



ORIGINAL ARTICLE

Musicians at the Cocktail Party: Neural Substrates of Musical Training During Selective Listening in Multispeaker Situations

Sebastian Puschmann ^{1,2}, Sylvain Baillet^{1,2} and Robert J. Zatorre^{1,2,3}

¹Montreal Neurological Institute, McGill University, Montreal, Quebec, H3A 2B4, Canada, ²Centre for Research on Brain, Language and Music, Montreal, Quebec, H3G 2A8, Canada and ³International Laboratory for Brain, Music and Sound Research, Montreal, Quebec, H2V 2J2, Canada

Address correspondence to Sebastian Puschmann, Montreal Neurological Institute, Cognitive Neuroscience Unit, 3801 Rue University, Montreal, Quebec, H3A 2B4, Canada. Email: sebastian.puschmann@mail.mcgill.ca  orcid.org/0000-0002-8552-5353

Abstract

Musical training has been demonstrated to benefit speech-in-noise perception. It is however unknown whether this effect translates to selective listening in cocktail party situations, and if so what its neural basis might be. We investigated this question using magnetoencephalography-based speech envelope reconstruction and a sustained selective listening task, in which participants with varying amounts of musical training attended to 1 of 2 speech streams while detecting rare target words. Cortical frequency-following responses (FFR) and auditory working memory were additionally measured to dissociate musical training-related effects on low-level auditory processing versus higher cognitive function. Results show that the duration of musical training is associated with a reduced distracting effect of competing speech on target detection accuracy. Remarkably, more musical training was related to a robust neural tracking of both the to-be-attended and the to-be-ignored speech stream, up until late cortical processing stages. Musical training-related increases in FFR power were associated with a robust speech tracking in auditory sensory areas, whereas training-related differences in auditory working memory were linked to an increased representation of the to-be-ignored stream beyond auditory cortex. Our findings suggest that musically trained persons can use additional information about the distracting stream to limit interference by competing speech.

Key words: auditory cognition, MEG, selective attention, speech

Introduction

Mastering a musical instrument places high demands on auditory, motor, and cognitive skills, making musicians an ideal model of training-induced brain plasticity (Herholz and Zatorre 2012; Strait and Kraus 2014). Patel (2011) suggests that—due to a partial overlap of underlying brain networks—musical training may also benefit speech processing in challenging listening situations. In line with this concept, many studies have reported superior performance in musicians when listening to

speech in noise in various target and noise background configurations (for a review, see Coffey, Mogilever et al. 2017).

The musician advantage in processing speech in noise can be partially attributed to superior low-level processing of auditory information. Previous work demonstrates increased frequency-following responses (FFR) from the brainstem and auditory cortices, shorter response latencies, and increased response fidelity in musicians (Bidelman et al. 2014; Strait et al. 2014; Coffey et al. 2016). Moreover, musicians demonstrate

more robust neural representations of stimulus inputs on both brainstem and cortical levels when presented with speech in noise, suggesting that auditory sensory processing is more resilient to interference (Parbery-Clark, Skoe and Kraus 2009; Du and Zatorre 2017).

Another line of research suggests that effects of musical training on speech-in-noise performance may also be mediated by training-induced increases in auditory working memory (Kraus et al. 2012). Across a wide range of tasks, musicians have been repeatedly demonstrated to show superior auditory working memory performance (Parbery-Clark, Skoe, Lam et al. 2009; Cohen et al. 2011). In the context of speech-in-noise processing, working memory is thought to foster the use of contextual semantic or phonological cues, thus facilitating the formation of input predictions and the filling of gaps in the perceived speech stream (Rönnberg et al. 2013; Zekveld et al. 2013). Subjects with high working memory capacity show reduced effort when listening to speech in noise, as well as reduced susceptibility to distractor sounds (Dalton et al. 2009; Rudner et al. 2012; Tsuchida et al. 2012). Sörqvist et al. (2010) reported top-down effects of working memory on early processing of auditory information at the brainstem level, suggesting a possible link between superior auditory working memory and more robust encoding of speech stimuli, especially in musicians.

A limitation in the literature is that effects of musical training on speech-in-noise understanding have been primarily studied using short target stimuli and energetic masker backgrounds (Coffey, Mogilever et al. 2017). It is still an open question whether and to what extent the musician advantage in these tasks generalizes to more complex listening settings with informational masking (Boebinger et al. 2015; Swaminathan et al. 2015; Baskent and Gaudrain 2016).

Here we combined behavioral measures and magnetoencephalography (MEG) recordings to investigate how musical training modulates listening success and brain processing

during sustained selective listening to continuous speech in a “cocktail party” scenario (Cherry 1953). We used a stimulus decoding technique to reconstruct the speech envelope from MEG source time series and to quantify the cortical tracking of selectively attended and ignored speech streams (Ding and Simon 2012a; O’Sullivan et al. 2015). Auditory working memory scores and FFRs, serving as a marker of low-level coding fidelity, were obtained to investigate differential contributions of these factors on speech tracking and selective listening success.

Previous work provides evidence that the auditory cortex tracks the slow temporal fluctuations in speech streams (Ding and Simon 2012a, 2012b; Kubanek et al. 2013). During selective listening, the ongoing auditory cortex response is dominated by the attended speech stream, whereas the masker background is relatively suppressed (Teder et al. 1993; Ding and Simon 2012a, 2012b; Mesgarani and Chang 2012; Puvvada and Simon 2017). Hierarchically higher brain regions involved in speech processing primarily track the attended speech stream (Zion Golumbic et al. 2013). Several studies have shown a positive association between the accuracy of tracking the attended speech stream and listening performance (Ding and Simon 2013; O’Sullivan et al. 2015; Puschmann et al. 2017). We therefore reasoned that if musical training promotes cocktail party listening, it may do so by amplifying attentional modulations in speech processing, leading to an increased cortical representation of the to-be-attended speech stream and/or a stronger suppression of the to-be-ignored speech. An alternative hypothesis however would be that musical training may be associated with an enhanced representation of both to-be-attended and to-be-ignored speech streams. We also hypothesized that in musically trained subjects, superior low-level coding fidelity, as measured via the FFR, would be associated with a more accurate early cortical tracking of speech in auditory cortex, whereas musical training-related increases in auditory working memory would be related to the persistence of speech representations at later cortical processing stages.

Table 1. Musical training

Age at onset (years)	Total training (h)	Intense training (years)
–	0	0
–	0	0
–	0	0
–	0	0
14	400	1
10	300	2
11	11 900	3
11	1200	4
6	2600	9
8	3100	10
7	3500	10
6	7400	10
7	4800	11
7	8300	11
6	9200	14
5	12 900	14
9	5100	16
5	12 600	17
6	7000	18
7	30 400	18

The table states the individual age at the onset of musical training, the reported number of practice hours throughout the lifetime (rounded to hundreds), and the time period of intense musical training (i.e., number of years in which subjects practiced at least 3 times per week for at least 1 h). All measures were assessed using the Montreal Musical History questionnaire.

Material and Methods

Participants

A total of 20 right-handed volunteers (11 females; mean age: 21 ± 3 years; age range: 19–27 years) participated in the experiment. Participants with different degrees of musical training were recruited from an existing database and were selected to have participated in previous MRI studies, allowing us to use existing structural T1 images for an anatomical coregistration of the obtained MEG data. All participants were native speakers of English and had no history of neurological, psychiatric, or hearing-related disorder. The experimental procedures were approved by the Research Ethics Board of the Montreal Neurological Institute and Hospital and written informed consent was obtained from all participants.

The Montreal Musical History questionnaire (Coffey et al. 2011) was used to assess the onset age of musical training, the duration of musical training, and the total hours of training received during the life span (Table 1). All statistical analyses on the effect of musical training on task performance and brain function were based on the duration of intense musical training, which was defined as the number of years in which participants practiced at least 3 times per week for at least 1 h. In musically trained individuals, this measure was significantly correlated with the total number of training hours (Pearson’s

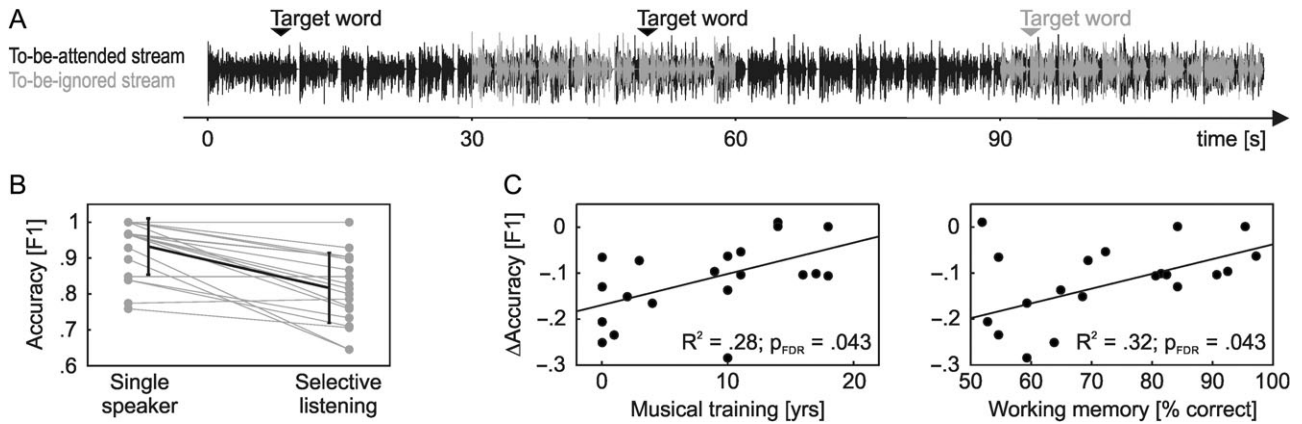


Figure 1. Selective listening task and behavior. (A) Participants performed a sustained selective listening task in which they attended to 1 of 2 simultaneously presented audio streams. Selective-listening trials alternated with single-speaker trials, in which only the to-be-attended speech stream was present. Listening performance was measured using a target word detection task. Target words occurring in the to-be-ignored speech stream (gray) had to be ignored. (B) The figure depicts the individual target detection accuracy (gray) as well as group averages (black; \pm standard deviation) in single speaker and selective-listening trials of the experiment. (C) The duration of musical training and auditory working memory scores were associated with individual differences in the relative enhancement of target detection accuracy during selective listening.

