# The Impact of Selective Attention and Musical Training on the Cortical Speech Tracking in the Delta and Theta Frequency Bands

**Alina Schüller*, Annika Mücke*, Jasmin Riegel, and Tobias Reichenbach**

## Abstract

■ Oral communication regularly takes place amidst background noise, requiring the ability to selectively attend to a target speech stream. Musical training has been shown to be beneficial for this task. Regarding the underlying neural mechanisms, recent studies showed that the speech envelope is tracked by neural activity in auditory cortex, which plays a role in the neural processing of speech, including speech in noise. The neural tracking occurs predominantly in two frequency bands, the delta and the theta bands. However, much regarding the specifics of these neural responses, as well as their modulation through musical training, still remain unclear. Here, we investigated the delta- and theta-band cortical tracking of the speech envelope of target and distractor speech using magnetoencephalography (MEG) recordings. We thereby assessed both musicians and nonmusicians to explore potential differences between these groups. The cortical speech tracking was quantified through source-reconstructing the MEG data and subsequently relating the speech envelope in a certain frequency band to the MEG data using linear models. We thereby found the theta-band tracking to be dominated by early responses with comparable magnitudes for target and distractor speech, whereas the delta band tracking exhibited both earlier and later responses that were modulated by selective attention. Almost no significant differences emerged in the neural responses between musicians and nonmusicians. Our findings show that only the speech tracking in the delta but not in the theta band contributes to selective attention, but that this mechanism is essentially unaffected by musical training. ■

## INTRODUCTION

Auditory selective attention refers to the ability of the human brain to segregate spatiotemporally overlapping speech streams into distinct auditory objects and to selectively attend one of them (Ding & Simon, 2012). However, this ability requires significant cognitive resources and can be impeded by several factors, such as hearing impairment (Coffey, Mogilever, & Zatorre, 2017; Kraus & Chandrasekaran, 2010), speech-related learning impairment (Kraus & Chandrasekaran, 2010), or age-related decline in the ability to attend to a target speaker (Parbery-Clark, Strait, Anderson, Hittner, & Kraus, 2011; Souza, Boike, Witherell, & Tremblay, 2007).

In contrast, musical training may prevent or even counteract difficulties in speech-in-noise perception (Coffey et al., 2017; Du & Zatorre, 2017; Parbery-Clark et al., 2011; Strait & Kraus, 2011; Kraus & Chandrasekaran, 2010). Possible explanations for this hypothesis include that musical training can enhance brain plasticity (Du & Zatorre, 2017) and functional connectivity (Puschmann, Regev, Baillet, & Zatorre, 2021; Du & Zatorre, 2017), increase the auditory working memory (Clayton et al., 2016; Strait & Kraus, 2011; Parbery-Clark, Skoe, & Kraus, 2009), and thus improve auditory fitness (Kraus & Chandrasekaran, 2010), especially when the training started at a young age (Zuk et al., 2013; Kraus & Chandrasekaran, 2010).

However, despite the importance of speech-in-noise comprehension for human oral communication and social life, the underlying neural processes, as well as the neural mechanisms leading to declined or improved abilities, are not yet fully understood. Recent investigations have employed noninvasive neuroimaging through electroencephalography or magnetoencephalography (MEG) while participants listen to naturalistic speech in noise. Through subsequent statistical modeling, these recordings allow quantifying ongoing neural responses to repetitive, rhythmic aspects of speech stimuli, often referred to as neural speech tracking (Chen et al., 2023; Gillis, Van Canneyt, Francart, & Vanthornhout, 2022; Brodbeck & Simon, 2020; Ding & Simon, 2012).

Perhaps the most robust neural tracking emerges in response to the speech envelope, a signal in the low-frequency range, between 1 and 15 Hz, that traces the amplitude fluctuations of a speech signal (Brodbeck & Simon, 2020; Ding & Simon, 2013). The neural tracking of the speech envelope does not simply reflect the bottom–up processing of the acoustic input, but is also

Friedrich-Alexander-Universität Erlangen-Nürnberg
*These authors contributed equally to this work.

influenced by higher cognitive processes, in particular by selective attention, which enhances the tracking (Brodbeck & Simon, 2020; O'Sullivan et al., 2015; Ding & Simon, 2012). Moreover, modulating the neural tracking through transcranial alternating current stimulation can impact and even enhance speech-in-noise comprehension (Keshavarzi, Kegler, Kadir, & Reichenbach, 2020; Kadir, Kaza, Weissbart, & Reichenbach, 2019; Wilsch, Neuling, Obleser, & Herrmann, 2018; Zoefel, Archer-Boyd, & Davis, 2018; Riecke, Formisano, Sorger, Başkent, & Gaudrain, 2018).

Different functions have been attributed to cortical tracking in the theta band (4–8 Hz) and delta band (1–4 Hz; Ding & Simon, 2014). Theta-band tracking probably reflects the parsing of lower level speech components, such as syllables and phonemes (Mai & Wang, 2023; Brodbeck, Presacco, & Simon, 2018; Di Liberto, O'Sullivan, & Lalor, 2015; Ding & Simon, 2014), and reflects the acoustic clarity (Etard & Reichenbach, 2019). Delta-band tracking is associated with neural responses to words, reflects higher level linguistic processing and can inform on speech comprehension (Mai & Wang, 2023; Etard & Reichenbach, 2019; Ding & Simon, 2013). Theta-band tracking has indeed been found to be more sensitive to stationary background noise, whereas delta band tracking seems to be robust as long as the speech stimulus is still intelligible (Van Hirtum, Somers, Verschueren, Dieudonné, & Francart, 2023; Ding & Simon, 2013). However, the specific contributions of delta- and theta-band tracking to selective attention have not yet been fully clarified.

The neural specialization that may allow musicians (Ms) to achieve better speech-in-noise comprehension has been investigated through noninvasive neuroimaging as well, mostly focusing on the evaluation of short auditory stimuli (Zendel, Tremblay, Belleville, & Peretz, 2015; Parbery-Clark, Anderson, Hittner, & Kraus, 2012; Zendel & Alain, 2009; Parbery-Clark et al., 2009). These investigations found, in particular, enhanced subcortical responses of people with musical training (Musacchia, Sams, Skoe, & Kraus, 2007; Parbery-Clark, Anderson, et al., 2012; Parbery-Clark, Tierney, Strait, & Kraus, 2012), as well as a larger right-hemispheric recruitment of neural resources in auditory cortex (Jantzen, Howe, & Jantzen, 2014).

However, there has only been very little work investigating the impact of musical training on cortical speech tracking. As a notable exception, Puschmann, Baillet, and Zatorre (2019) found that in musically trained individuals, attentional modulation of the cortical speech tracking was less pronounced, suggesting that the ignored stream is represented stronger in cortical activity (Puschmann et al., 2019). However, they did not investigate delta and theta band cortical tracking separately.

