vocals. When masked by familiar music, response latencies to speech were less susceptible to informational masking, suggesting concurrent neural tracking of speech was easier during music known to the listener. These differential effects of music familiarity were further exacerbated in listeners with less musical ability. Our neuroimaging results and their dependence on listening skills are consistent with early attentional-gain mechanisms where familiar music is easier to tune out (listeners already know the song’s expectancies) and thus can allocate fewer attentional resources to the background music to better monitor concurrent speech material.



Citation: Brown, J.A.; Bidelman, G.M. Familiarity of Background Music Modulates the Cortical Tracking of Target Speech at the “Cocktail Party”. *Brain Sci.* **2022**, *12*, 1320. <https://doi.org/10.3390/brainsci12101320>

Academic Editor: Antoine Shahin

Received: 27 August 2022

Accepted: 27 September 2022

Published: 29 September 2022

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: auditory evoked potentials (ERPs); speech in noise (SIN); familiarity; music perception

1. Introduction

Listeners are constantly faced with the challenge of listening to speech in noisy environments. This so-called “cocktail party” problem is often studied using noise or multi-talker babble maskers. However, many realistic cocktail party scenarios also involve music (e.g., coffee shops, concerts), which is not often considered in studies of auditory scene analysis. The effect of background music on concurrent speech/linguistic tasks is mixed and dependent on many factors, including type of music, participant characteristics, task structure (reviewed by [1]), and cognitive faculties. For example, music presented concurrently with a memorization task impairs performance, but only for listeners who prefer to study without background music [2]. Fast-tempo background music increases spatial processing speed and linguistic processing accuracy [3], but can also disrupt reading comprehension [4]. Music with vocals also more negatively affects concurrent tasks across various cognitive modalities [2,5–10]. Linguistic content in the masker introduces informational masking, which in turn interferes with cognitive resources needed to complete the task.

In typical (i.e., speech-on-speech) cocktail party tasks, the familiarity of a talker can be advantageous for speech recognition [11,12] when the familiar voice is either the target or the masker [13]. A familiar voice is retained implicitly [12], which allows for more efficient processing of novel words or sentences spoken in that voice [14]. However, the role of familiarity for music in perceptual–cognitive tasks is not well known. It is also worth noting that many studies define “familiarity” inconsistently (e.g., exposure training of naïve listeners versus songs that listeners already know), which limits comparisons between results. Still, familiar music maskers can improve various linguistic behavioral measures [15,16], but may be detrimental to foreign language learning [17]. Previous work from our lab [5] has shown a negative familiarity effect of background noise on concurrent speech recognition. In a music-on-speech cocktail party task, we found speech recognition performance was worse during familiar compared to unfamiliar music maskers, likely due to the increased cognitive load of the familiar music (i.e., those songs were more distracting). However, that prior work was solely behavioral and did not provide insight into the neural underpinnings of those perceptual–cognitive effects.

Besides indexical attributes of the signal, demographic properties of the *listener* also modulate cocktail party perception [18]. In particular, music expertise has been widely shown to alter auditory-cognitive brain structure and function, providing a “musician advantage” in various listening skills [19]. This is especially evident in speech-in-noise tasks where musicians show better degraded speech perception and more successful suppression of acoustic distractors [20–24]. Musicians’ improved speech-in-noise abilities might result from their superiority juggling multiple auditory streams [21,25] and lesser susceptibility to informational masking than their non-musician peers [24,26]. Musicians have more experience with auditory stream segregation (e.g., parsing a melody from harmonies, hearing one’s own melody in an orchestra), which in turn seems to enhance the parsing of degraded speech [27]. However, in contrast to speech-on-speech, musicians are more affected by background music than non-musicians [28]. Importantly, the “musician advantage” for speech-in-noise listening is not dependent on formal musical training. Mankel and Bidelman [29] demonstrated similar effects in highly musical people without formal musical training, indicating that superior cocktail party listening skills may be attributed more to general listening abilities rather than music experiences/training, *per se*.

Extending this prior work, the current study sought to further investigate the role of background music familiarity and presence of vocals on concurrent speech perception. In a variant of the cocktail party task, participants listened to an audiobook (speech) in the presence of popular music maskers that varied in their familiarity to the listener and content of the original audio recording (i.e., with or without vocals). Here, we refer to familiarity as acquired in repeated exposure in everyday life, including on the radio. Our design departs from previous studies investigating the effects of music-on-speech intelligibility which have predominantly used synthesized music or music-like stimuli [30,31] which, though easier to control, are not as ecologically valid as the popular music used here. Participants answered comprehension questions about the story to ensure task engagement with target speech material. We simultaneously recorded multichannel EEG and measured neural tracking to the speech envelopes using temporal response functions (TRFs) [32]. In accordance with our previous behavioral work [5], we hypothesized that the perception and neural tracking of speech would diminish (i.e., lower comprehension scores, TRFs with weaker amplitudes and longer latencies) when concurrent background music was familiar to listeners and when it contained vocals. These findings would suggest that speech perception suffers from a concurrent linguistic masker even from a different domain (i.e., music), as well as stronger attentional (mis)allocation to background music when it is familiar to the listener.

2. Materials and Methods

2.1. Participants

The sample included $n = 17$ young adults ages 21–32 ($M = 25$, $SD = 2.6$ years, 4 male). This sample size is comparable with previous studies investigating continu-

ous speech processing with TRFs [33–35]. 12 participants reported having musical training ($M = 10.25$ years, $SD = 5.1$). All participants showed audiometric thresholds better than 20 dB HL (octave frequencies, 250–8000 Hz) and reported English as their native language. Listeners were primarily right-handed (mean 75% laterality using the Edinburgh Handedness Inventory [36]). Each was paid for their time and gave written informed consent in compliance with a protocol approved by the IRB at the University of Memphis.

2.2. Stimuli and Task

During EEG recording, we measured neural speech tracking and comprehension by presenting a continuous audiobook in different background music conditions. The audiobook (taken from librivox.org, accessed on 9 August 2021) was *Doctor* by Murray Leinster, read by a male speaker. Silences longer than 300 milliseconds were shortened to decrease extended silence in the stimulus while still sounding like natural speech [37]. The speech signal was RMS amplitude normalized and separated into 20 successive 2 min segments.