$r = 0.58, P = 0.019$) and with the onset age at training start ($r = -0.78, P < 0.001$).

Stimulus and Task

Audio recordings of 2 detective stories by Sir Arthur Conan Doyle, “The yellow face” and “The Greek interpreter,” served as the to-be-attended and the to-be-ignored speech streams, respectively. Both stories were read by nonidentical male speakers. The speech streams were adjusted for root mean square intensity and cut into consecutive 30-s segments. In each 30-s segment, the speech streams were tagged with a sinusoidal amplitude modulation of 95 Hz (to-be-ignored stream) or 105 Hz (to-be-attended stream) modulation rate and a 50% modulation depth. Both speech streams were presented diotically, thus appearing to originate from a central sound source. Stimulus level was set to 70 dB(A).

The selective listening task is depicted in Figure 1A. Selective-listening trials, in which both speech streams interfered with each other, alternated with single-speaker trials, in which only the to-be-attended speech stream was present. The trial duration was set to 30 s for both conditions. During each trial, a target word was presented visually on a projection screen. The target word appeared 5 s after trial onset and was presented for 20 s. For the remaining time, a fixation cross was displayed to stabilize eye gaze.

Participants were instructed to press a button on a response pad whenever the target word occurred in the to-be-attended speech stream. Listeners were instructed not to respond for target words occurring in the to-be-ignored stream. The target word was changed after each task trial. In the selective-listening condition, the target word occurred within the to-be-attended and the to-be-ignored stream with equal probability. In the single-speaker condition, the target word was absent in half of the trials.

The task was split into 3 runs, each consisting of 10 selective-listening and 10 single-speaker trials. Runs started with a single-speaker trial to facilitate the initial identification of the to-be-attended speech stream. The to-be-attended stream and the to-be-detected target words were identical across participants. After finishing the experiment, participants

were asked to freely recall the content of the to-be-attended story.

Additional Behavioral Testing

Prior to MEG recording, several other tests were carried out. A standard pure tone audiogram (frequency range: 0.125–8 kHz) was obtained from all participants. The pure tone average (PTA; i.e., the mean hearing threshold obtained at 500, 1000, 2000, and 4000 Hz, averaged over both ears) served as measure of individual hearing loss.

The Hearing in Noise test (HINT; Nilsson et al. 1994) was used to assess the 50% speech reception threshold for sentences in noise (SRT_{Noise}). We applied an adaptive 1-up-1-down staircase procedure in which the presentation level of the speech stimuli was varied adaptively in steps of 3, 1, or 0.5 dB. The step size was decreased after the second and fourth turning point. The noise level was set to 65 dB(A). The signal-to-noise ratio (SNR) in the first trial was set to 6 dB. The staircase consisted of at least 23 trials and 10 turns. A training run was used to familiarize subjects with the task. The average SNR obtained across all trials following the fourth turning point served as measure of SRT_{Noise} .

Complementing the HINT measures, the speech reception threshold for sentences in a concurrent speech background (SRT_{Speech}) was obtained using a task design adapted from Swaminathan et al. (2015). In each trial, participants listened to 1 of 3 simultaneously presented 5-word sentences, uttered by 3 nonidentical male speakers. All sentences were taken from a closed set matrix corpus developed by Kidd et al. (2008) and had an identical grammatical structure (i.e., <name><verb><number><color><noun>). In each trial, 3 speakers were chosen randomly from a pool of 11 male voices. Participants were instructed to repeat the sentence beginning with the word “Mike.” The target sentence was presented at the fixed level of 55 dB(A). SNR in the first trial was set to 15 dB. The level of the competing sentences was adjusted parametrically in steps of 4 or 1 dB (after 2 turns) in a 1-up-1-down staircase procedure. The staircase consisted of at least 23 trials and 10 turns. SRT_{Speech} was calculated from the average SNR of all trials following the fourth turning point. A training run was used to familiarize participants with the task.

Performance in a tonal sequence manipulation task (Foster et al. 2013; Albouy et al. 2017) was used as an index of auditory working memory performance. In each trial of this task, a 3-tone piano sequence was presented. After a 2-s retention period, the sequence was repeated in reversed temporal order. Participants had to judge whether a single tone was changed in the second presentation (in positions 1–3, pitch change of 2 or 3 semi tones equally distributed). The pitch change preserved the melodic contour of the sequence. The task consisted of 108 trials, half of them containing a pitch change. The presentation order was controlled in such a way that similar trial types (same, different) could not be presented more than 3 times in a row. The auditory working memory score was computed as the percentage of correct judgements given across all trials.

Behavioral Data Analysis

Target word detection accuracy in single-speaker and selective-listening trials of the selective listening task was quantified using the F1 score (Van Rijnsbergen 1979). The F1 score represents the harmonic mean of recall (i.e., number of detected target words divided by the total number of targets words) and precision (number of detected target words divided by the total number of button presses). Group-level differences in target word detection accuracy between single-speaker and selective-listening conditions were assessed using a paired *t*-test.

To account for variability in baseline target detection performance between participants (Fig. 1B), the difference in F1 scores obtained in the single-speaker and selective-listening trials ($\Delta F1$) served as measure of individual selective listening performance in our study. We tested for linear relationships between $\Delta F1$ and audiological measures, the duration of musical training, and auditory working memory scores using robust regression with default settings as implemented in MATLAB (i.e., bisquare weighting, tuning constant set to 4.685). Robust regression was preferred to ordinary least-squares regression to reduce the effect of outliers on the parameter estimation. The goodness of the robust regression fit is stated in terms of R^2 values. Regression slopes are considered to differ significantly from zero when passing a statistical threshold of $P_{FDR} < 0.05$, corrected for multiple comparisons using the Benjamini-Hochberg method for false discovery rate correction (5 tests included; Benjamini and Hochberg 1995).

Robust regression was used to assess potential relationships between musical training and both auditory working memory scores and audiological measures. A significance criterion of $P_{FDR} < 0.05$ was applied (FDR correction over 4 tests). Given that previous studies suggest an increase of auditory working memory performance with increasing levels of musical training, we tested for positive relationships only. A Sobel test (Sobel 1982; MacKinnon et al. 1995) was performed to test whether auditory working memory acts as a mediator to convey effects of musical training on selective listening performance.

MEG Data Acquisition

MEG data were acquired during the selective listening task using a 275-channel whole-head MEG system (CTF/VMS, Port Coquitlam, British Columbia, Canada). Data were recorded with a sampling rate of 2400 Hz, an antialiasing filter with a 600 Hz cut-off, and third-order spatial gradient noise cancellation. Horizontal and vertical electrooculograms, and an electrocardiogram were acquired with bipolar montages. The head position inside the MEG sensor helmet was determined with coils

fixated at the nasion and the preauricular points (fiducial points). For anatomical registration with anatomical MRI data, the spatial positions of the fiducial coils and of about 150 scalp points were obtained using a 3D digitizer system (Polhemus Isotrack, Polhemus, Colchester, VT, USA). Participants were seated in upright position in a sound-attenuated, magnetically shielded recording room. Auditory stimulation was delivered via insert earphones (E-A-RTONE 3A, 3M, St. Paul, MN, USA). The earphones were equipped with customized prolonged air tubes (1.5 m) to increase the distance between audio transducers and MEG sensors, thus minimizing potential stimulation artifacts in the MEG signal, which was confirmed with pilot testing using a foam head.

MEG Data Preprocessing

MEG data preprocessing was performed in Brainstorm (Tadel et al. 2011), in observance of good-practice guidelines (Gross et al. 2013). Line noise artifacts were removed using notch filters at the power frequency (60 Hz) and its first 3 harmonics. Artifacts related to eye movements and cardiac activity were pruned from the data using independent components analysis. For this procedure, a copy of the data was offline filtered between 1 and 40 Hz. A principal component analysis was performed to reduce the dimensionality of the MEG data to 40 dimensions and 40 independent components were computed using the extended infomax algorithm implemented in Brainstorm. The demixing matrix obtained from this procedure was applied to the original unfiltered MEG dataset and independent components which sensor topography and/or time series reflected eye blinks, lateral eye movements, and cardiac activity were removed. No further data cleaning was performed.

The participant's structural T1 MRI images were automatically segmented and labeled using Freesurfer (Dale et al. 1999; Fischl et al. 1999, 2004), and coregistered to the MEG data in Brainstorm using the digitized head points. An OPENMEEG boundary element method head model and a surface-based minimum norm source model with depth weighting (order: 0.5; maximal amount: 10) was computed for each participant, also using Brainstorm. Dipole orientation was constrained to be normal to the cortical surface. Noise covariance matrices for the source reconstruction process were estimated from 2-min empty room recordings obtained for each participant.