Here, we sought to differentiate the modulation of the neural tracking in the delta and in the theta band through selective attention, as well as how these individual responses may be shaped by musicianship. We hypothesized that attentional modulation only acts on the higher level processing associated to the delta band, but not on the lower level processing in the theta band, because the attentional effects have been observed comparatively late, at delays of 140 msec and later (Brodbeck, Jiao, Hong, & Simon, 2020). We further hypothesized that the attentional modulation of the delta-band tracking is more pronounced in Ms, contributing to an enhanced behavioral performance in speech-in-noise listening.

## METHODS

### Experimental Design

We used an MEG data set from 52 participants (26 female, 26 male) aged 24.1 ± 3.1 years that was acquired in the scope of two of our previous studies (Riegel, Schüller, & Reichenbach, 2024; Schüller, Schilling, Krauss, Rampp, & Reichenbach, 2023). All participants were right-handed and native German speakers. They had no history of neurological disease or hearing impairment. The study was approved by the ethics board of the University Hospital Erlangen (Registration No. 22–361-S).

Participants listened to two German audiobooks simultaneously, both of which were narrated by male speakers (Figure 1). Two audiobooks were used alternatingly as to-be-attended audiobooks, which the participants were instructed to focus on. These two audiobooks will be referred to as the target audiobooks in the following. The first target audiobook was *Frau Ella* written by Florian Beckerhoff and read by Peter Jordan (Speaker 1). The second target audiobook was *Den Hund überleben* written by Stefan Hornbach and read by Pascal Houdus (Speaker 2). Simultaneously, a randomly selected excerpt from an unrelated audiobook read by the other speaker, respectively, was presented as a to-be-ignored audiobook. These will be referred to as distractor audiobooks. The first distractor audiobook was thus read by Speaker 2, Pascal Houdus, and was titled *Looking for Hope* by Colleen Hoover (translated into German by Katarina Ganslandt). The second distractor audiobook was *Darum* by Daniel Glattauer and read by Speaker 1, Peter Jordan. All of the employed audiobooks were published by Hörbuch Hamburg and are available in stores. Speaker 1 read with a mean word frequency in the first target audio of 3.6 Hz with an average syllable frequency of 5.8 Hz. In the second target audio, Speaker 2 showed comparable speech characteristics with a mean word frequency of 3.7 Hz and a mean syllable frequency of 5.7 Hz. The audio data, as well as the MEG recordings, will be published upon submission.

The experiment was split into 10 trials. Each trial consisted of listening to one chapter from one of the target audiobooks. Simultaneously, a random duration-matched excerpt from the corresponding distractor audio stream was presented. Once the audiobook chapter ended, participants had to answer three single-choice listening comprehension questions about this chapter to ensure that the participants did indeed focus on the target audio. After

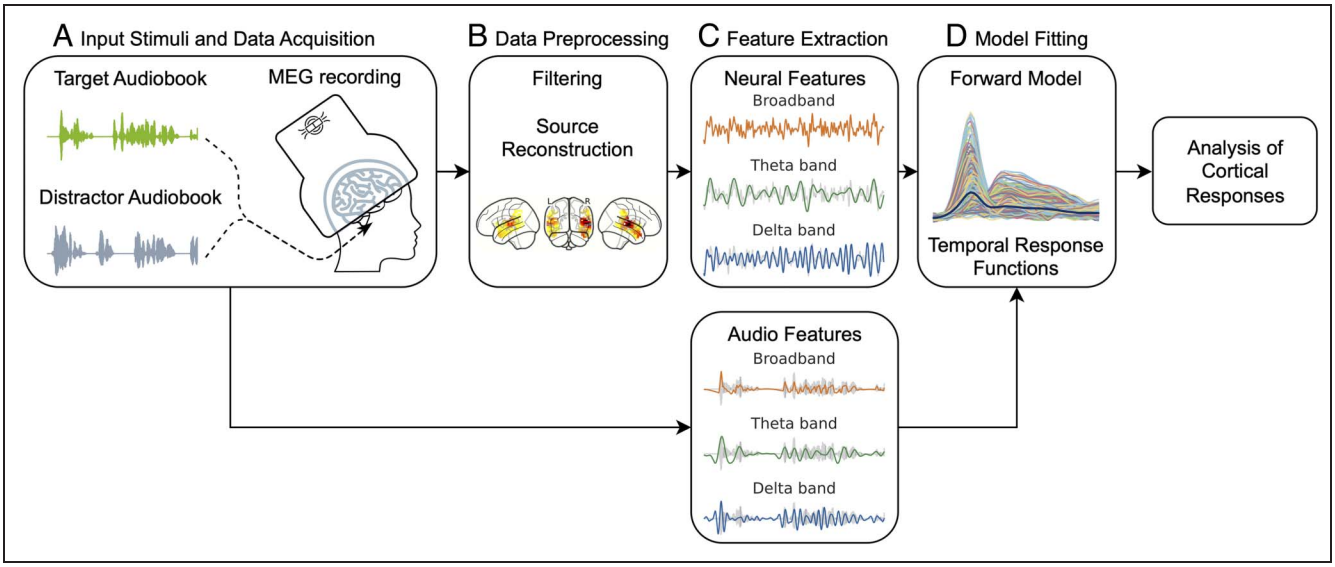


**Figure 1.** Overview of the experimental setup and the data processing pipeline. (A) Participants were presented with two audiobooks while attending one and ignoring the other. Simultaneously, MEG was measured. (B) On the basis of the acquired data, source reconstruction with focus on auditory cortex was performed. (C) From the source-level data, the frequency bands of interest were extracted using bandpass filtering. The corresponding audio features were extracted from the acoustic envelopes of the distractor and target audio, respectively. (D) Lastly, temporal response functions (TRFs) were calculated by training a forward model to estimate the neural features from the audio input features. The resulting TRF magnitudes were compared between the target and distractor condition, as well as between Ms and NMs.

each trial, the target audio, and thus the target and distractor speakers, were switched. This was repeated until the first five chapters of both target audios were presented. The alternation served to avoid a potential speaker bias within each session, while maintaining the continuity of the audiobook's narrative. The chapters had varying lengths between 3 and 5 min, resulting in a total recording duration of 37 min. The audios were presented diotically at equal sound pressure levels of 67 db SPL(A).

MEG data were acquired with the participants in a supine position and with their eyes focused on a fixation cross displayed on a screen above. Their head was placed in a 248-channel MEG system (4D Neuroimaging). Before the start of the measurement, the head shape of each participant was captured using a digitizer (Polhemus). In addition, the head position relative to the MEG system was recorded using five integrated head-position indicator coils. MEG data were recorded with a sampling frequency of 1017.25 Hz. During online processing, an analog bandpass filter (1–200 Hz) and a noise reduction algorithm (4D Neuroimaging) using 23 reference sensors were applied to eliminate environmental noise.