Music stimuli were a subset of those used in Brown and Bidelman [5], which had been previously identified as being familiar (“Just Dance” by Lady Gaga) or unfamiliar (“Play with Fire” by Hilary Duff) to a cohort of young, normal-hearing listeners. We used a machine learning algorithm trained to separate instrumental and vocal tracks (lalal.ai) to isolate the instrumentals from the full unprocessed song [5]. All music files (two songs each with the full song and isolated instrumentals) were sampled at 44100 Hz, converted to mono for diotic presentation, and RMS amplitude normalized after removing silences to equate the sound level across clips.

For each participant, the 20 audiobook segments were presented sequentially, separated into four runs (totaling eight minutes of listening per condition). Each run contained five audiobook segments randomly presented with each song condition (familiar with vocals, familiar without vocals, unfamiliar with vocals, and unfamiliar without vocals) and in silence. In all, listeners heard each song condition four times. Participants were instructed to listen to the audiobook and ignore the background music. Each trial was two minutes, after which participants answered one comprehension question presented on a computer screen (20 questions total across the full experiment). After completing the task, participants rated each song on a scale from 1 (*Not at all familiar*) to 5 (*Extremely familiar*) to gauge prior familiarity with the music stimuli.

After completing the EEG task, participants completed the Profile of Music Perception Skills (PROMS) [38] to measure their musical listening skills. We have previously shown that high-PROMS-scoring individuals (“musical sleepers”) have enhanced speech processing akin to trained musicians despite having no formal musical training [29]. The PROMS contains eight subtests focusing on different musical domains (e.g., rhythm, melody, timbre, etc.). For each subtest, participants heard two tokens (e.g., two melodies) and indicated whether they were the same or different. The scores were on a 5-point Likert scale, where correctly identifying “definitely same/different” was given one point and “probably same/different” was worth one-half point. The maximum possible test score was 80 (8 subtests, 10 items each).

2.3. EEG Recording Procedures

Participants were seated in an electrically shielded, sound-attenuated booth for the duration of the experiment. Continuous EEG recordings were obtained from 64 channels aligned in the 10-10 system [39] and digitized using a sample rate of 500 Hz (SynAmps RT amplifiers; Compumedics Neuroscan, Charlotte, NC, USA). Contact impedances were maintained <10 k Ω . Music and speech stimuli were each presented diotically at 70 dB SPL through electromagnetically shielded ER-2 insert headphones (Etymotic Research, Elk Grove Village, IL, USA), resulting in a signal-to-noise ratio of 0 dB. Stimulus presentation was controlled with a custom MATLAB program (v. 2021a; MathWorks, Natick, MA, USA) and routed through a TDT RZ6 digital signal processor (Tucker-Davis Technologies, Alachua, FL, USA). EEGs were re-referenced to average mastoids and pre-processed in

accordance with the recommendations provided by Crosse, et al. [40]. Data from 0 to 1000 milliseconds after the onset of each two min epoch were discarded to avoid transient brain responses in the subsequent analysis [40]. Epochs were then concatenated for each of the five conditions, resulting in eight minutes of continuous data per condition.

2.4. Behavioral Data Analysis

We logged the comprehension question response at the end of each presentation. Questions were scored as a binary “correct” or “incorrect” label.

2.5. Electrophysiological Data Analysis: Temporal Response Functions (TRFs)

We analyzed continuous neural tracking to the speech signal using the Temporal Response Function toolbox in MATLAB [32]. The TRF is a linear function representing the (deconvolved) impulse response to a continuous stimulus. To measure EEG tracking to the speech, we extracted the temporal envelope of the audiobook via the Hilbert transform. EEG data were down-sampled to 250 Hz, then filtered between 1 and 30 Hz to isolate cortical activity to the low-frequency speech envelope. EEG and stimulus signals were both z-score normalized. As with conventional event-related potentials (ERPs), TRFs were computed for each participant to account for inherent inter-subject variability in neural response tracking [40]. We used 6-fold cross-validation to derive TRFs per condition. Ridge regression [41] was used to identify the optimal λ smoothing parameter of the forward model for the speech-only condition. Model tuning was conducted using the speech-only condition to optimize the TRF to the clean (unmasked) speech. We used a fronto-central channel cluster (F1, Fz, F2, FC1, FCz, FC2, C1, Cz, C2) to further optimize the model fit to canonical topography of auditory ERPs. For each participant, the optimal λ was taken as the ridge parameter yielding the highest reconstruction of simulated neural response (i.e., correlation r-value between actual EEG and TRF-derived responses). We then used each participant’s optimal λ parameter to derive TRFs for all other conditions. This approach preserves response consistency within subjects while avoiding model overfitting [40]. The resulting TRF waveforms represent the EEG signal at each electrode changes in response to a unit change in the speech stimulus envelope.

We analyzed TRFs (i.e., RMS amplitude and latency) between 100–150 ms corresponding to the “N1” wave of the canonical auditory ERP (see Figure 1). The N1 was selected as it reflects the early arrival of sound information in auditory cortex and is also modulated by attention [42–44]. Previous EEG studies have also demonstrated that noise has the largest effect on speech TRFs within this time window [45]. To further investigate possible hemisphere differences, we created two homologous channel clusters over front right (Fz, F2, F4, F6, F8, FC6, FT8) and front left (Fz, F1, F3, F5, F7, FC5, FT7) scalp regions.

2.6. Statistical Analyses

Statistics were run in R using the lme4 package [46]. For all analyses, we used mixed-effects models with fixed factors of familiarity (2 levels: familiar, unfamiliar) and song condition (2 levels; with vocals, without vocals). Subjects and trial served as random factors (where applicable). Because the behavioral response was a binary score (correct vs. incorrect), we analyzed those data using a generalized linear mixed-effects model ANOVA with binomial link function. Note that the Wald statistic is used to determine significant factor(s) effects for these models instead of conventional *F*-statistics (used in lme4 package; [46]). Peak amplitudes and latencies were normally distributed and thus analyzed using conventional linear mixed models and *F*-statistics. Multiple comparisons were corrected with Tukey adjustments. For all measures (score, latency, and amplitude), there were no hemisphere differences (all *p* values > 0.17), so subsequent analyses used data pooled between the two electrode clusters.