FFR Analysis

The ongoing speech signal was tagged with an amplitude modulation. Previous work shows that the ongoing auditory response locks to such periodic signal amplitude fluctuations, resulting in spectral peaks at the modulation frequency (Picton et al. 2003; Bharadwaj et al. 2014). Individual differences in spectral power at the amplitude modulation rate have been related to behavioral measures of auditory temporal acuity (Purcell et al. 2004; Bharadwaj et al. 2015), suggesting that the FFR power can serve as a neural marker of auditory temporal coding fidelity. Please note that this subtype of the FFR is also known as envelope following response (Shinn-Cunningham et al. 2017).

MEG data demonstrate that FFRs can be measured not only at brainstem and midbrain levels, but also from auditory cortex (Schoonhoven et al. 2003; Coffey et al. 2016). We here obtained cortical FFRs from left and right Heschl's gyrus. A copy of the cleaned MEG sensor data was bandpass-filtered between 60 and 120 Hz, downsampled to 600 Hz sampling rate, and

epoched into continuous 2-s intervals. Each of the resulting epochs contained 190 or 210 full cycles (phase onset: 0) of the sinusoidal 95 and 105 Hz amplitude modulations applied to the speech stimuli. Averaging across epochs and conditions is thought to preserve the entrained oscillation at the modulation frequency while attenuating stimulus-independent intrinsic oscillations. MEG sources of the condition averages (single-speaker/selective-listening trials) were projected into the standard MNI space using Brainstorm. Welch's power spectral density was estimated for all elementary sources within left and right Heschl's gyrus regions and subsequently averaged within hemispheres. Anatomical labels were obtained automatically from Freesurfer, based on the Desikan-Killiany atlas (Desikan et al. 2006). To account for overall power differences across individuals, the spectral power at the tagged modulation frequencies is reported relative to the individual baseline power (i.e., mean spectral power between 60 and 120 Hz).

A 2×3 repeated-measures ANOVA tested for differences in FFR power related to hemisphere (2 levels: left/right) and listening condition (3 levels: single/to-be-attended/to-be-ignored stream). Post hoc paired t-tests were used to investigate the statistically significant main effect of listening condition in more detail. Results of the post hoc t-tests were reported as statistically significant when passing a threshold of $P_{FDR} < 0.05$ (FDR correction over 3 tests). Pearson correlation was used to investigate the stability of individual differences in FFR power across hemispheres ($P_{FDR} < 0.05$, FDR correction over 3 tests).

Since previous work suggests that selective attention may modulate FFR amplitudes (Lehmann and Schönwiesner 2014), FFR power measured during the single-speaker phase, which lacks any attention manipulation, served as unbiased measure of auditory coding fidelity. FFR power was averaged over both hemispheres for further analyses. Linear relationships between individual FFR power and the duration of musical training or task performance were analyzed using robust regression. Based on our hypotheses, we only tested for positive relationships. Results are reported as statistically significant for $P_{FDR} < 0.05$ (FDR correction over 2 tests).

Speech Envelope Reconstruction

Speech envelope fluctuations are typically most pronounced in the delta and theta frequency range, with spectral peaks occurring between 4 and 7 Hz (Houtgast and Steeneken 1985; Giraud and Poeppel 2012). Following up on previous studies applying speech envelope reconstruction methods, we therefore used only low frequency components of the MEG response for the speech envelope reconstruction (Ding and Simon 2012a, 2012b; Puschmann et al. 2017; Puvvada and Simon 2017). MEG sensor data were bandpass-filtered from 1 to 8 Hz, epoched from 0 to 30 s relative to the onset of each single-speaker and selective-listening trial, and downsampled to 64 Hz to reduce computational demands. MEG sources for each listening condition were projected into the standard MNI space and mean signal time courses were extracted from 102 regions-of-interest (ROIs), covering the entire cortical surface (Fig. 3A). The ROIs were based on anatomical labels provided by the Desikan-Killiany atlas. Large anatomical regions were subdivided into cohesive portions of about 25 mm² surface area. ROI time courses were z-transformed to equalize means and standard deviations across regions and trials.

Amplitude envelopes of the speech streams were obtained using a Hilbert transform, followed by 1–8 Hz bandpass filtering. Filtering was performed using a third order Butterworth filter

and the `filtfilt` function in MATLAB for zero-phase digital filtering of the data. The speech streams were subsequently cut into the corresponding 30-s intervals, downsampled to 64 Hz, and z-transformed.

The reconstruction of the speech envelope time course from the concurrently measured cortical MEG data was performed using the multivariate temporal response function toolbox for MATLAB (Crosse et al. 2016). A ridge regression was used for a linear backward mapping between the 102 MEG source time courses and envelope fluctuations of the speech stimuli. Sets of 102 regression weights were computed for each trial and for all time lags between 0 and 500 ms following the speech onset. The ridge parameter λ was optimized for each stream and each time lag using a search grid and a leave-one-out cross-validation with the goal to minimize the mean squared error (grid values: $\lambda = 10^{-2}, 10^{-1}, \dots, 10^8$). Regression models were estimated separately for each stream.

The speech envelope reconstruction was performed using a leave-one-out cross-validation procedure on the subject level (Mirkovic et al. 2016; Puschmann et al. 2017). This means that the speech envelope reconstruction in a selected trial and for a given stream was based on the mean regression weights obtained for this stream in all other but this trial. The leave-one-out cross-validation procedure ensured that the reconstruction did not depend on trial-specific properties of the recorded MEG data but was rather related to a general and trial-independent mapping between sound envelope and MEG response. Pearson's correlation between the reconstructed and the original speech envelope were computed to quantify the accuracy of the envelope reconstruction. The obtained r values were converted to normally distributed r_z values using Fisher's z transformation and averaged across trials.

Analysis of Overall Mean Reconstruction Accuracy

For the to-be-attended and to-be-ignored speech stream, r_z curves were averaged over all time lags from 0 to 500 ms and analyzed as a function of musical training, auditory working memory scores, and FFR power using robust regression. Results were treated as statistically significant when passing a threshold of $P_{FDR} < 0.05$ (FDR correction over 6 tests). To test whether the tracking of the to-be-ignored stream increases relative to the tracking of the to-be-attended stream with an increasing duration of musical training, auditory working memory, or FFR power, the ratio of speech envelope reconstruction accuracies obtained in both conditions $r_z(\text{Ign})/r_z(\text{Att})$ was computed and analyzed as a function of the 3 covariates using Spearman's correlation (one-tailed test; $P_{FDR} < 0.05$; FDR correction over 3 tests).

Time-Window-of-Interest Analysis

We analyzed relationships between the 3 covariates of interest and r_z scores obtained for the to-be-ignored speech stream within different time windows of interest, which were selected based on the r_z peaks in the single-speaker condition (early: 15–80 ms; intermediate: 90–175 ms; late: 250–450 ms). The r_z values were averaged within each time window of interest. Again, robust regression was applied to study relationships to musical training, auditory working memory, and FFR power ($P_{FDR} < 0.05$; FDR correction over 9 tests).

Spatial Pattern of Speech Envelope Tracking

To reveal which cortical regions showed robust speech envelope tracking in the different listening conditions, we computed envelope reconstruction accuracies for individual ROIs and

over all time lags from 0 to 500 ms. The regressor matrix consisted of multiple copies of a single ROI time series which were shifted in time, covering the entire time window of interest. This approach therefore resulted in a single r_z value for each ROI and the entire time window. The pattern of brain areas showing significant speech envelope tracking in the single-speaker condition served as anatomical mask for a spatial analysis of speech tracking. For thresholding, we computed P values corresponding to the mean r values obtained in each region. Regions were reported to show robust speech tracking for $P_{FDR} < 0.05$ (one-tailed; FDR correction over 102 regions).

ROI Analysis

The speech envelope reconstruction pipeline was re-computed for each ROI and the lag range from 0 to 500 ms. For each ROI within the anatomical mask of interest, reconstruction accuracies were averaged within early (15–80 ms), intermediate (90–175 ms), and late (250–450 ms) time windows of interest. Robust regression was used to investigate the relationship between r_z scores and musical training, auditory working memory scores, or FFR power within the different time windows. Based on the results of the preceding analysis, we only tested for positive relationships to the covariates. Results were then reported as statistically significant when passing a threshold of $P_{FDR} < 0.05$ (FDR correction over 59 tests).

Results

Behavioral Data

Figure 1B depicts individual (gray) and mean (black) target word detection accuracies during the selective listening experiment, for both single-speaker and selective-listening trials. Overall, there was a significant decrease in target detection accuracy when the to-be-ignored speech stream superimposed the to-be-attended stream ($t[19] = 6.27$, $P < .001$). On the individual level, we however observed a high variability in the extent to which the additional to-be-ignored speech input affected task performance, with $\Delta F1$ ranging from -0.28 to 0.01 . Robust regression showed that both the duration of musical training ($R^2 = 0.28$, slope = 0.007 year^{-1} , $P_{FDR} = 0.043$) and auditory working memory capacity ($R^2 = 0.32$, slope = $0.003 \text{ \% correct}^{-1}$, $P_{FDR} = 0.043$) were positively related to individual $\Delta F1$ values. As shown in Figure 1C, both a longer duration of musical training as well as higher auditory working memory scores were associated with smaller changes in target detection performance between the single-speaker and selective-listening conditions. In contrast, none of the audiological measures explained the variability in target word detection accuracy (PTA: $R^2 = 0.10$, $P_{FDR} = 0.284$; SRT_{Noise} : $R^2 = 0.02$, $P_{FDR} = 0.763$; SRT_{Speech} : $R^2 < 0.01$, $P_{FDR} = 0.813$). As in previous studies on the effects of musical training, the auditory working memory score of our participants was positively associated with the duration of their musical training ($R^2 = 0.32$, slope = $1.4 \text{ \% correct/year}$, $P_{FDR} = 0.009$, one-tailed). There was no relationship between musical training and any of the audiological measures (PTA: $R^2 = 0.08$, $P_{FDR} = 0.474$; SRT_{Noise} : $R^2 = 0.01$, $P_{FDR} = 0.642$; SRT_{Speech} : $R^2 = 0.01$, $P_{FDR} = 0.642$).