The audio stimuli were presented to the participants using a customized setup that was validated and described in more detail in previous work (Schilling et al., 2021). In brief, it consisted of two flexible tubes that were connected to loudspeakers and led into the magnetically shielded MEG chamber, connecting to earphones that the participants wore. Simultaneously, the presented audio signal was fed as an additional input channel to the MEG data logger. This enabled the synchronization of MEG recordings and the auditory stimuli with a 1 precision.

## Assessment of Musical Training

Participants were assigned as M or nonmusicians (NMs) based on their history of musical training. This categorization was performed using previously established criteria (Riegel et al., 2024), namely, starting age of playing an instrument, total years of undergoing musical training, and whether they were currently practicing.

To be classified as an M, a participant had to start playing before the age of 7 years for a total period of 10 or more years. Furthermore, they had to regularly play an instrument at the time of the study. On the other hand, participants were classified as a NM if they did not undergo musical training any longer than 3 years in sum and only started aged older than 7 years.

On the basis of these criteria, we acquired data from 25 NMs (12 female, 13 male, aged 24.5 ± 3.4 years) and 18 Ms (10 female, 8 male, aged 24.1 ± 3.1 years). Nine participants (4 female, 5 male, aged 24.2 ± 3.9 years) did not fit into either the M or NM category. For the comparison between Ms and NMs, these nine participants were excluded, resulting in data from 43 participants. This number of participants is in line with previous studies that investigated speech processing with regard to the musicality of an individual (Parbery-Clark, Anderson, et al., 2012; Parbery-Clark, Tierney, et al., 2012; Musacchia et al., 2007). An analysis that disregarded musical training was conducted using the data from all 52 participants.

## Data Analysis

Data processing and analysis were performed in Python using the libraries *MNE* (Gramfort et al., 2013), *SciPy* (Virtanen et al., 2020), and *statsmodels* (Seabold & Perktold, 2010).

### *MEG Data Preprocessing*

First, the acquired MEG recordings were cut to the intervals of interest where an acoustic stimulus was present. These pieces were then concatenated.

The MEG data were further processed with a notch filter (50 Hz) to remove a powerline interference. Afterward, the frequency range of interest was extracted using a Butterworth bandpass filter (1–20 Hz, order $n = 4$). This filter was applied forward and backward to prevent introducing a filter shift. To increase computational efficiency, the filtered data were downsampled to 100 Hz.

For source reconstruction, we used the average MRI brain template *fsaverage* provided by the FreeSurfer software package (Fischl, 2012). The computational steps were performed using the Python package MNE (Gramfort et al., 2013). The *fsaverage* template was aligned to the participant-specific head positioning and head shape that was collected before each measurement.

Next, a volume source space consisting of an equidistant grid of sources was created. As our analysis focused on auditory processes, sources were exclusively created in auditory cortex and its surrounding region, including the middle temporal gyrus, the transverse temporal gyrus, the superior temporal gyrus and its banks, the supramarginal gyrus, and the insular lobe, in both the left and right hemispheres. For the subcortical segmentation and cortical parcellation of the brain regions within the volume, the *aparc+aseg* template from FreeSurfer was employed. Sources were selected to have a distance of 5 between one another, resulting in 525 distinct source points.

Using the head model and the source space, a forward solution was calculated estimating the magnetic field strength at each MEG channel produced by the sources. To account for the different types of tissue in the brain and their varying conductivities, the boundary element method model provided by FreeSurfer for the *fsaverage* template was employed. The resulting leadfield matrix characterizes the sensitivity of each MEG channel to the activity of each of the selected sources.

A spatial filter was then applied to reconstruct the source activity from the sensor-level data and the estimated leadfields. Here, a linearly constrained minimum variance beamformer (Van Veen, Van Drongelen, Yuchtman, & Suzuki, 1997) was employed to estimate the activation of each source independently while deducting environmental noise measured in the empty MEG chamber.

For analyzing neural tracking in the delta and theta bands, the data were further filtered in the corresponding frequency range. To this end, we employed a forward-backward Butterworth bandpass filter (order $n = 4$) in the range of 1–4 Hz and 4–8 Hz, respectively.

### *Speech Envelope*

The audio files had a sampling frequency of 44.1 kHz. To retrieve the acoustic speech envelope, the analytic representation was obtained by applying a Hilbert transform. The magnitude of this analytic signal was then calculated, as it represents the instantaneous amplitude and hence the envelope of this signal.

To analyze the neural tracking of the audio, we filtered the speech envelope in three different frequency bands. On the one hand, a broad range of frequencies ranging from 1 and 20 Hz was extracted, which is referred to as *broadband* feature in the following. On the other hand, two narrower frequency bands were analyzed as well, namely, the *delta band* (1–4 Hz) and the *theta band* (4–8 Hz). The envelope was then filtered within the desired frequency bands using a Butterworth bandpass filter (order $n = 4$) applied forward and backward to prevent filtering delays using the SciPy function *sosfiltfilt* (Virtanen et al., 2020). The resulting features were then resampled to 100 to match the sampling rate of the preprocessed, source-level MEG data.

### *Temporal Response Functions (TRFs)*

To analyze the neural tracking of the extracted audio features, we determined the relation between them and the neural data. Using the extracted audio features as input, we calculated forward models to predict the corresponding neural responses to both the target and distractor audio and for each frequency band.

For this purpose, a linear model was trained that predicted the neural response $y_t^{(s)}$ for every point $t$ in time at every neural source $s$ based on a linear combination of acoustic feature values $x_{t-\tau}$ covering the time interval $[t - \tau_{\max}, t - \tau_{\min}]$. For each time delay $\tau$, a coefficient $\alpha_\tau^{(s)}$ was estimated. $\tau_{\min}$ was chosen to be negative to verify the plausibility of the resulting model weights and to gain information about the noise level present in the TRF. As not all variation in the neural signal can be explained by the auditory input, a residual $\epsilon_t^{(s)}$ remains for every investigated time lag:

$$y_t^{(s)} = \sum_{\tau=\tau_{\min}}^{\tau_{\max}} \alpha_\tau^{(s)} x_{t-\tau} + \epsilon_t^{(s)}. \quad (1)$$

The coefficients $\alpha_\tau^{(s)}$ are referred to as the TRF of the source point $s$. The relative magnitudes of these coefficients characterize to which extent each delay contributes to the neural signal and, thus, how strong the neural response to the auditory input is after a certain time lag.

Local maxima in the magnitudes of the TRFs indicate that neural responses are particularly pronounced at these time delays.