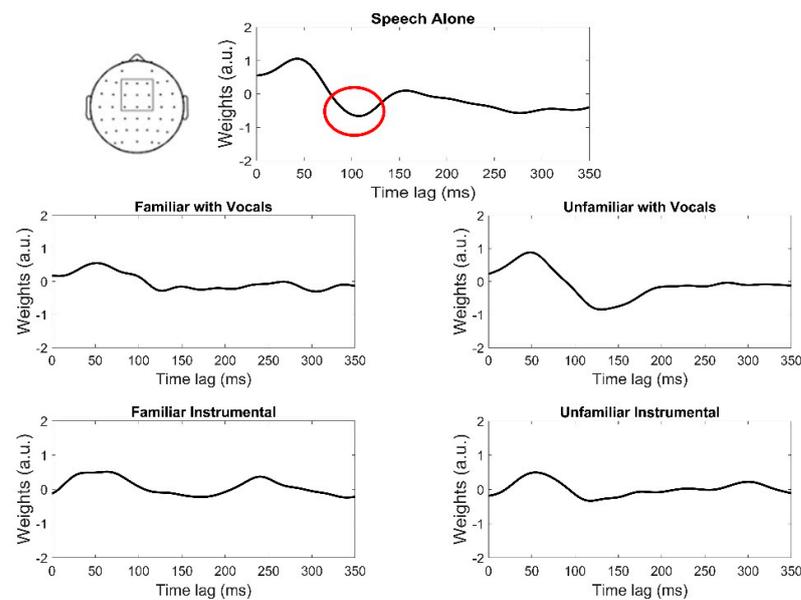


Figure 1. Grand average TRF waveforms (plotted at channel FCz) across music masker conditions. The model was trained using a fronto-central cluster of electrodes (top left). Red circle = region of interest for analyses corresponding to the auditory N1.

3. Results

3.1. Behavioral Data

Participants showed stark differences in their familiarity ratings across music selections ($t(16) = 19.13, p < 0.001$), validating our first stimulus manipulation. Speech comprehension scores were subject to a strong masking effect; comprehension was (expectedly) better for speech presented in silence compared to all other music-masked conditions ($t(338) = 3.23, p = 0.001$). An ANOVA based on the generalized binomial model showed no interaction between familiarity and song condition ($\chi^2(1, N = 17) = 1.26, p = 0.26$). There was no main effect of familiarity ($\chi^2(1, N = 17) = 0.41, p = 0.52$). There was, however, a significant effect of song condition ($\chi^2(1, N = 17) = 5.42, p = 0.02$), whereby behavioral performance was overall poorer during vocal vs. non-vocal music maskers (Figure 2). These results confirm the effectiveness of music in masking target speech recognition as well as the added hinderance of music containing vocal (linguistic) information [5].

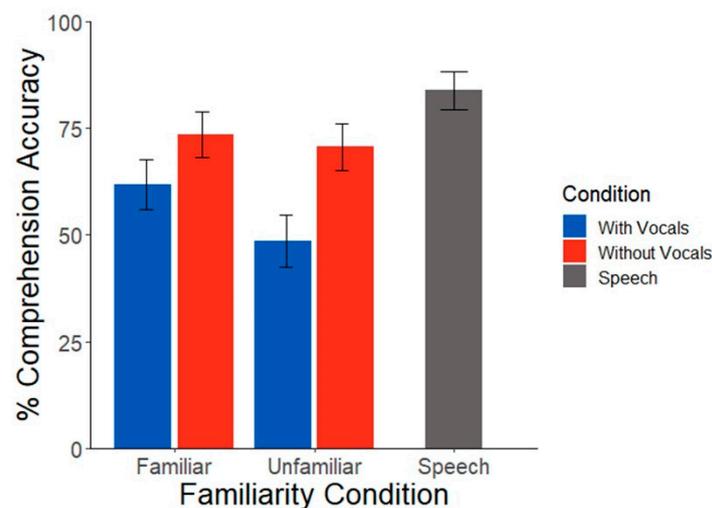


Figure 2. Speech comprehension scores during background music as a function of music familiarity and vocal condition. Speech recognition was poorer during concurrent vocal vs. non-vocal music. Error bars represent ± 1 s.e.m.

3.2. Electrophysiological Data

TRF latency and amplitude are shown across conditions in Figure 3. Music-masked speech showed longer latencies than speech presented in silence ($t(83) = 3.40, p = 0.001$). An ANOVA conducted on TRF latencies revealed an interaction between familiarity and condition ($F(1,48) = 6.18, p = 0.016, \eta^2 = 0.11$). Post hoc tests showed this interaction was attributable to longer speech-evoked TRF latencies in music with vocals than instrumentals alone ($t(48) = 2.54, p = 0.015$), but only in unfamiliar music. This vocal vs. instrumental latency difference was not observed for familiar music ($p = 0.32$).

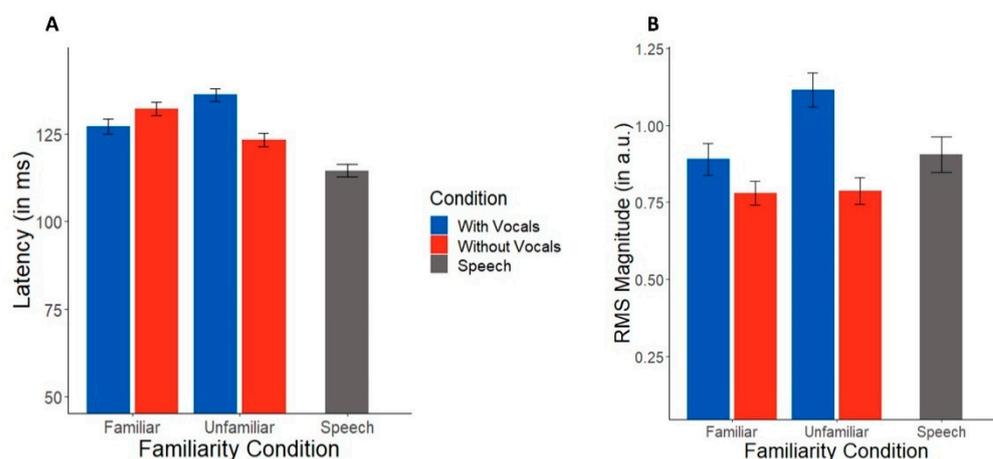


Figure 3. TRF N1 latencies and magnitudes across music masking conditions. (A) Latencies were longer for songs with vocals than instrumentals only during unfamiliar music. (B) Neural tracking of target speech was also stronger during music with vocals in both familiarity conditions. Error bars represent ± 1 standard error of the mean.