Given the relationship between musical training and auditory working memory scores, we hypothesized that auditory working memory may act as a mediator variable to convey the effect of musical training on task performance. A Sobel test however revealed no significant mediation effect ($t[19] = 1.03$, $P = 0.156$). This result suggests that effects of musical training

on task performance cannot be solely attributed to differences in auditory working memory induced by musical training.

Cortical FFR Analysis

Figure 2A depicts the power spectrum density estimates obtained from left and right Heschl's gyrus sources in the single-speaker condition and during selective listening. The power spectra show bilateral peaks at the amplitude of the modulation frequencies applied to the to-be-attended (105 Hz) and to-be-ignored (95 Hz) speech streams. Figure 2B illustrates the mean power spectral density at 105 Hz for each MEG sensor in the single-speaker condition. The power topography is consistent with the pattern of MEG activity generated by an auditory cortex source (Coffey et al. 2016), providing evidence that the measured FFR response is primarily generated in cortical auditory sensory areas.

A repeated-measures ANOVA was used to test for effects of hemisphere (left/right Heschl's gyrus) and condition (single stream at 105 Hz/to-be-attended stream at 105 Hz/to-be-ignored stream at 95 Hz) on FFR power. The analysis revealed a main effect of listening condition ($F[2,38] = 5.14$, $P = 0.011$) as well as a significant main effect of hemisphere ($F[1,19] = 5.32$, $P = 0.032$), but no condition-by-hemisphere interaction ($F[2,38] = 1.63$, $P = 0.210$). As shown in Figure 2C, FFR power was reduced for the to-be-attended stream ($4.6 \pm 4.5 \text{ dB}$) as compared with both the single speaker ($6.1 \pm 5.4 \text{ dB}$; $P_{FDR} = 0.028$) and to-be-ignored speech stream ($5.6 \pm 5.0 \text{ dB}$; $P_{FDR} = 0.048$). No differences in FFR power were observed between single-speaker and to-be-ignored streams ($P_{FDR} = 0.426$). In line with data by Coffey et al. (2016), FFR power was overall higher in right ($5.8 \pm 5.1 \text{ dB}$) than in left auditory cortex ($4.6 \pm 4.5 \text{ dB}$). In all listening conditions, individual FFR responses were however highly correlated across hemispheres (single stream: $r = 0.98$, $P_{FDR} < 0.001$; to-be-attended stream: $r = 0.65$, $P_{FDR} = 0.002$; to-be-ignored stream: $r = 0.73$, $P_{FDR} < 0.001$).

The FFR power at 105 Hz as obtained in the single-stream condition, averaged over both hemispheres, served as a measure of temporal coding fidelity in auditory cortex. A robust regression analysis revealed that the individual FFR power was positively related to the duration of musical training ($R^2 = 0.21$, slope = 0.40 dB/year , $P_{FDR} = 0.041$, one-tailed; see Fig. 2D), as predicted by previous studies. In contrast to musical training and working memory scores, FFR power was not significantly related to $\Delta F1$ scores in the selective listening task ($R^2 = 0.11$, $P_{FDR} = 0.074$, one-tailed). In an exploratory analysis, we subsequently confirmed that a similar relationship between FFR power and musical training can qualitatively also be observed for the to-be-attended speech stream ($R^2 = 0.15$, slope = 0.28 dB/year , $P = 0.047$, uncorrected, one-tailed). No such effect was found for the to-be-ignored speech stream ($R^2 = 0.08$, $P = 0.109$, uncorrected, one-tailed).

Speech Envelope Reconstruction

The analyses presented above provide evidence that individual differences in musical training can explain differences in selective listening performance. As predicted by prior research, musical training was further found to modulate auditory working memory scores as well as FFR power measured in auditory cortex. Our results below clarify how these variables affect the cortical tracking of the to-be-attended and to-be-ignored speech streams during selective listening. Figure 3B depicts the accuracy of speech envelope tracking r_z across conditions, for a

range of relative time lags (from 0 to 500 ms) between speech input and cortical signals. Overall, speech tracking was strongest in the single-speaker condition, as expected. In selective-listening trials, higher speech envelope reconstruction accuracies were achieved for the to-be-attended than for the to-be-ignored stream, demonstrating that the cortex predominantly tracks the to-be-attended speech stream.

Analysis of Overall Mean Reconstruction Accuracy

Speech tracking accuracies, averaged over the tested time-lag range from 0 to 500 ms, were analyzed as a function of musical training, auditory working memory scores, and FFR power. Figure 3C shows that the r_z scores for the to-be-ignored speech stream were positively related to all covariates (musical training: $R^2 = 0.48$, slope = 0.002 year^{-1} , $P_{\text{FDR}} = 0.004$; working memory: $R^2 = 0.29$, slope = $0.001 \text{ \% correct}^{-1}$, $P_{\text{FDR}} = 0.043$; FFR power: $R^2 = 0.26$, slope = 0.002 dB^{-1} , $P_{\text{FDR}} = 0.043$). No statistically significant relationship was found for the to-be-attended speech stream. To test whether the representation strength of the to-be-ignored stream increases relative to the to-be-attended speech stream with musical training, working memory scores, or FFR power, we calculated the ratio between r_z values in both conditions and performed Spearman correlations with all covariates. As shown in Figure 3D, significant effects were found for the duration of musical training ($P = 0.48$, $P_{\text{FDR}} = 0.023$, one-tailed) and working memory ($P = 0.51$, $P_{\text{FDR}} = 0.023$, one-tailed), but not for FFR power ($P = 0.33$, $P_{\text{FDR}} = 0.080$, one-tailed).

In an additional exploratory analysis, we tested for a direct positive association between envelope reconstruction accuracies and our behavioral measure of target word detection accuracy. For the to-be-attended speech stream, robust regression provides some evidence for the expected positive relationship between r_z scores and $\Delta F1$ ($R^2 = 0.17$, slope = 1.7 , $P = 0.035$, uncorrected, one-tailed). No such effect was found for the to-be-ignored stream ($R^2 = 0.07$, $P = 0.130$, uncorrected, one-tailed).

Spatial Pattern of Speech Envelope Tracking

Figure 3E depicts the brain regions with robust speech envelope tracking for all tested conditions (single-speaker, to-be-attended, and to-be-ignored streams). Speech envelope reconstruction accuracies were computed individually for each stream, and over the entire tested lag range between speech input and cortical signals (0 to 500 ms). In the single-speaker condition, we observed robust speech envelope tracking in auditory sensory regions and adjacent temporal areas, the inferior parietal lobe, the motor and somatosensory cortices, and large portions of prefrontal cortex. Speech tracking accuracy was generally reduced for the to-be-attended speech stream during selective listening, albeit still involving auditory sensory, motor, and ventrolateral prefrontal cortex. This was in clear contrast with the regions tracking the to-be-ignored stream, which were restricted to early auditory areas (at $P_{\text{FDR}} < 0.05$, one-tailed).

Time-Window-of-Interest Analysis

Complementing the analyses over the entire tested lag range, we also tested for effects restricted to early, intermediate, and late stages of information processing. We defined 3 time windows of interest based on the peaks in r_z scores obtained in the single-speaker condition (early: 15–80 ms; intermediate: 90–175 ms; late: 250–450 ms). The mean reconstruction accuracies were derived in all 3 time windows. Since, for the entire lag