However, unconstrained linear regression is not the optimal approach for speech features as an input signal, as it likely contains temporal correlations causing numerical instability when fitting the model (Crosse, Di Liberto, Bednar, & Lalor, 2016). Furthermore, because of the high amounts of input data and parameters, the model is prone to overfitting. To address these issues and thus retrieve more robust and reliable results, we employed ridge regression (Schüller et al., 2023; Kegler, Weissbart, & Reichenbach, 2022; Etard, Kegler, Braiman, Forte, & Reichenbach, 2019; Biesmans, Das, Francart, & Bertrand, 2017). The TRF coefficients $\alpha^{(s)} \in \mathbb{R}^{(\tau_{\max}-\tau_{\min})\times 1}$ were accordingly approximated as

$$\alpha^{(s)} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}^{(s)} \tag{2}$$

where $\boldsymbol{X} \in \mathbb{R}^{\Delta t\times(\tau_{\max}-\tau_{\min})}$ contains the speech envelope information over the entire stimulus length $\Delta t$, $\boldsymbol{y}^{(s)} \in \mathbb{R}^{T\times 1}$ describes the neural recording at source $s$, $\lambda$ is a predefined regularization parameter, and $\boldsymbol{I}$ refers to the identity matrix. A Python implementation of this TRF coefficient estimation method that was previously developed and utilized by our group (Kegler et al., 2022; Etard et al., 2019) was used. To retrieve meaningful results, all speech features and neural features were scaled using $z$ score normalization.

To select the regularization parameter, a subject-wise, five-fold cross-validation testing 12 parameters ranging from $10^{-5}$ to $10^{5}$ ($\lambda \in [0, 1e{-}5, 1e{-}4, 1e{-}3, 1e{-}2, 0.1, 0.5, 1, 5, 10, 1e2, 1e5]$) was applied. The respective optimal $\lambda$ was calculated based on the prediction that yielded the highest correlation (Pearson correlation coefficient $r$) with the measured neural signal. Because of similar results across all participants, the mean optimal $\lambda$ from the search space was used for all TRF coefficients per investigated frequency band to ensure comparability of the results. The resulting regularization parameters are listed in Table 1. Two separate forward models were trained, one that employed the distractor sound to predict the MEG signal and another that used the target sound for reconstructing the MEG measurement. This was done for each frequency band (broadband, delta and theta) to yield a corresponding TRF. The chosen regularization parameters for target and distractor TRFs were, however, equal within the frequency bands.

The time lag intervals were selected based on visual inspection of the TRF estimates and are depicted for all frequency bands in Table 1. A step size of 10 msec between each time step was used, which corresponds to the sampling frequency of 100 Hz of the employed features.

**Table 1.** Overview of the Parameters Employed for Computing the TRFs

| *Frequency Band* | $\tau_{min}$ | $\tau_{max}$ | $\lambda$ |
|---|---|---|---|
| Broadband | −100 | 700 | 5 |
| Delta band | −300 | 1300 | 100 |
| Theta band | −200 | 700 | 5 |

### *TRF Peak Extraction*

Distinct TRF coefficients were estimated for all participants and the two listening modes *target* and *distractor*. Our main focus was to identify the most prominent responses and corresponding time lags and compare their intensities across different groups and conditions. Thus, to simplify the subsequent analysis, only the magnitudes of the individual TRFs were considered further. The TRF magnitudes were min–max normalized at the population level. In particular, we determined both the maximal and the minimal value of the TRFs across time lags and participants and used these two values to normalize the TRFs. The normalized TRFs of most participants therefore had maxima that were below 1.

To determine the magnitude and average latency of the distinct peaks in the TRFs, the average TRF magnitudes were investigated at the population level. They were calculated by averaging the TRF magnitudes over all source points, all participants, and both presented speakers, for both target and distractor listening modes. On the basis of these participants' average responses, the most prominent peaks and their latencies were identified for each frequency band and listening mode.

As the TRF peak latencies differed from one individual to another, we found that information loss occurred when individual peak magnitude values were retrieved by simply considering participants' average peak latencies. Thus, we also extracted a set of individual TRF peak latencies and magnitudes from the source-averaged TRF magnitudes for each individual participant. To avoid distorting the results with very early or late outlier peaks that may appear in individual TRFs because of noise or artifacts, we defined a fixed search interval around each average peak per frequency band. For every peak type, all local maxima were identified and the respective maximum closest to the average was selected as the individual peak. If no local maximum was found in the search interval, the individual TRF was excluded from the analysis of the respective peak.

### *Attentional Modulation of Cortical Responses*

To investigate the impact of attention on the cortical responses, TRF magnitudes were investigated on the single-subject level. For each participant, cortical responses to the target and distractor audios were retrieved, using the envelope from the target and distractor streams as input features, respectively. Thus, the two listening modes *target* and *distractor* were compared with

evaluate differences in cortical tracking of the respective speech envelopes. TRF magnitudes were obtained by averaging over all sources. From these subject-level TRFs, individual TRF peaks characterized by peak amplitudes and peak latencies were extracted as described above and compared between listening modes.

To quantify the attentional modulation of the cortical responses for each participant and to furthermore compare this measure between the NM and M groups, an attentional modulation index (*AI*) was calculated for each participant, speaker, frequency band, and TRF peak $p$ as

$$AI^{(p)} = \frac{A_{\text{tar}}^{(p)} - A_{\text{dis}}^{(p)}}{A_{\text{tar}}^{(p)} + A_{\text{dis}}^{(p)}} \tag{3}$$

where $A_{\text{tar}}^{(p)}$ describes the individual peak amplitude at peak $p$ for the target and $A_{\text{dis}}^{(p)}$ for the distractor audio feature, respectively. Values of *AI* close to zero indicate similar response strengths to the distractor and target stream and, hence, no attentional modulation. In contrast, positive values of *AI* suggest the existence of attentional modulation by attenuation of the distractor stream or amplification of the target stream, whereas negative values of *AI* imply the opposite effect, that is, a stronger cortical representation of the distractor than the target stream.

## Statistical Analysis

All statistical tests were performed using the Python package *statsmodels* (Seabold & Perktold, 2010) and the *stats*-module from the Python package SciPy (Virtanen et al., 2020). The significance level was set to $p < .05$ for all tests.

### *Significance of Peaks in Population-level TRFs*

The population-level TRFs for both the target and the distractor listening mode were obtained by averaging over the magnitudes of the TRF coefficient from all participants and sources. The intervals in which these population-level TRF magnitudes were significantly different from noise were determined using a bootstrapping approach that compared the magnitudes to a distribution of noise models. Participant- and source-specific noise models were generated by fitting the ridge regression model on a reversed version of the auditory input features. Therefore, the noise models did not contain any meaningful relationship between the input feature and the predicted neural output.