In contrast to latency measures, TRF amplitudes were not affected by masking and concurrent music (masking effect: $t(83) = 0.11, p = 0.91$). An ANOVA conducted on TRF RMS amplitudes, indicating the strength of speech tracking, showed a sole main effect of condition ($F(1,48) = 5.13, p = 0.028, \eta^2 = 0.10$); responses were larger in music with vocals as compared to instrumentals. There was no effect of familiarity ($F(1,48) = 1.45, p = 0.23$) nor a condition*familiarity interaction ($F(1,48) = 1.28, p = 0.26$).

3.3. Neural Speech Tracking as a Function of Listeners' Musicality

We next asked whether speech-envelope tracking amidst music (as indexed via TRFs) varied as a function of listeners' music-listening skills (as indexed by their PROMS scores). Previous studies demonstrate that individuals who lack formal music training but who nonetheless have superior auditory skills show advantages with speech identification and cocktail party processing [29,47]. As in Mankel and Bidelman [29], we divided our participants into two groups—"high PROMS" and "low PROMS"—using a median split of their PROMS musicality scores. The groups did not differ in age ($t(15) = 1.28, p = 0.22$), years of education ($t(15) = 0.82, p = 0.43$), or sex (Fisher's exact test; $p = 0.294$). The high PROMS group had 11.4 ($SD = 5.1$) years of musical training as compared to the low PROMS group ($M = 3.6, SD = 5.2$ years; $t(15) = 3.13, p = 0.01$). To account for this difference in training, we ran our omnibus ANOVA models with the three factors of interest (familiarity, vocals condition, and PROMS level) and years of training as a covariate. In addition to dichotomizing the musicality variable [48], we also ran models treating PROMS score as a continuous variable.

3.3.1. Behavioral Data

We first tested for differences between PROMS scores and familiarity ratings to assess whether high vs. low musicality listeners were more/less familiar with the stimuli used

in our experiment. There was no group difference for either the familiar ($t(15) = 1.63$, $p = 0.125$) or unfamiliar ($t(15) = 1.24$, $p = 0.236$) stimuli, indicating listeners were similarly (un)familiar with our song selections.

Analysis of speech comprehension during the EEG task revealed a significant condition \times group interaction ($\chi^2(1) = 4.12$, $p = 0.042$), as well as a marginal familiarity \times condition \times group interaction ($\chi^2(1) = 3.63$, $p = 0.056$) (Figure 4). Group differences were also partially driven by years of musical training ($\chi^2(1) = 11.15$, $p = 0.001$). To help interpret these complex interactions, we conducted separate 2-way ANOVAs by group to assess the impact of music familiarity and condition on speech recognition in low vs. high PROMS listeners. High PROMS listeners' comprehension was invariant to condition and familiarity effects (all p values > 0.065). However, the low PROMS group showed a familiarity \times condition interaction ($\chi^2(1) = 3.63$, $p = 0.042$). Low musicality listeners showed poorer comprehension in music with vocals than without vocals but only during unfamiliar music ($z = 2.44$, $p = 0.01$). This vocal effect was not present for familiar music ($z = 2.20$, $p = 0.69$). When musicality was treated as continuous variable, there were no significant effects of any variable of interest on comprehension score (all p values > 0.311).

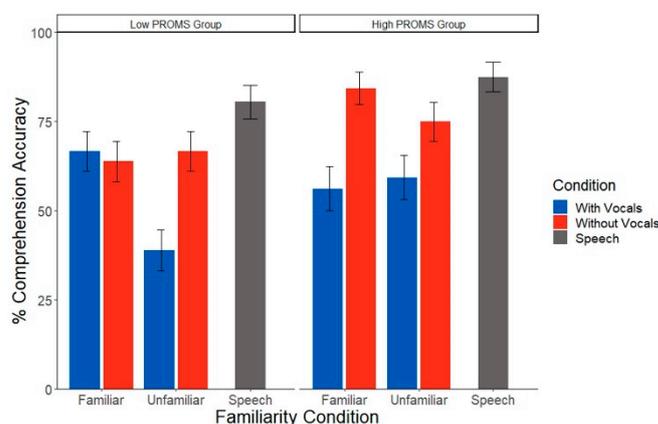


Figure 4. Speech comprehension performance varies with listeners' musical skill level. Less musical listeners performed worse in music with vocals only in the unfamiliar music condition. However, more musical listeners did not show any effects of familiarity or song condition. PROMS = Profile of Music Perception Skills test. Error bars represent ± 1 standard error of the mean.

3.3.2. Electrophysiological Results

TRF latencies showed a significant interaction between familiarity and condition ($F(1,45) = 6.00$, $p = 0.018$, $\eta^2 = 0.12$), where latencies for vocals were longer than instrumentals for only the unfamiliar music. There was no effect of musicality ($F(1,15) = 1.85$, $p = 0.194$). However, when treating musicality as a continuous variable, there was a familiarity \times musicality interaction ($F(1,45) = 10.17$, $p = 0.003$, $\eta^2 = 0.18$). TRF latencies were longer for unfamiliar music ($t(45) = 2.60$, $p = 0.013$) for the less musical listeners (i.e., lower PROMS scores) but were longer for familiar songs in higher PROMS scoring individuals ($t(45) = 2.63$, $p = 0.012$).

For TRF amplitudes, the omnibus ANOVA revealed an overall effect of condition ($F(1,45) = 4.88$, $p = 0.032$), where all listeners showed larger amplitudes for music with vocals (Figure 5B). There were no effects of familiarity or PROMS group (all p values > 0.1). There was also not an effect of musical training ($F(1,15) = 0.03$, $p = 0.88$), indicating again that the condition effect is not driven by experience. Treating musicality as a continuous variable showed no significant effects for any of the variables of interest (all p values > 0.067).