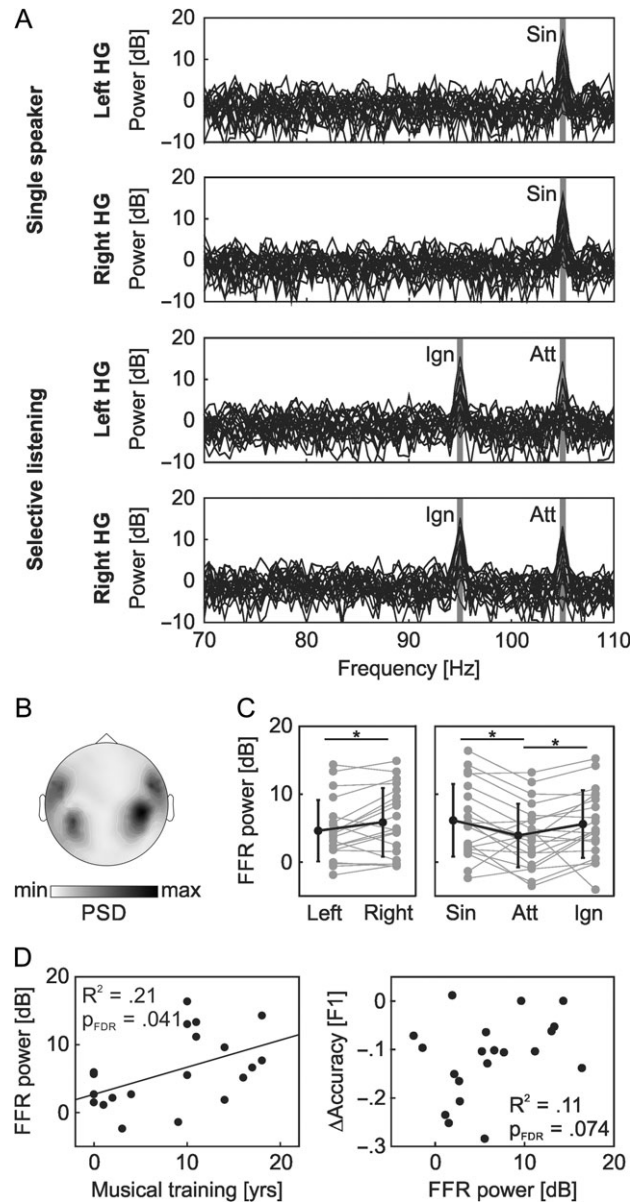


Figure 2. FFR analysis: The single/to-be-attended (Sin/Att) and the to-be-ignored (Ign) speech stream were tagged with an amplitude modulation of 105 or 95 Hz modulation frequency. (A) The spectral power induced at the modulation frequencies was investigated in surface ROIs covering left and right Heschl's gyrus. The figure depicts the individual power spectrum for each participant, with response peaks emerging at both modulation frequencies. (B) The group-level sensor topography of the power spectral density, obtained in the single-speaker condition at 105 Hz (i.e., the modulation frequency of the speech stream), is consistent with the pattern of MEG activity typically generated by an auditory cortex source. (C) FFR power varied significantly as between hemispheres (left/right Heschl's gyrus) and as a function of listening condition (Sin = single-stream at 105 Hz, Att = to-be-attended stream at 105 Hz, Ign = to-be-ignored stream at 95 Hz). The figure shows individual data (gray) and group averages (black; \pm standard deviation). Significant differences between conditions are marked by asterisks. (D) The FFR power obtained in the single stream condition, averaged over both hemispheres, served as a measure of temporal coding fidelity in auditory cortex. This measure was significantly related to the duration of musical training, but did not significantly explain individual differences in task performance.

range from 0 to 500 ms, musical training, auditory working memory, and FFR power were only related to the tracking of the to-be-ignored stream, analyses were restricted to this stream. Figure 4A shows that effects of musical training on

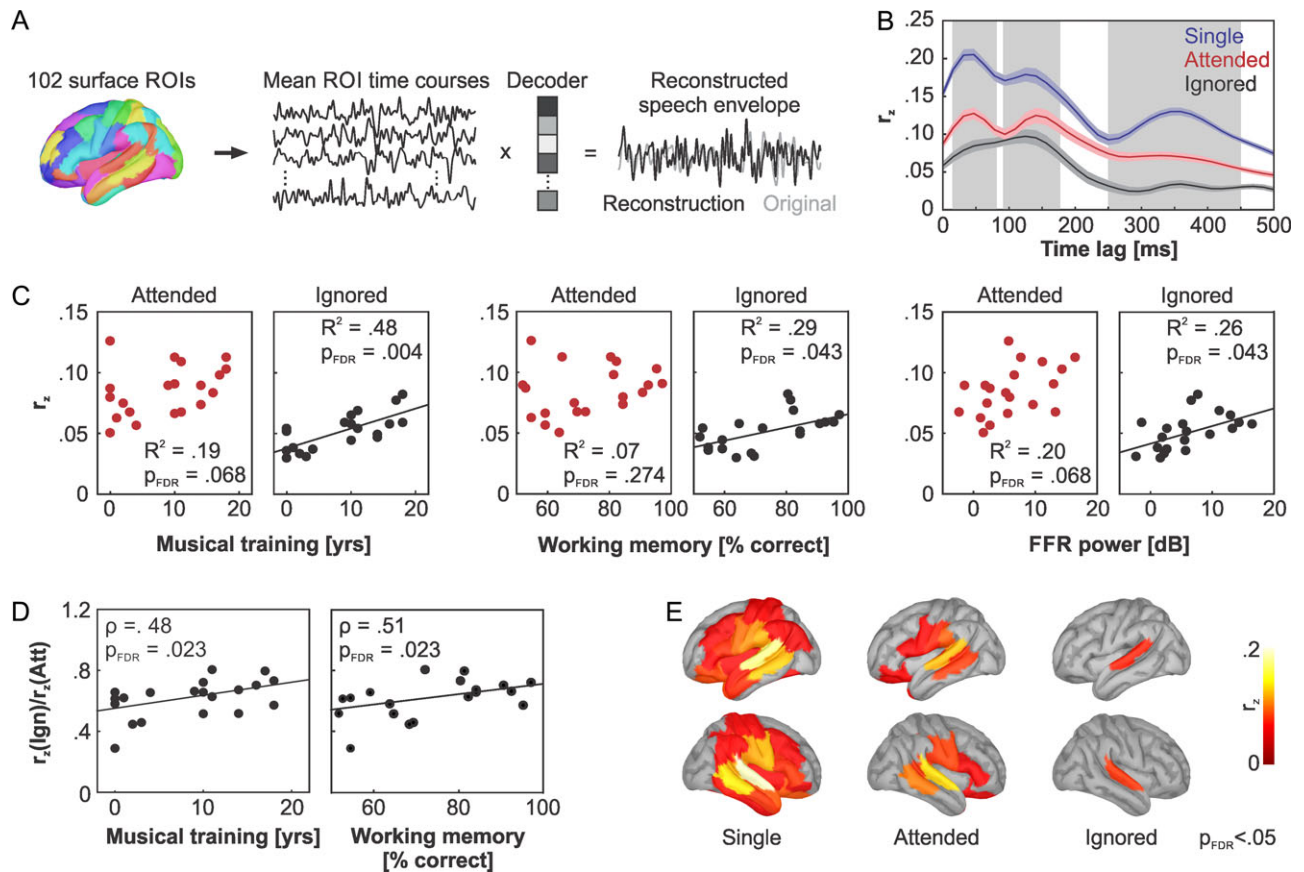


Figure 3. Speech envelope reconstruction: (A) Mean MEG time courses were extracted from 102 surface ROIs. Backward modeling of the speech envelope from the MEG source time courses was performed using a linear decoder and a leave-one-out cross validation. Fisher z-transformed Pearson's correlation coefficients between the original and the reconstructed speech envelope served as measure of reconstruction accuracy. (B) Mean speech envelope reconstruction accuracy (\pm standard error of the mean) as a function of stream and relative time lag between sound input and MEG response. Gray boxes highlight early (15–80 ms), intermediate (90–175 ms), and late (250–450 ms) time windows of interest. The time windows were chosen based on peaks in envelope reconstruction accuracy obtained for the single-speaker condition. (C) Relationship between the individual mean speech envelope reconstruction accuracy, averaged from 0 to 500 ms, and musical training, auditory working memory scores, and FFR power. Black lines show statistically significant robust regression slopes obtained for the to-be-ignored speech stream. (D) Ratio between individual estimates of speech reconstruction accuracy for the to-be-ignored and to-be-attended speech stream as a function of musical training and auditory working memory. The relative representation strength of the to-be-ignored stream increased with an increasing duration of musical training and higher working memory scores. (E) ROIs showing a robust tracking of the speech envelope in the single-speaker condition as well as for the to-be-attended and the to-be-ignored stream during selective listening ($P_{\text{FDR}} < 0.05$, one-tailed).

speech envelope reconstruction accuracy for the to-be-ignored speech stream were persistent over time, across all tested time windows (early: $R^2 = 0.28$, slope = 0.002 year^{-1} , $P_{\text{FDR}} = 0.036$; intermediate: $R^2 = 0.34$, slope = 0.003 year^{-1} , $P_{\text{FDR}} = 0.020$; late: $R^2 = 0.43$, slope = 0.001 year^{-1} , $P_{\text{FDR}} = 0.008$). Auditory working memory was related to tracking of the to-be-ignored stream in the intermediate, but not in the early and late time windows of interest (early: $R^2 = 0.06$, $P_{\text{FDR}} = 0.324$; intermediate: $R^2 = 0.50$, slope = $0.001 \text{ \% correct}^{-1}$, $P_{\text{FDR}} = 0.005$; late: $R^2 = 0.11$, $P_{\text{FDR}} = 0.190$). There was no significant relation between FFR power and speech envelope reconstruction accuracy in any of the 3 time windows, although some tendency was detected in the earliest window (early: $R^2 = 0.23$, $P_{\text{FDR}} = 0.055$; intermediate: $R^2 = 0.18$; $P_{\text{FDR}} = 0.091$; late: $R^2 = 0.04$, $P_{\text{FDR}} = 0.404$).

ROI Analysis

We mapped the brain regions in which tracking accuracy of the to-be-ignored speech stream was related to the duration of musical training, auditory working memory performance, and temporal coding fidelity in auditory cortex, as indexed by measures of FFR power. For each ROI, the mean speech envelope

reconstruction accuracy within the 3 time windows of interest (15–80, 90–175, and 250–450 ms) was computed. The analysis was restricted to brain regions showing a robust speech envelope tracking in the single-speaker condition (Fig. 3E, left column). We concentrated on those time windows in which a relationship between speech tracking accuracy and 1 of the 3 covariates of interest was found.

We observed significant effects of musical training on the tracking of the to-be-ignored speech stream within individual ROIs only in the early (15–80 ms) and late (250–450 ms) time windows (at $P_{\text{FDR}} < 0.05$, one-tailed; Fig. 4B). In both time windows, speech envelope reconstruction accuracies in left posterior temporal sulcus increased with the duration of musical training. Similar trends were observed in adjacent auditory sensory and somatosensory regions as well as—only for the early time window—in the precuneus, inferior parietal regions, and left inferior frontal cortex ($P < 0.05$, uncorrected, one-tailed). Note that, in the intermediate time window, effects of musical training tended to shift from auditory sensory areas towards hierarchically higher cortical regions ($P < 0.05$, uncorrected, one-tailed).