The resulting individual noise model weights were randomly shuffled across neural sources and participants and then averaged in the same manner as the individual TRFs to generate the participants' average noise over all regarded time lags. This procedure was repeated 10,000 times to obtain a noise model distribution. Using this distribution, empirical $p$ values were calculated for each TRF and time lag based on the amount of noise-level values with a lower magnitude than the investigated actual TRF magnitude. To account for multiple comparisons across all time lags, the resulting $p$ values were revised using the Bonferroni correction.

### *Significance of Model Performance*

We assessed the predictive performance of the TRF model for each frequency band and each condition (target and distractor) by calculating the Pearson correlation coefficient between the predicted and measured MEG source activity for both target and distractor conditions in each frequency band. To statistically compare the so-obtained model prediction scores, we calculated null model performance scores in the same way using the above described noise models. We then performed a Wilcoxon signed-rank test to evaluate the statistical difference between the actual model performance and the null model performance on the population level.

To test whether the model performance of the target model significantly differed from the distractor model performance, we conducted a Wilcoxon signed-rank test as well.

### *Significance of Lateralization*

To determine how the spatial distribution of cortical responses to natural speech differs between earlier and later time delays and across the investigated frequency bands, source-specific average TRFs were calculated. This was achieved by averaging the source-level TRFs over all participants and target audios. Only the target audios were incorporated in this analysis, as we wanted to analyze the spatial distribution of responses independent of possible top–down effects introduced by attentional modulation. As a result, the cortical responses to the target speech envelope for each of the 525 source points in the area of auditory cortex were retrieved. Snapshots from the source-level TRFs were extracted at each of the previously determined average response peak latencies. The momentary spatial distribution of intensities in the source space at these time lags was visualized through brain plots.

To assess whether potential differences in these responses occurring between sources in the right and left hemispheres were statistically significant, a two-sided Mann–Whitney U test was performed. This test was undertaken with the TRF magnitudes at the latencies of the population-level TRF peaks in all frequency bands. To correct for multiple comparisons for the different peaks within a frequency band, the Bonferroni correction was applied.

To further quantify the extent of lateralization at the response peaks yielding significant differences between both hemispheres, a participants' average lateralization index $LI^{(p)}$ for every peak $p$ and frequency band was

calculated as already applied similarly in previous literature (Seghier, 2008):

$$LI^{(p)} = \frac{R^{(p)} - L^{(p)}}{R^{(p)} + L^{(p)}} \quad (4)$$

where $R^{(p)}$ and $L^{(p)}$ describe the sum of peak magnitudes from sources in the right and left hemispheres, respectively. Thus, a negative $LI^{(p)}$ value indicates left-lateralized neural activity, whereas a positive $LI^{(p)}$ value implies a right-lateralized neural processing. The computation of the LI allows the comparison with other lateralized responses reported in the literature.

### *Significance of Attentional Modulation*

The degree to which target and distractor speech were tracked differently in auditory cortex was assessed at the population level using different characteristics of the TRF peaks, such as peak amplitude, peak latency, and the AI. When a certain TRF peak was not present in one listening mode for a particular participant, for example, for the distractor audio, the respective peak was also excluded for the respective other mode, for example, the target audio, to allow a mode-wise comparison within participants. Statistical significance of differences in these metrics within the same group of participants, that is, when comparing target and distractor speech responses for all participants, were assessed using a two-sided pairwise $t$ test if the underlying data of the respective characteristic was normally distributed. Otherwise, a two-sided Wilcoxon signed-rank test was applied. The normality of the analyzed samples was determined using the Shapiro–Wilk test.

When comparing characteristics between Ms and NMs, a two-sided unpaired $t$ test was utilized for normally distributed characteristics. Alternatively, for samples that did not follow a normal distribution, a two-sided Mann–Whitney U test was utilized. For all tests within a frequency band, the Bonferroni correction was applied to compensate for multiple comparisons across the different TRF peaks.

### *Relationship of Listening Comprehension and Cortical Response Strength*

During the study, participants answered single-choice listening comprehension questions regarding the target audio stream. The percentage of correct answers per participant and target audiobook was captured as a behavioral metric to quantify the performance of the individual participant. To relate the cortical response to this behavioral measure, Spearman's correlation coefficients were calculated between the comprehension score and each TRF peak per feature and listening mode. The Bonferroni correction was used to adjust for multiple comparisons between different TRF peaks for all correlations within a frequency band. Furthermore, to evaluate the listening performance of Ms and NMs on a purely behavioral level, a two-sided unpaired $t$ test was performed to compare the comprehension scores of the two groups.

### *Relationship of Musicality Features and Cortical Response Strength*

As detailed above, participants were asked about different aspects of their musical training: their starting age of playing an instrument, total years of undergoing musical training, amount of current training, and the total number of instruments played. (Note that only the first three aspects were used to classify the participants into the three categories.)

Using a linear regression model, we assessed if the TRF peak characteristics related to these reported musicality aspects. A regression model was estimated for each peak using ordinary least squares. Because of the large number of estimates, this analysis was only performed for the delta and theta bands. All independent and dependent variables were scaled beforehand using $z$ score normalization. The musicality aspects served as independent variables, whereas the peak values were used as the dependent predicted features.

## RESULTS

### Temporal and Spatial Distribution of Cortical Responses to Target Speech

As a first step, we assessed the temporal and spatial distribution of the neural speech tracking in the three employed frequency bands at the population level (Figure 2). For the broadband responses, the average magnitude of the TRFs exhibited peaks at the time lags 110 msec and 240 msec (Figure 2A). In the delta band, peak neural responses emerged at the time lags 100 msec, 270 msec, and 540 msec. The cortical tracking in the theta band led to a main peak at 130 msec. The symmetrical sidelobes around this peak are caused by the narrow bandwidth of the speech feature and were therefore not regarded as independent peaks.

In the following, responses peaking around 100 msec and 200 msec are referred to as $M100_{TRF}$ and $M200_{TRF}$, respectively. Furthermore, the late TRF response peak observed in the delta band is referred to as $M400_{TRF}$ in analogy to the N400 component in ERPs (Kutas & Federmeier, 2011). We thus found a $M100_{TRF}$ in all three frequency bands, a $M200_{TRF}$ in the broadband and the delta band, and a $M400_{TRF}$ solely in the delta band.

$M100_{TRF}$ showed a highly significant right-lateralization in all three frequency bands, with $LI = 0.26$ ($U = 18864$, $p < .001$, corrected for two comparisons) in the broadband, $LI = 0.36$ ($U = 11133$, $p < .001$, corrected for three comparisons) in the delta band, and $LI = 0.15$ ($U = 22770$, $p < .001$) in the theta band (Figure 2B). A similar degree of right lateralization was observed for $M200_{TRF}$ with $LI = 0.26$ ($U = 13988$, $p < .001$, corrected for two comparisons) in the broadband, and $LI = 0.36$ ($U = 9577$, $p < .001$, corrected for three comparisons) in the delta band.