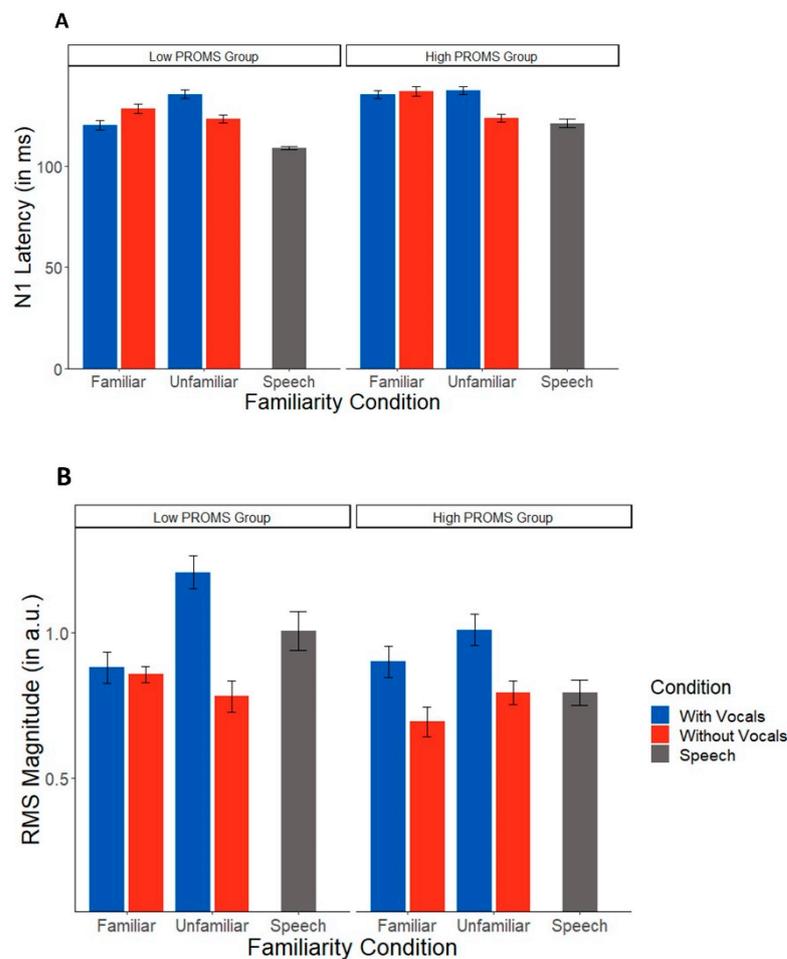


Figure 5. Speech TRFs vary as a function of musical aptitude (controlling for musical training). (A) Latencies were shorter for instrumental music than vocals only in the unfamiliar condition. (B) Instrumental music produced smaller speech responses than songs with vocals in both groups. PROMS = Profile of Music Perception Skills test. Error bars represent ± 1 standard error of the mean.

4. Discussion

As an innovative extension of the cocktail party problem, we compared speech comprehension and neural tracking of target speech amidst various music backdrops. We also manipulated the familiarity and vocals (i.e., song with lyrics or only instrumentals) of the music to evaluate how content and listeners' familiarity of concurrent music backgrounds affect concurrent speech perception and its neural processing. Perception and neural encoding of speech was worse during music with vocals than solely instrumentals. However, the impact of vocals on speech coding varied based on the familiarity of the background music. These findings indicate the monitoring of speech concurrent with music containing vocals might be more challenging for unfamiliar tunes. Our data suggest it is more difficult (i.e., harder for the brain to suppress the music masker) during certain types (unfamiliar, vocal) of music backdrops, likely through increased susceptibility to linguistic interference and/or misallocation of attention between speech and music streams. Moreover, these effects were exacerbated when accounting for musicality, suggesting listeners' inherent auditory skills also impact their cocktail party speech processing. In accordance with our previous behavioral study [5], we found speech comprehension was further impaired in music containing vocals than in instrumental music. We attribute this decline to the informational masking introduced by the linguistic content of the vocals [49–51].

At the neural level, TRFs showed that brain tracking of speech was modulated by both familiarity and song condition. Overall, the presence of music prolonged TRF latencies

to speech, which is consistent with well-known masking effects observed in previous auditory EEG studies [45,52]. More critically, we found that condition and familiarity had an interactive effect on speech-evoked TRFs; neural latencies were strongly modulated by vocals relative to instrumental music, but only for unfamiliar music. Responses were also larger during concurrent vocal compared to instrumental music. Intuitively, larger evoked responses are typically associated with a stronger representation or encoding of the speech signal (i.e., “bigger is better”) [53]. However, we found here that songs with vocals showed *worse* comprehension and longer neural latencies in conjunction with larger N1 responses. Indeed, TRF latencies were negatively correlated with speech recognition performance. It is well-established that larger N1 amplitudes are a marker of increased attention [42], analogous to the M100 peak in TRFs derived from neuromagnetic recordings [54]. Thus, the larger N1 responses to target speech we find in these difficult music conditions may reflect attentional load due to the increased listening demand of parsing speech from concurrent (especially unfamiliar) music. Larger N1 may also reflect increased listening effort [55]. Indeed, overexaggerated N1 to speech is also indicative of increased listening effort during speech processing, as observed in older adults with cognitive impairments [56]. The stronger effects on speech processing we observe during unfamiliar vocal music might therefore reflect the influences of selective attention [42], with increased effort needed to maintain that attention in more difficult listening conditions. This may also be comparable to a recent study that found a larger N400 during reading comprehension masked by music, reflecting increased semantic processing effort [57]. Still, the latency of neural effects observed here (~100 milliseconds) suggests that music challenges speech perception much earlier in the processing hierarchy.

Interestingly, we show that speech tracking at the “cocktail party” varies depending on the inherent musical skills of the listener. We found that low PROMS listeners were more impacted by unfamiliar music than the high PROMS group. Several studies have showed that more familiar music facilitates concurrent linguistic tasks by increasing arousal [58] and generating expectancies [16,52]. Musical ability is associated not only with speech-in-noise processing advantages [24,29,59], but also parsing complex auditory scenes [25,60]. Indeed, there is some evidence that trained musicians also more successfully deploy attention in auditory and even non-auditory perceptual tasks [24,61,62], including those related to cocktail party listening [21]. Moreover, non-musicians are more susceptible to informational and linguistic masking [21,26]—though see [63]. Here, low PROMS listeners may have been more distracted by the unfamiliar background music as a more challenging listening condition, which then becomes exacerbated by the presence of vocals. In this sense, less musical listeners (i.e., those with poorer auditory perceptual skills) might experience increased informational masking compared to their more musical peers. Our findings thus support notions that musical ability impacts cocktail party speech listening and one’s susceptibility to informational masking [21,26,64] but extend prior work by demonstrating such effects are not necessarily attributable to musical experience, per se, but instead depend on inherent listening abilities.