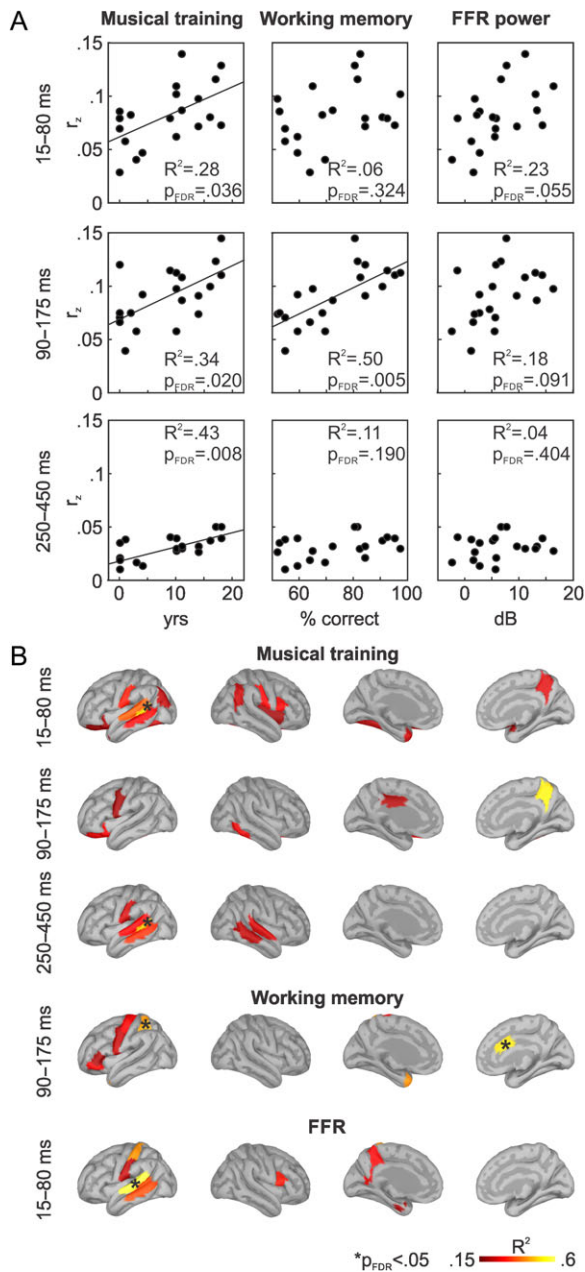


Figure 4. The tracking of the to-be-ignored speech stream was investigated within 3 time windows of interest, which were defined based on peaks in speech envelope reconstruction accuracy for the single-speaker condition (shown in Fig. 3B). (A) Scatter plots show mean r_z values obtained within the time windows of interest as a function of musical training, auditory working memory scores, and FFR power. Black lines depict statistically significant robust regression slopes. (B) ROIs showing a positive relationship between mean speech tracking accuracy and the corresponding covariate within the respective time window are highlighted on the brain surface (at $P < 0.05$, uncorrected, one-tailed). Asterisks mark ROIs in which a statistically significant relationship can be observed (i.e., $P_{FDR} < 0.05$, one-tailed).

Effects of auditory working memory performance on the tracking of the to-be-ignored stream in the intermediate lag range (90–175 ms) were found over the left superior parietal lobe and the anterior cingulate gyrus ($P_{FDR} < 0.05$, one-tailed). Similarly, speech envelope reconstruction accuracies in left motor cortex and parts of left inferior frontal gyrus, encompassing Broca's area, tended to increase with increasing

working memory performance ($P < 0.05$, uncorrected, one-tailed). Speech envelope tracking in auditory sensory brain areas was not related to auditory working memory.

Although effects of FFR power on cortical speech tracking were not significant in any of the 3 time windows of interest, we analyzed which brain regions contributed most in the first time window, in which some tendency was detected. Individual differences in FFR power were predominantly associated with envelope reconstruction accuracies in the left auditory cortex, in particular in the posterior portion of the superior temporal lobe ($P_{FDR} < 0.05$, one-tailed). Similar positive trends were found in left somatosensory cortex, the precuneus, and right inferior frontal lobe ($P < 0.05$, uncorrected, one-tailed). No effects of FFR power were observed in right auditory cortex.

Discussion

Our data provide evidence that musical training is related to both improved behavioral performance and enhanced cortical speech tracking during selective listening in a cocktail party situation. Individuals with longer duration of intense musical training showed a reduced distracting effect of competing speech on target word detection accuracy. Surprisingly, this enhancement was associated with an increased cortical speech envelope reconstruction accuracy of the to-be-ignored speech stream, indicating a more balanced cortical tracking of attended and ignored speech in musically trained persons.

Musical Training is Associated With Better Selective Listening Performance

Overall, our participants performed worse in the presence of competing speech, as expected. Yet, the deleterious effect of the to-be-ignored speech stream on task performance decreased with increasing duration of musical training (Fig. 1C). This result suggests a higher resilience against auditory distractors in musically trained individuals during cocktail party listening. While multiple studies reported a musician advantage when listening to speech in noise (for a review, see Coffey, Mogilever et al. 2017), previous work on selective listening in cocktail party settings provided no consistent evidence for beneficial effects of musical training (Boebinger et al. 2015; Swaminathan et al. 2015; Baskent and Gaudrain 2016; Clayton et al. 2016). Most of these studies however tested for musical training-related effects on speech reception thresholds in speech backgrounds, whereas our experiment assessed performance during sustained selective listening at suprathreshold level. It is noteworthy that when applying the experimental design introduced by Swaminathan et al. (2015) and no spatial separation of sound sources, we observed no effect of musical training on speech reception thresholds in a speech background. This observation suggests that our present findings of a beneficial effect of musical training in cocktail party listening cannot be explained by changes in speech reception thresholds in speech backgrounds.

Similarly, our data provide no additional evidence for reduced speech-in-noise perception thresholds in musicians. Unlike the majority of studies reporting such effects, we however did not compare groups of trained musicians and nonmusicians (Coffey, Mogilever et al. 2017) but treated musical training as a continuous covariate in a heterogeneous group of participants, including both musicians and nonmusicians.

Effect of Musical Training on Cortical Speech Tracking

During selective listening, speech envelope reconstruction of the to-be-ignored speech from MEG source time courses was degraded with respect to the attended speech stream (Fig. 3B). This result is in accord with previous work applying similar reconstruction approaches, and in line with the view that the attended speech stream is more strongly represented in the ongoing neural response than ignored speech (Ding and Simon 2012a, 2012b; Mesgarani and Chang 2012; Puschmann et al. 2017). Recent work suggests that attention-related modulations on cortical representations of speech occur at the level of auditory cortex, but only after initial sensory processing of the acoustic scene (Puvvada and Simon 2017).

In our study, robust cortical tracking of the to-be-attended speech stream was observed in auditory sensory regions, motor and somatosensory cortex, and inferior frontal brain regions (Fig. 3E). In contrast, tracking of the to-be-ignored speech stream was—on the group level—restricted to auditory cortex, in spite of possible field spread of MEG source imaging (Brodbeck et al. 2018). These observations are in line with electrocorticographic recordings reported by Zion Golumbic et al. (2013).

Previous work on cocktail party listening provide evidence for a link between selective listening performance and the cortical tracking of attended speech (O'Sullivan et al. 2015; Puschmann et al. 2017). The intelligibility of speech in noise has also been related to the accuracy of cortical speech envelope tracking (Ding and Simon 2013). Further, behavioral benefits of musicians in speech in noise perception were reported to be associated with a superior encoding of stimulus information, yielding more robust speech representations at both brainstem and cortical levels, including auditory sensory as well as premotor and frontal areas (Parbery-Clark, Skoe and Kraus 2009; Du and Zatorre 2017). Based on these findings, we expected to observe a positive relationship between the duration of musical training and the tracking of the to-be-attended speech stream. However, although our data provide some evidence for the previously reported relationship between the tracking of attended speech and behavior, effects of musical training on the envelope reconstruction accuracies obtained for the to-be-attended stream were relatively weak in our dataset. Instead, we observed a strong positive relationship between the individual duration of musical training and the tracking of the to-be-ignored speech stream (Fig. 3C). Further, a longer duration of intense musical training was associated with a more balanced cortical representation of both speech streams (Fig. 3D).

While positive effects of musical training on the tracking of the ignored speech were most robust over the left auditory cortex, we observed similar tendencies in motor and somatosensory cortices, the inferior parietal lobe, and inferior frontal brain regions (Fig. 4B). Increased tracking of ignored speech was persistent up until late processing time windows, suggesting that the increased resilience against distractor interference in musically trained subjects did not rely on an efficient filtering of the attended information (Fritz et al. 2007). Our findings are consistent with and extend recent fMRI results by Du and Zatorre (2017) on speech-in-noise processing in musicians and nonmusicians. Their data showed that musicians exhibit a more robust multivariate encoding of speech features in the auditory cortex and adjacent temporal regions in high signal-to-noise conditions, but also in motor and somatosensory cortices, and in parts of the inferior frontal lobe, especially as signal-to-noise conditions become more difficult. We here demonstrate that, in musically trained individuals, the same brain regions also reveal

more robust temporal envelope representations of the to-be-ignored speech during cocktail party listening.