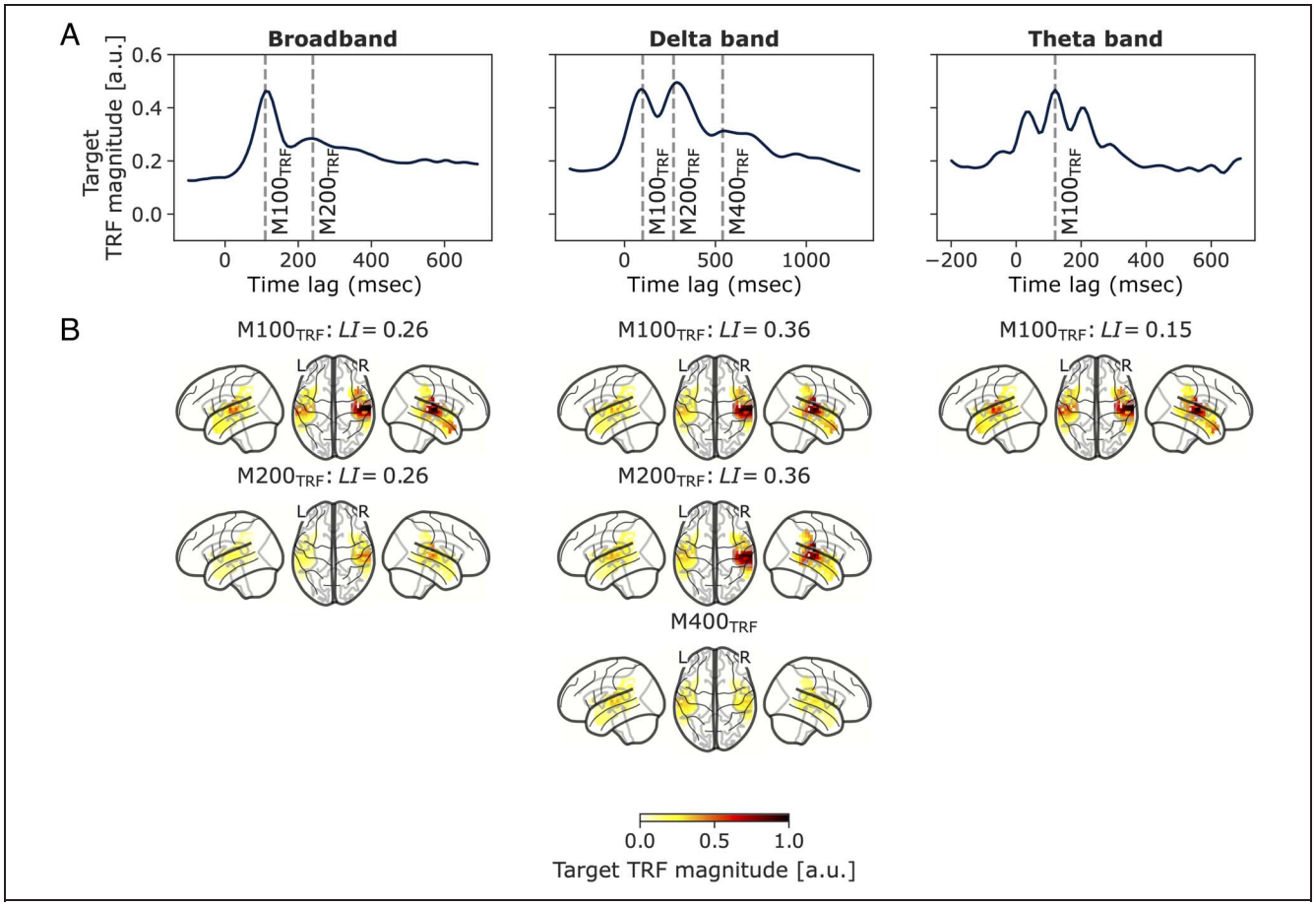


**Figure 2.** Population-level magnitudes of the averaged TRFs of target speech and their spatial distribution for the distinct frequency bands. (A) The mean magnitudes of the TRFs across all source channels and participants for target speech for each investigated frequency band. On the basis of these functions, the time lags at which cortical responses were maximal (dotted vertical lines) were extracted and assigned to one of the response types $M100_{TRF}$, $M200_{TRF}$, or $M400_{TRF}$. (B) The distribution of TRF magnitudes at these maximum RTs in the source volume ranging from low activation (yellow) to high activation (dark red). For brainplots showing significant differences in activations of right and left hemispheres, the respective LIs are reported. The colored area corresponds to the investigated source ROI.

In contrast, $M400_{TRF}$ showed a bilateral distribution of activations ($LI = 0.01$, $U = 33656$, $p = .99$, corrected for three comparisons).

### Attentional Modulation of Cortical Responses

In the next analysis step, TRF responses to the distractor audios were compared with the target responses at the population level (Figure 3). For all three frequency bands, the TRF magnitudes were significantly higher than those of the noise model for the majority of the analyzed positive time lags. Moreover, the model performances obtained by comparing the predicted and measured source activity in all three frequency bands for both conditions, target as well as distractor, significantly exceeded the null model performances (Wilcoxon signed-rank test, $p < .001$ for all frequency bands and conditions).

The model performance scores obtained for each frequency band and each condition (target or distractor), averaged across participants, are displayed in Table 2. To investigate the difference in the predictive power for the TRF models based on target speech and the TRF models based on distractor speech in the three frequency bands, we conducted a Wilcoxon signed-rank test. The resulting statistics with $p$ values are shown in Table 2, indicating significant differences in the model performances for the broadband and the delta band, with the model for the target condition significantly outperforming the model for the distractor condition. For the theta band, however, no significant difference in model performances regarding target versus distractor conditions was found.