Our results can be explained by certain attentional load theories [65], which posit that selective attention is comprised of a passive perception mechanism and cognitive control mechanism. In a more demanding listening situation, more cognitive resources are recruited, meaning that there is less cognitive capacity to suppress distractors. In this study, listening to the target speech in more distracting music maskers (i.e., unfamiliar/vocal songs) was a difficult task, creating a larger cognitive load. This results in a diminished capacity to suppress the music masker, which makes it more difficult to attend to the target speech. The directional differences between the musicality groups also implies that cognitive load is variable and dependent on listening expertise. Under this interpretation, speech perception in unfamiliar music may have created a larger cognitive load for less musical listeners than for the high PROMS group, resulting in longer N1 latencies and worse comprehension scores.

In our previous study [5], we found that comprehension was *worse* in unfamiliar music, which is the opposite of our current results. Though the two studies used the same music maskers, the task here was more complex, assessing speech comprehension rather than target word recognition (as in Brown and Bidelman, 2022). Additionally, the task in this study was completed post-stimulus while the previous word-recognition task required increased online processing (i.e., monitoring key words in sentences). These task differences may contribute to the contradictory findings on familiarity effects. Indeed, the specific nature of the task can yield differential effects of music on concurrent speech processing, sometimes with opposite directions of effects [5,15–17,59].

It is of note that most prior studies used years of formal music training (self-reported) as a metric for defining musicians and non-musicians, while we solely used aptitude scores. Though high- and low-PROMS groups were separable based on their years of music training, musicality group differences remained significant while controlling for training, meaning that the effects found in this study likely result from some combination of experience and natural auditory skills. Our data are consistent with emerging notions that listeners' inherent, rather than acquired, musicality affect their speech-in-noise processing abilities [29,47].

In this experiment, we used pop songs as music maskers in order to preserve the ecological validity of the study. The trade-off of doing so is that the stimuli are less controlled than lab-synthesized sounds, and the two songs used here have inherently different acoustic qualities. Previous studies show that features such as tempo and loudness affect concurrent reading comprehension [4], and the stimuli used in this study were matched in tempo (beats per minute) and intensity (RMS). Furthermore, both songs were of the same genre and featured a female vocalist. Importantly, the behavioral and neural effects found here differ as a function of listeners' musicality. Such differential effects across listeners suggest that acoustic differences cannot be the sole driver of our neural effects and consequently must be rooted in psychological rather than physical attributes of the stimulus.

5. Conclusions

In summary, our combined behavioral and neuroimaging results demonstrate that speech tracking is negatively affected both by familiarity and the presence of vocals in concurrent music. Furthermore, we show that these effects are modulated by musical ability, whereby less-musical listeners are more susceptible to these different background music characteristics. By using naturalistic, continuous stimuli, we simulated a realistic listening scenario, thus further adding to our understanding of the cocktail party phenomenon. Our findings also qualify prior studies by suggesting that in addition to general arousal, familiarity and internal structure of music (e.g., presence or absence of vocals) might affect concurrent cognitive-linguistic processing.