One possible explanation for the ability of musically trained persons to uphold representations of both attended and ignored sound streams is that musical training and performance typically require not only segregating one sound from a background, but also attending to and integrating multiple different musical streams (Disbergen et al. 2018). However, it is unknown how increased tracking of the to-be-ignored speech stream may benefit listening success on the attended channel. On the one hand, keeping track of both the attentional foreground and background may facilitate predictions of how both streams evolve over time and, thus, stabilize auditory stream segregation during cocktail party listening (Elhilali and Shamma 2008). In consequence, intrusions of the irrelevant speaker may be limited, thus reducing the overall listening effort. On the other hand, it may enable predictions on the temporal progression of the to-be-ignored stream, allowing better anticipation of time intervals of low acoustic energy in the noise background. However, although it is widely believed that listeners tend to “listen in the dips” of the noise background to improve speech perception, previous work does not provide strong evidence for effects of masker predictability on listening success (Cooke 2006; Jones and Litovsky 2008).

Relationship to Low-Level Sensory Processing and Auditory Working Memory

Musical training was previously demonstrated to benefit both low-level sensory processing of auditory information and higher cognitive functions, in particular auditory working memory (Kraus et al. 2012; Strait and Kraus 2014). In line with these findings, we also observed positive relationships between the duration of musical training and both auditory working memory performance and cortical FFR, which can be seen as a marker of temporal coding fidelity of low-level auditory processing. Differences in temporal coding fidelity and in auditory working memory however affected the cortical tracking of speech in different ways.

Individual differences in FFR power were positively related to the cortical tracking of the to-be-ignored stream. In contrast to effects found for musical training and working memory scores, there was no significant change in the relative cortical representation of attended and ignored speech with increasing FFR power. This suggests that musical training-related changes in FFR power improve the cortical tracking of both streams similarly. The association between FFR power and speech envelope tracking was strongest at early time points (i.e., the 15–80 ms time window) and in left auditory sensory brain regions, indicating that superior temporal coding of auditory information enabled more robust tracking at early stages of auditory sensory processing (Fig. 4B). This effect may be related to a higher fidelity of stimulus representations within the auditory system *per se* (Strait et al. 2014), but also to a superior stream segregation, resulting in more stable representations of both speech streams. In our task, participants presumably relied on spectral cues for early input-based stream segregation. Previous work provides evidence for a link between the enhanced temporal coding of auditory information in musicians and fine pitch discrimination abilities (Bidelman et al. 2011; Coffey et al. 2016; Coffey, Chepesiuk et al. 2017). Also, musicians were reported to show superior pitch-based concurrent stream segregation (Zendel and Alain 2008).

Similar to the duration of musical training, auditory working memory was a good predictor of behavioral performance (Fig. 1C). On the neural level, auditory working memory performance was positively related to the cortical tracking of the to-be-ignored, but not of the to-be-attended, speech stream, similar to the effect of musical training. Working memory-related modulations of speech tracking were strongest at intermediate time lags (i.e., the 90–175 ms time window), and primarily detected in left superior parietal and anterior cingulate regions, with similar trends in the left motor cortex and left inferior frontal lobe (Fig. 4A,B). This observation suggests that auditory working memory, in contrast to the fidelity of temporal information coding, does not amplify the early cortical representation of speech in auditory sensory areas, but rather strengthens the representation of the to-be-ignored speech input at later processing stages.

Auditory working memory was assessed using a tonal sequence manipulation task (Foster et al. 2013; Albouy et al. 2017), measuring a nonverbal tonal component of auditory working memory (Schulze and Koelsch 2012). Still, the tonal auditory working memory measure was associated with both the individual outcome of the verbal behavioral task and cortical speech tracking. Other studies however provided evidence for a similar musical training-related increase in verbal auditory working memory (Parbery-Clark, Skoe and Kraus 2009; Strait et al. 2012), suggesting that musical training can benefit different components of auditory working memory. Tonal and verbal working memory networks were previously reported to share a set of common core structures (Schulze et al. 2011), but can also be selectively impaired (Albouy et al., 2013).

Recent MEG data demonstrated that the motor system is involved in generating temporal predictions of auditory inputs to facilitate auditory stream segregation (Morillon and Baillet 2017). In the context of speech processing, delta-band inputs from left inferior frontal cortex and left motor cortex into left auditory cortex were found to be associated with increased neural coupling with the speech envelope in auditory cortex (Park et al. 2015). Similarly, cortical oscillations in orbitofrontal regions were shown to modulate the entrainment of auditory cortex by the speech envelope (Keitel et al. 2017). Taken together, these findings provide evidence for top-down predictive signaling from frontal and motor cortices directed to the auditory cortex, leading to superior speech envelope tracking. The superior parietal regions were, in contrast, previously reported to receive bottom-up signals from auditory sensory areas, related to the neural tracking of the speech envelope (Keitel et al. 2017).

Our data show that subjects with superior auditory working memory performance maintain increased representations of the ignored speech stream during selective listening, and that this is associated with superior parietal lobe areas, the left motor cortex, and left inferior frontal regions. We suggest that the latter brain regions generate input predictions of both the attended and ignored speech streams during cocktail party listening, thus facilitating ongoing stream segregation and, potentially, listening “in the dips” of the perceptual background. Listeners with inferior auditory working memory performance may not have the capacity to keep and process representations of both input streams, and therefore may only rely on predictions of the to-be-attended stream. However, future work is necessary to test this hypothesis specifically and in greater detail. Also, it remains open whether prediction signals of attended and ignored speech input are generated independently or interdependently of each other.

Summary

In conclusion, we here demonstrate that the positive effect of musical training on speech-in-noise perception extends to selective listening in a cocktail party setting. Subjects with a long duration of intense musical training benefited from a superior low-level temporal coding fidelity in auditory cortex and increased auditory working memory, allowing them to keep robust neural representations of both the attended and the ignored speech streams, even at later stages of processing beyond auditory cortex. We suggest that these neural enhancements may allow musically trained individuals to generate predictions of both the attended and ignored speech input, thus facilitating ongoing auditory stream segregation. Our data challenge the view that successful selective listening relies solely on the efficient filtering of attended input. Instead, a more balanced processing of both perceptual foreground and background may be advantageous to reduce distractor interference and the overall listening effort in cocktail party situations.

Funding

Research scholarship of the Deutsche Forschungsgemeinschaft (DFG; PU590/1) to S.P., a Discovery grant from the Natural Sciences and Engineering Research Council of Canada (NSERC; 436355-13), an operating grant from the National Institutes of Health (NIH; 2R01EB009048-05) and a Platform Support Grant from the Brain Canada Foundation (PSG15-3755) to S.B., and a Foundation Grant from the Canadian Institutes of Health Research (CIHR; FDN143217) to R.J.Z.

Notes

The authors thank Philippe Albouy for providing the auditory working memory test, Elizabeth Bock for supporting the MEG data acquisition, Emily B.J. Coffey for her advice on the FFR analysis, and Gerald D. Kidd Jr. for granting access to the matrix-style speech corpus. *Conflict of Interest*: None declared.

References

- Albouy P, Schulze K, Caclin A, Tillmann B. 2013. Does tonality boost short-term memory in congenital amusia? *Brain Res.* 1537:224–232.
- Albouy P, Weiss A, Baillet S, Zatorre RJ. 2017. Selective entrainment of theta oscillations in the dorsal stream causally enhances auditory working memory performance. *Neuron.* 94:193–206.
- Baskent D, Gaudrain E. 2016. Musician advantage for speech-on-speech perception. *J Acoust Soc Am.* 139:EL51–EL56.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc Ser B.* 57:289–300.
- Bharadwaj HM, Lee AK, Shinn-Cunningham BG. 2014. Measuring auditory selective attention using frequency tagging. *Front Integr Neurosci.* 8:6.
- Bharadwaj HM, Masud S, Mehraei G, Verhulst S, Shinn-Cunningham BG. 2015. Individual differences reveal correlates of hidden hearing deficits. *J Neurosci.* 35:2161–2172.
- Bidelman GM, Krishnan A, Gandour JT. 2011. Enhanced brainstem encoding predicts musicians’ perceptual advantages with pitch. *Eur J Neurosci.* 33:530–538.
- Bidelman GM, Weiss MW, Moreno S, Alain C. 2014. Coordinated plasticity in brainstem and auditory cortex contributes to