Analogous to the target TRFs, the distractor TRFs in the broadband condition exhibited both $M100_{TRF}$ (time lag 110 msec) and $M200_{TRF}$ (time lag 200 msec). The same responses were also visible in the distractor TRFs in the delta band, with $M100_{TRF}$ and $M200_{TRF}$ occurring with average delays of 80 msec and 250 msec, respectively. In addition, a weak $M400_{TRF}$ response was extracted at 520 msec to allow a direct comparison with the target responses. As for the target TRFs, the distractor TRFs in

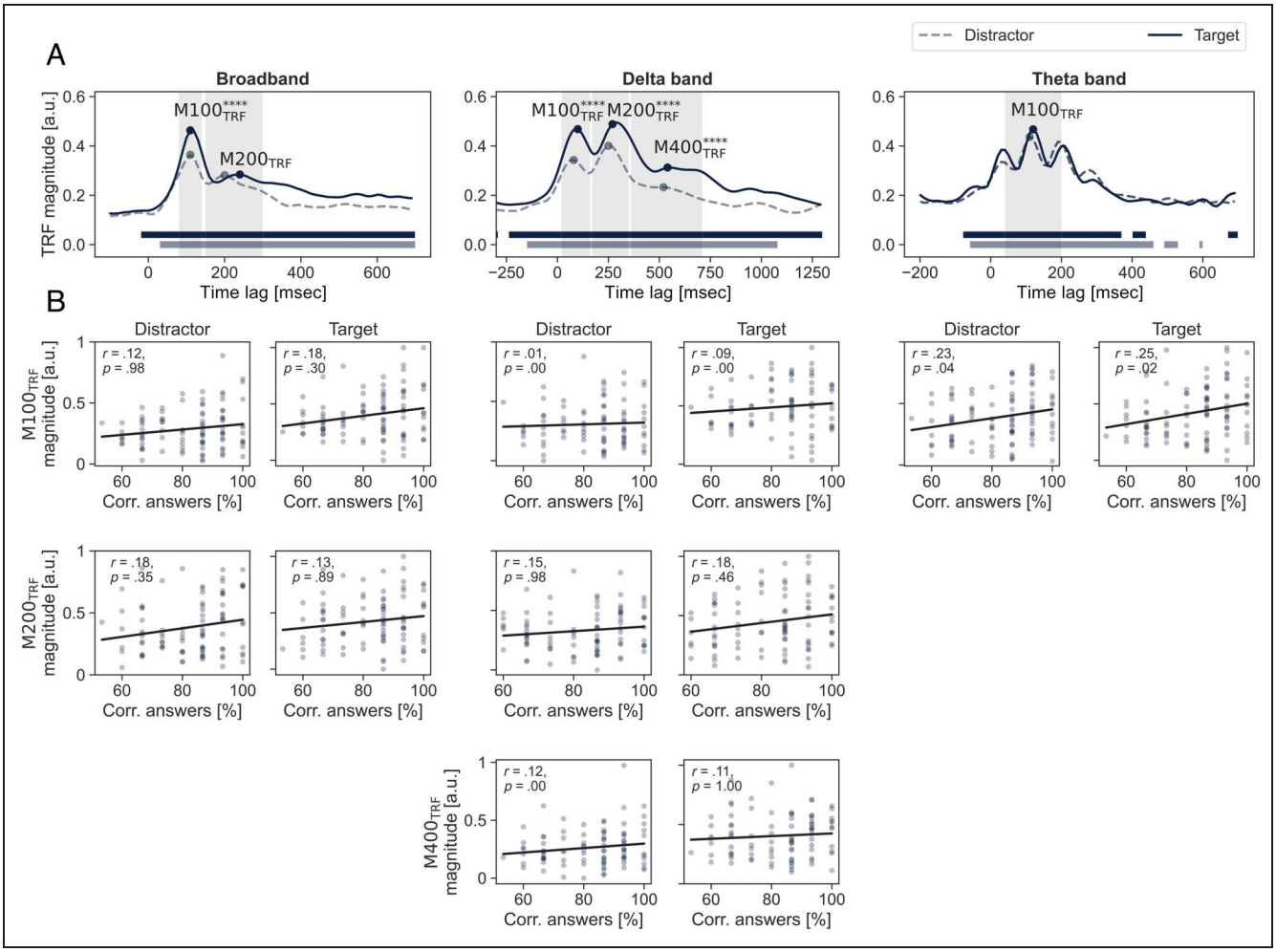


**Figure 3.** Population-level TRF magnitudes in the target and distractor condition, as well as their relation to the comprehension scores. (A) The TRF magnitudes in the target condition (black solid) are larger than those in the distractor condition (gray dashed) for the broadband and the delta-band response, but not in the theta band. The horizontal bars indicate the areas where the TRF magnitudes significantly differ from the noise models. Peaks with significant differences in amplitude between the target and distractor condition are marked with asterisks indicating their significance level (*$0.01 \leq p < .05$; **$0.001 \leq p < .01$, ***$0.0001 \leq p < .001$, ****$p < .0001$). The search areas for the participant-wise TRF response peaks are shaded in gray. (B) The individual magnitudes at each peak were extracted per participant and related to the percentage of correct comprehension question answers through Spearman's correlation coefficient $r$. A significant positive correlation between response magnitude and comprehension was found in the theta band, both for the target and the distractor condition.

the theta band showed a single main peak, $M100_{TRF}$ (time lag 120 msec).

To further quantify the differences between responses in the target and distractor condition, individual time lags were extracted for each participant, listening scenario, and response type. The utilized search intervals for each frequency band are depicted in Table 3. For the broadband feature, a $M100_{TRF}$ was found in 94% of all participant-level TRFs, whereas a $M200_{TRF}$ was present in 91% of broadband TRFs. In the delta band, $M100_{TRF}$ appeared in 84%,

**Table 2.** Average Model Performance Scores for the Target TRF Model and the Distractor TRF Model of Each Frequency Band

| | *Broadband* | *Delta Band* | *Theta Band* |
|---|---|---|---|
| Target | 0.024 | 0.035 | 0.016 |
| Distractor | 0.017 | 0.021 | 0.015 |
| $z$, $p$ value (target vs. distractor) | 402, <0.001 | 266, <0.001 | 2530, 0.37 |

The statistical difference evaluation between target and distractor conditions are quantified by the corresponding $z$ statistic and $p$ value of a Wilcoxon signed-rank test.

**Table 3.** Search Intervals for the Peaks in the Subject-level TRF Magnitudes for the Different Frequency Bands and the Corresponding Proportion of Participants for Whom the Respective Peak Was Found

| *Frequency Band* | *TRF Response* | *Search Interval* | *Detected Peaks (%)* |
|---|---|---|---|
| Broadband | $M100_{TRF}$ | [80, 140] | 94 |
| | $M200_{TRF}$ | [150, 300] | 91 |
| Delta band | $M100_{TRF}$ | [20, 160] | 84 |
| | $M200_{TRF}$ | [170, 350] | 91 |
| | $M400_{TRF}$ | [360, 710] | 90 |
| Theta band | $M100_{TRF}$ | [40, 200] | 100 |

$M200_{TRF}$ in 91%, and $M400_{TRF}$ in 90% of all participant-level TRFs. In the theta band, a $M100_{TRF}$ was identified for every participant.