Author Contributions: Conceptualization, J.A.B. and G.M.B.; methodology, J.A.B. and G.M.B.; software, G.M.B.; validation, J.A.B.; formal analysis, J.A.B.; investigation, J.A.B.; resources, G.M.B.; data curation, J.A.B.; writing—original draft preparation, J.A.B.; writing—review and editing, J.A.B. and G.M.B.; visualization, J.A.B.; supervision, G.M.B.; project administration, J.A.B. and G.M.B.; funding acquisition, G.M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Institute on Deafness and Other Communication Disorders, grant number R01DC016267.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of The University of Memphis (ID 2370, approved 15 April 2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data are available upon request from the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kämpfe, J.; Sedlmeier, P.; Renkewitz, F. The impact of background music on adult listeners: A meta-analysis. *Psychol. Music* **2011**, *39*, 424–448. [\[CrossRef\]](#)
2. Crawford, H.J.; Strapp, C.M. Effects of vocal and instrumental music on visuospatial and verbal performance as moderated by studying preference and personality. *Personal. Individ. Differ.* **1994**, *16*, 237–245. [\[CrossRef\]](#)
3. Angel, L.A.; Polzella, D.J.; Elvers, G.C. Background music and cognitive performance. *Percept. Mot. Ski.* **2010**, *110*, 1059–1064. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Thompson, W.F.; Schellenberg, E.G.; Letnic, A.K. Fast and loud background music disrupts reading comprehension. *Psychol. Music* **2011**, *40*, 700–708. [\[CrossRef\]](#)
5. Brown, J.A.; Bidelman, G.M. Song properties and familiarity affect speech recognition in musical noise. *Psychomusicol. Music Mind Brain* **2022**, *32*, 1–6. [\[CrossRef\]](#)
6. Darrow, A.A.; Johnson, C.; Agnew, S.; Fuller, E.R.; Uchisaka, M. Effect of preferred music as a distraction on music majors' and nonmusic majors' selective attention. *Bull. Counc. Res. Music Educ.* **2006**, *170*, 21–31. [\[CrossRef\]](#)
7. Martin, R.C.; Wogalter, M.S.; Forlano, J.G. Reading comprehension in the presence of unattended speech and music. *J. Mem. Lang.* **1988**, *27*, 382–398. [\[CrossRef\]](#)
8. Perham, N.; Currie, H. Does listening to preferred music improve reading comprehension performance? *Appl. Cogn. Psychol.* **2014**, *28*, 279–284. [\[CrossRef\]](#)
9. Lee, E.K.; Lee, S.E.; Kwon, Y.S. The effect of lyrical and non-lyrical background music on different types of language processing—An ERP study. *Korean J. Cogn. Sci.* **2020**, *31*, 155–178. [\[CrossRef\]](#)
10. Vasilev, M.R.; Kirkby, J.A.; Angele, B. Auditory Distraction During Reading: A Bayesian Meta-Analysis of a Continuing Controversy. *Perspect. Psychol. Sci.* **2018**, *13*, 567–597. [\[CrossRef\]](#)
11. Souza, P.; Gehani, N.; Wright, R.; McCloy, D. The advantage of knowing the talker. *J. Am. Acad. Audiol.* **2013**, *24*, 689–700. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Yonan, C.A.; Sommers, M.S. The Effects of Talker Familiarity on Spoken Word Identification in Younger and Older Listeners. *Psychol. Aging* **2000**, *15*, 88–99. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Johnsrude, I.S.; Mackey, A.; Hakyemez, H.; Alexander, E.; Trang, H.P.; Carlyon, R.P. Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychol. Sci.* **2013**, *24*, 1995–2004. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Pisoni, D.B. Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Commun.* **1993**, *13*, 109–125. [\[CrossRef\]](#)
15. Feng, S.; Bidelman, G.M. Music listening and song familiarity modulate mind wandering and behavioral success during lexical processing. In Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci 2015), Pasadena, CA, USA, 22–25 July 2015.
16. Russo, F.A.; Pichora-Fuller, M.K. Tune in or tune out: Age-related differences in listening to speech in music. *Ear Hear.* **2008**, *29*, 746–760. [\[CrossRef\]](#)
17. De Groot, A.M.B.; Smedinga, H.E. Let the music play!: A short-term but no long-term detrimental effect of vocal background music with familiar language lyrics on foreign language vocabulary learning. *Stud. Second Lang. Acquis.* **2014**, *36*, 681–707. [\[CrossRef\]](#)
18. Bidelman, G.M.; Dexter, L. Bilinguals at the “cocktail party”: Dissociable neural activity in auditory-linguistic brain regions reveals neurobiological basis for nonnative listeners' speech-in-noise recognition deficits. *Brain Lang.* **2015**, *143*, 32–41. [\[CrossRef\]](#)
19. Strait, D.L.; Kraus, N. Can you hear me now? Musical training shapes functional brain networks for selective auditory attention and hearing speech in noise. *Front. Psychol.* **2011**, *2*, 113. [\[CrossRef\]](#)
20. Bidelman, G.M.; Krishnan, A. Effects of reverberation on brainstem representation of speech in musicians and non-musicians. *Brain Res.* **2010**, *1355*, 112–125. [\[CrossRef\]](#)
21. Bidelman, G.M.; Yoo, J. Musicians Show Improved Speech Segregation in Competitive, Multi-Talker Cocktail Party Scenarios. *Front. Psychol.* **2020**, *11*, 1927. [\[CrossRef\]](#)
22. Coffey, E.B.J.; Mogilever, N.B.; Zatorre, R.J. Speech-in-noise perception in musicians: A review. *Hear. Res.* **2017**, *352*, 49–69. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Parbery-Clark, A.; Skoe, E.; Kraus, N. Musical experience limits the degradative effects of background noise on the neural processing of sound. *J. Neurosci.* **2009**, *29*, 14100–14107. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Yoo, J.; Bidelman, G.M. Linguistic, perceptual, and cognitive factors underlying musicians' benefits in noise-degraded speech perception. *Hear. Res.* **2019**, *377*, 189–195. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Zendel, B.R.; Alain, C. Concurrent sound segregation is enhanced in musicians. *J. Cogn. Neurosci.* **2009**, *21*, 1488–1498. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Oxenham, A.J.; Fligor, B.J.; Mason, C.R.; Kidd, G., Jr. Informational masking and musical training. *J. Acoust. Soc. Am.* **2003**, *114*, 1543–1549. [\[CrossRef\]](#)
27. Parbery-Clark, A.; Skoe, E.; Lam, C.; Kraus, N. Musician enhancement for speech-in-noise. *Ear Hear.* **2009**, *30*, 653–661. [\[CrossRef\]](#)
28. Patston, L.; Tippett, L. The Effect of Background Music on Cognitive Performance in Musicians and Nonmusicians. *Music Backgr. Music Cogn.* **2011**, *29*, 173–184. [\[CrossRef\]](#)