- enhanced categorical speech perception in musicians. *Eur J Neurosci.* 40:2662–2673.
- Boebinger D, Evans S, Rosen S, Lima CF, Manly T, Scott SK. 2015. Musicians and non-musicians are equally adept at perceiving masked speech. *J Acoust Soc Am.* 137:378–387.
- Brodbeck C, Presacco A, Simon JZ. 2018. Neural source dynamics of brain responses to continuous stimuli: speech processing from acoustics to comprehension. *Neuroimage.* 172:162–174.
- Cherry EC. 1953. Some experiments on the recognition of speech, with one and with two ears. *J Acoust Soc Am.* 25: 975–979.
- Clayton KK, Swaminathan J, Yazdanbakhsh A, Zuk J, Patel AD, Kidd G Jr. 2016. Executive function, visual attention and the cocktail party problem in musicians and non-musicians. *PLoS One.* 11:e0157638.
- Coffey EB, Chepesiuk AM, Herholz SC, Baillet S, Zatorre RJ. 2017. Neural correlates of early sound encoding and their relationship to speech-in-noise perception. *Front Neurosci.* 11:479.
- Coffey EB, Herholz SC, Chepesiuk AM, Baillet S, Zatorre RJ. 2016. Cortical contributions to the auditory frequency-following response revealed by MEG. *Nat Commun.* 7:11070.
- Coffey EB, Herholz SC, Scala S, Zatorre RJ. 2011. Montreal Music History Questionnaire: a tool for the assessment of music-related experience in music cognition research. In: *The Neurosciences and Music IV: Learning and Memory, Conference.* Edinburgh, UK.
- Coffey EB, Mogilever NB, Zatorre RJ. 2017. Speech-in-noise perception in musicians: a review. *Hear Res.* 352:49–69.
- Cohen MA, Evans KK, Horowitz TS, Wolfe JM. 2011. Auditory and visual memory in musicians and nonmusicians. *Psychon Bull Rev.* 18:586–591.
- Cooke M. 2006. A glimpsing model of speech perception in noise. *J Acoust Soc Am.* 119:1562–1573.
- Crosse MJ, Di Liberto GM, Bednar A, Lalor EC. 2016. The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Front Hum Neurosci.* 10:604.
- Dale AM, Fischl B, Sereno MI. 1999. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage.* 9:179–194.
- Dalton P, Santangelo V, Spence C. 2009. The role of working memory in auditory selective attention. *Q J Exp Psychol (Hove).* 62:2126–2132.
- Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT. 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage.* 31:968–980.
- Ding N, Simon JZ. 2012a. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc Natl Acad Sci USA.* 109:11854–11859.
- Ding N, Simon JZ. 2012b. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J Neurophysiol.* 107:78–89.
- Ding N, Simon JZ. 2013. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *J Neurosci.* 33:5728–5735.
- Disbergen NR, Valente G, Formisano E, Zatorre RJ. 2018. Assessing top-down and bottom-up contributions to auditory stream segregation and integration with polyphonic music. *Front Neurosci.* 12:121.
- Du Y, Zatorre RJ. 2017. Musical training sharpens and bonds ears and tongue to hear speech better. *Proc Natl Acad Sci USA.* 114:13579–13584.
- Elhilali M, Shamma SA. 2008. A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. *J Acoust Soc Am.* 124:3751–3771.
- Fischl B, Sereno MI, Dale AM. 1999. Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. *Neuroimage.* 9:195–207.
- Fischl B, Van Der Kouwe A, Destrieux C, Halgren E, Ségonne F, Salat DH, Busa E, Seidman LJ, Goldstein J, Kennedy D. 2004. Automatically parcellating the human cerebral cortex. *Cereb Cortex.* 14:11–22.
- Foster NE, Halpern AR, Zatorre RJ. 2013. Common parietal activation in musical mental transformations across pitch and time. *Neuroimage.* 75:27–35.
- Fritz JB, Elhilali M, David SV, Shamma SA. 2007. Auditory attention-focusing the searchlight on sound. *Curr Opin Neurobiol.* 17:437–455.
- Giraud A-L, Poeppel D. 2012. Cortical oscillations and speech processing: emerging computational principles and operations. *Nat Neurosci.* 15:511.
- Gross J, Baillet S, Barnes GR, Henson RN, Hillebrand A, Jensen O, Jerbi K, Litvak V, Maess B, Oostenveld R, et al. 2013. Good practice for conducting and reporting MEG research. *Neuroimage.* 65:349–363.
- Herholz SC, Zatorre RJ. 2012. Musical training as a framework for brain plasticity: behavior, function, and structure. *Neuron.* 76:486–502.
- Houtgast T, Steeneken HJ. 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J Acoust Soc Am.* 77:1069–1077.
- Jones GL, Litovsky RY. 2008. Role of masker predictability in the cocktail party problem. *J Acoust Soc Am.* 124:3818–3830.
- Keitel A, Ince RAA, Gross J, Kayser C. 2017. Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks. *Neuroimage.* 147:32–42.
- Kidd G Jr., Best V, Mason CR. 2008. Listening to every other word: examining the strength of linkage variables in forming streams of speech. *J Acoust Soc Am.* 124: 3793–3802.
- Kraus N, Strait DL, Parbery-Clark A. 2012. Cognitive factors shape brain networks for auditory skills: spotlight on auditory working memory. *Ann N Y Acad Sci.* 1252:100–107.
- Kubaneck J, Brunner P, Gunduz A, Poeppel D, Schalk G. 2013. The tracking of speech envelope in the human cortex. *PLoS One.* 8:e53398.
- Lehmann A, Schönwiesner M. 2014. Selective attention modulates human auditory brainstem responses: relative contributions of frequency and spatial cues. *PLoS One.* 9:e85442.
- MacKinnon DP, Warsi G, Dwyer JH. 1995. A simulation study of mediated effect measures. *Multivariate Behav Res.* 30:41–62.
- Mesgarani N, Chang EF. 2012. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature.* 485:233–236.
- Mirkovic B, Bleichner MG, De Vos M, Debener S. 2016. Target speaker detection with concealed EEG around the ear. *Front Neurosci.* 10:349.
- Morillon B, Baillet S. 2017. Motor origin of temporal predictions in auditory attention. *Proc Natl Acad Sci USA.* 114: E8913–E8921.
- Nilsson M, Soli SD, Sullivan JA. 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.
- O’Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC. 2015.

- Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb Cortex*. 25:1697–1706.
- Parbery-Clark A, Skoe E, Kraus N. 2009. Musical experience limits the degradative effects of background noise on the neural processing of sound. *J Neurosci*. 29:14100–14107.
- Parbery-Clark A, Skoe E, Lam C, Kraus N. 2009. Musician enhancement for speech-in-noise. *Ear Hear*. 30:653–661.
- Park H, Ince RA, Schyns PG, Thut G, Gross J. 2015. Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Curr Biol*. 25:1649–1653.
- Patel AD. 2011. Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. *Front Psychol*. 2:142.
- Picton TW, John MS, Dimitrijevic A, Purcell D. 2003. Human auditory steady-state responses. *Int J Audiol*. 42:177–219.
- Purcell DW, John SM, Schneider BA, Picton TW. 2004. Human temporal auditory acuity as assessed by envelope following responses. *J Acoust Soc Am*. 116:3581–3593.
- Puschmann S, Steinkamp S, Gillich I, Mirkovic B, Debener S, Thiel CM. 2017. The right temporoparietal junction supports speech tracking during selective listening: evidence from concurrent EEG-fMRI. *J Neurosci*. 37:11505–11516.
- Puvvada KC, Simon JZ. 2017. Cortical representations of speech in a multitalker auditory scene. *J Neurosci*. 37:9189–9196.
- Rudner M, Lunner T, Behrens T, Thoren ES, Rönnerberg J. 2012. Working memory capacity may influence perceived effort during aided speech recognition in noise. *J Am Acad Audiol*. 23:577–589.
- Rönnerberg J, Lunner T, Zekveld A, Sörqvist P, Danielsson H, Lyxell B, Dahlstrom O, Signoret C, Stenfelt S, Pichora-Fuller MK, et al. 2013. The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Front Syst Neurosci*. 7:31.
- Schoonhoven R, Boden CJ, Verbunt JP, de Munck JC. 2003. A whole head MEG study of the amplitude-modulation-following response: phase coherence, group delay and dipole source analysis. *Clin Neurophysiol*. 114:2096–2106.
- Schulze K, Koelsch S. 2012. Working memory for speech and music. *Ann N Y Acad Sci*. 1252:229–236.
- Schulze K, Zysset S, Mueller K, Friederici AD, Koelsch S. 2011. Neuroarchitecture of verbal and tonal working memory in nonmusicians and musicians. *Hum Brain Mapp*. 32:771–783.
- Shinn-Cunningham BG, Varghese L, Wang L, Bharadwaj H. 2017. Individual differences in temporal perception and their implications for everyday listening. In: Kraus N, Anderson S, White-Schwoch T, Fay RR, Popper AN, editors. *The frequency-following response: a window into human communication*. New York (NY): Springer International Publishing. p. 159–192.
- Sobel ME. 1982. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol Methodol*. 13: 290–312.
- Strait DL, Kraus N. 2014. Biological impact of auditory expertise across the life span: musicians as a model of auditory learning. *Hear Res*. 308:109–121.
- Strait DL, O’Connell S, Parbery-Clark A, Kraus N. 2014. Musicians’ enhanced neural differentiation of speech sounds arises early in life: developmental evidence from ages 3 to 30. *Cereb Cortex*. 24:2512–2521.
- Strait DL, Parbery-Clark A, Hittner E, Kraus N. 2012. Musical training during early childhood enhances the neural encoding of speech in noise. *Brain Lang*. 123:191–201.
- Swaminathan J, Mason CR, Streeter TM, Best V, Kidd G Jr., Patel AD. 2015. Musical training, individual differences and the cocktail party problem. *Sci Rep*. 5:11628.
- Sörqvist P, Ljungberg JK, Ljung R. 2010. A sub-process view of working memory capacity: evidence from effects of speech on prose memory. *Memory*. 18:310–326.
- Tadel F, Baillet S, Mosher JC, Pantazis D, Leahy RM. 2011. Brainstorm: a user-friendly application for MEG/EEG analysis. *Comput Intell Neurosci*. 2011:879716.
- Teder W, Kujala T, Näätänen R. 1993. Selection of speech messages in free-field listening. *Neuroreport*. 5:307–309.
- Tsuchida Y, Katayama J, Murohashi H. 2012. Working memory capacity affects the interference control of distractors at auditory gating. *Neurosci Lett*. 516:62–66.
- Van Rijsbergen CJ. 1979. *Information retrieval*. Newton, MA: Butterworth.
- Zekveld A, Rudner M, Johnsrude IS, Rönnerberg J. 2013. The effects of working memory capacity and semantic cues on the intelligibility of speech in noise. *J Acoust Soc Am*. 134: 2225–2234.
- Zendel BR, Alain C. 2008. Concurrent sound segregation is enhanced in musicians. *J Cogn Neurosci*. 21:1488–1498.
- Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, et al. 2013. Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. *Neuron*. 77:980–991.