29. Mankel, K.; Bidelman, G.M. Inherent auditory skills rather than formal music training shape the neural encoding of speech. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 13129–13134. [[CrossRef](#)]
30. Ekström, S.R.; Borg, E. Hearing speech in music. *Noise Health* **2011**, *13*, 277–285. [[CrossRef](#)]
31. Eskridge, E.; Galvin, J., III; Aronoff, J.; Li, T.; Fu, Q.J. Speech Perception with Music MASKers by Cochlear Implant Users and Normal Hearing Listeners. *J. Speech Lang. Hear. Res.* **2012**, *55*, 800–810. [[CrossRef](#)]
32. Crosse, M.J.; Di Liberto, G.M.; Bednar, A.; Lalor, E.C. The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci.* **2016**, *10*, 604. [[CrossRef](#)]
33. Forte, A.E.; Etard, O.; Reichenbach, T. The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *eLife* **2017**, *6*, e27203. [[CrossRef](#)] [[PubMed](#)]
34. Lalor, E.C.; Power, A.J.; Reilly, R.B.; Foxe, J.J. Resolving precise temporal processing properties of the auditory system using continuous stimuli. *J. Neurophysiol.* **2009**, *102*, 349–359. [[CrossRef](#)] [[PubMed](#)]
35. Ding, N.; Chatterjee, M.; Simon, J.Z. Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *Neuroimage* **2014**, *88*, 41–46. [[CrossRef](#)] [[PubMed](#)]
36. Oldfield, R. The Assessment and Analysis of Handedness: The Edinburgh Inventory. *Neuropsychologia* **1971**, *9*, 97–113. [[CrossRef](#)]
37. Ding, N.; Simon, J.Z. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J. Neurophysiol.* **2012**, *107*, 78–89. [[CrossRef](#)]
38. Law, L.N.; Zentner, M. Assessing musical abilities objectively: Construction and validation of the profile of music perception skills. *PLoS ONE* **2012**, *7*, e52508. [[CrossRef](#)]
39. Oostenveld, R.; Praamstra, P. The five percent electrode system for high-resolution EEG and ERP measurements. *Clin. Neurophysiol.* **2001**, *112*, 713–719. [[CrossRef](#)]
40. Crosse, M.J.; Zuk, N.J.; Liberto, G.M.D.; Nidiffer, A.R.; Molholm, S.; Lalor, E.C. Linear Modeling of Neurophysiological Responses to Naturalistic Stimuli: Methodological Considerations for Applied Research. *Front. Neurosci.* **2021**, *15*, 705621. [[CrossRef](#)]
41. Kulasingham, J.P.; Simon, J.Z. Algorithms for Estimating Time-Locked Neural Response Components in Cortical Processing of Continuous Speech. *IEEE Trans. Biomed. Eng.* **2022**, 1–9. [[CrossRef](#)]
42. Hillyard, S.A.; Hink, R.F.; Schwent, V.L.; Picton, T.W. Electrical signs of selective attention in the human brain. *Science* **1973**, *182*, 177–180. [[CrossRef](#)] [[PubMed](#)]
43. Näätänen, R.; Picton, T. The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiology* **1987**, *24*, 375–425. [[CrossRef](#)] [[PubMed](#)]
44. Picton, T.W.; Woods, D.L.; Proulx, G.B. Human auditory sustained potentials. I. The nature of the response. *Electroencephalogr. Clin. Neurophysiol.* **1978**, *45*, 186–197. [[CrossRef](#)]
45. Muncke, J.; Kuruvila, I.; Hoppe, U. Prediction of Speech Intelligibility by Means of EEG Responses to Sentences in Noise. *Front. Neurosci.* **2022**, *16*, 876421. [[CrossRef](#)] [[PubMed](#)]
46. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **2015**, *67*, 1–48. [[CrossRef](#)]
47. Mankel, K.; Barber, J.; Bidelman, G.M. Auditory categorical processing for speech is modulated by inherent musical listening skills. *NeuroReport* **2020**, *31*, 162–166. [[CrossRef](#)]
48. MacCallum, R.C.; Zhang, S.; Preacher, K.J.; Rucker, D. On the practice of dichotomization of quantitative variables. *Psychol. Methods* **2002**, *7*, 19–40. [[CrossRef](#)]
49. Brouwer, S.; Akkermans, N.; Hendriks, L.; van Uden, H.; Wilms, V. “Lass frooby noo!” the interference of song lyrics and meaning on speech intelligibility. *J. Exp. Psychol. Appl.* **2022**, *28*, 576–588. [[CrossRef](#)]
50. Kidd, G.; Colburn, H.S. Informational Masking in Speech Recognition. In *The Auditory System at the Cocktail Party*; Middlebrooks, J.C., Simon, J.Z., Popper, A.N., Fay, R.R., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 75–109.
51. Scharenborg, O.; Larson, M. Investigating the Effect of Music and Lyrics on Spoken-Word Recognition. *arXiv* **2018**, arXiv:1803.05058.
52. Lavie, N.; Hirst, A.; De Fockert, J.W.; Viding, E. Load theory of selective attention and cognitive control. *J. Exp. Psychol. Gen.* **2004**, *133*, 339–354. [[CrossRef](#)]
53. Dong, Y.; Zheng, H.-Y.; Wu, S.X.-Y.; Huang, F.-Y.; Peng, S.-N.; Sun, S.Y.-K.; Zeng, H. The effect of Chinese pop background music on Chinese poetry reading comprehension. *Psychol. Music* **2022**, *50*, 1–22. [[CrossRef](#)]
54. Bidelman, G.M.; Howell, M. Functional changes in inter- and intra-hemispheric auditory cortical processing underlying degraded speech perception. *Neuroimage* **2016**, *124*, 581–590. [[CrossRef](#)] [[PubMed](#)]
55. Key, A.P.F.; Dove, G.O.; Maguire, M.J. Linking brainwaves to the brain: An ERP primer. *Dev. Neuropsychol.* **2005**, *27*, 183–215. [[CrossRef](#)] [[PubMed](#)]
56. Ding, N.; Simon, J.Z. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 11854–11859. [[CrossRef](#)] [[PubMed](#)]
57. Strauss, D.J.; Corona-Strauss, F.I.; Trenado, C.; Bernarding, C.; Reith, W.; Latzel, M.; Froehlich, M. Electrophysiological correlates of listening effort: Neurodynamical modeling and measurement. *Cogn. Neurodyn.* **2010**, *4*, 119–131. [[CrossRef](#)]
58. Bidelman, G.M.; Lowther, J.E.; Tak, S.H.; Alain, C. Mild cognitive impairment is characterized by deficient hierarchical speech coding between auditory brainstem and cortex. *J. Neurosci.* **2017**, *37*, 3610–3620. [[CrossRef](#)]

59. Hennessy, S.; Mack, W.J.; Habibi, A. Speech-in-noise perception in musicians and non-musicians: A multi-level meta-analysis. *Hear. Res.* **2022**, *416*, 108442. [[CrossRef](#)]
60. Du, M.; Jiang, J.; Li, Z.; Man, D.; Jiang, C. The effects of background music on neural responses during reading comprehension. *Sci. Rep.* **2020**, *10*, 18651. [[CrossRef](#)]
61. Román-Caballero, R.; Martín-Arévalo, E.; Lupiáñez, J. Attentional networks functioning and vigilance in expert musicians and non-musicians. *Psychol. Res.* **2021**, *85*, 1121–1135. [[CrossRef](#)]
62. Bialystok, E.; DePape, A.M. Musical expertise, bilingualism, and executive functioning. *J. Exp. Psychol. Hum. Percept. Perform.* **2009**, *35*, 565–574. [[CrossRef](#)]
63. Weiss, M.W.; Trehub, S.E.; Schellenberg, E.G.; Habashi, P. Pupils Dilate for Vocal or Familiar Music. *J. Exp. Psychol. Hum. Percept. Perform.* **2016**, *42*, 1061–1065. [[CrossRef](#)] [[PubMed](#)]
64. Swaminathan, J.; Mason, C.R.; Streeter, T.M.; Best, V.; Kidd, G., Jr.; Patel, A.D. Musical training, individual differences and the cocktail party problem. *Sci. Rep.* **2015**, *5*, 11628. [[CrossRef](#)] [[PubMed](#)]
65. Bendixen, A. Predictability effects in auditory scene analysis: A review. *Front. Neurosci.* **2014**, *8*, 60. [[CrossRef](#)] [[PubMed](#)]