The amplitude of $M100_{TRF}$ was significantly higher in the target than in the distractor condition in the broadband ($z = 683$, $p < .001$, corrected for two comparisons) and the delta band ($z = 347$, $p < .001$, corrected for three comparisons). In the theta band, however, no significant differences between target and distractor response amplitude were found. The same applies to $M200_{TRF}$ in the broadband TRFs, where no considerable differences in target and distractor response amplitudes were discernible. In contrast, in the delta band, both $M200_{TRF}$ and $M400_{TRF}$ magnitudes were significantly greater in the target TRFs compared with the distractor TRFs ($M200_{TRF}$: $z = 1066$, $p < .001$; $M400_{TRF}$: $z = 699$, $p < .001$; corrected for three comparisons).

To further confirm the specificity of selective attention significantly modulating neural activity in the delta but not in the theta band, we conducted a paired $t$ test to compare the AI at the $M100_{TRF}$ between the theta and delta bands. The test revealed a significant difference in AI at the $M100_{TRF}$ between the theta and delta bands ($t = -5.30$, $p < .001$), indicating the theta band AI to be significantly lower than the delta band AI.

Furthermore, noticeable differences between target and distractor responses were visible not only in the TRF magnitudes but also in the time lags of the TRF peaks. Both $M100_{TRF}$ and $M200_{TRF}$ showed significant differences in delay, with the target response occurring later than the distractor response in the broadband ($M100_{TRF}$: $z = 596$, $p = .008$, $M200_{TRF}$: $z = 311$, $p < .001$, corrected for two comparisons), the delta band ($M100_{TRF}$: $z = 1002$, $p < .01$, $M200_{TRF}$: $z = 972$, $p < .001$, corrected for three comparisons), and also the theta band ($M100_{TRF}$: $z = 930$, $p < .001$). No significant differences in response peak time lags were found for $M400_{TRF}$. These differences were also mostly reflected in the mean TRF peak time lags shown in Figure 3A, where the target response peaks occurred with a delay of 10–40 compared with the distractor response peaks in all three frequency bands.

### Correlation between Neural Responses and Participant Behavior

To investigate whether the neural responses were related to behavior, we computed the Spearman correlation coefficient between the subject-level TRF magnitudes and the subject's percentage of correctly answered comprehension questions (Figure 3B). For both the target and the distractor listening mode, correlations were either close to zero, indicating that there was no relationship, or slightly positive ranging up to $r = .25$ ($M100_{TRF}$ for target audio input in the theta band). However, only the Spearman's correlation coefficients estimated for the theta band were significant, revealing similar positive correlations for both distractor ($r = .23$, $p = .04$, corrected for two comparisons) and target ($r = .25$, $p = .02$, corrected for two comparisons) modes regarding the $M100_{TRF}$ response.

The average comprehension score for Ms was 0.81, and the average score for NMs was 0.84. A statistical comparison revealed no significant difference between the two groups ($t = -0.95$, $p = .35$), indicating that both groups performed similarly on the task.

### Cortical Responses in Ms and NMs

To assess the putative influence of musical training on the cortical speech tracking, we compared the neural responses for the group of Ms to that of NMs (Figure 4). Similar to the population-level TRF magnitudes depicted in Figure 3A, for both groups, attention-induced differences in the TRF magnitudes are visible in both the broadband and in the delta band. At the same time, no considerable differences between distractor and target responses are discernible in the theta band.

For $M100_{TRF}$ peaks across all frequency bands and listening modes, the mean magnitude was smaller in Ms than

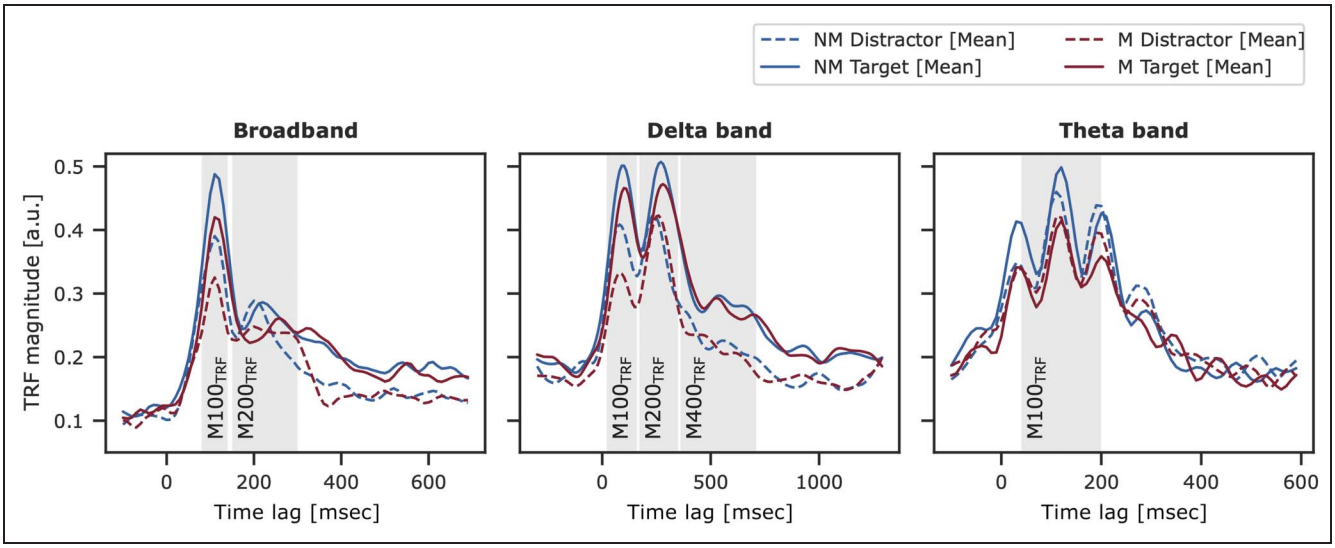


**Figure 4.** Magnitudes of the TRFs averaged across all sources for NMs (blue) and Ms (M, red) in the target condition (solid) and in the distractor case (dashed). The search areas for the subject-wise TRF peaks are shaded in gray. They were estimated based on the participants' average signal and are thus identical for M and NM and distractor and target responses. For all three frequency bands and both target and distractor responses, a tendency toward higher $M100_{TRF}$ peaks in NMs can be seen compared with the Ms.

in NMs. This difference was most pronounced for the distractor $M100_{TRF}$ response in the delta band with a mean increase of 23.0% in NMs compared with Ms. $M200_{TRF}$ is also slightly larger for NMs than Ms in both the target (9.6%) and distractor (16.5%) broadband TRF and the target delta-band TRF (7.4%). For the remaining TRF peaks, no noticeable differences emerged. A difference in latencies could be observed in the target $M200_{TRF}$ in the broadband response, as the Ms' average peaked 40 msec after the NMs at 260 msec. A similar effect was discernible in the delta-band $M200_{TRF}$ in the distractor condition, where the Ms' peak was delayed by 20 msec. Peak latencies were otherwise similar in both groups.

To assess the statistical significance of these differences, the subject-level TRF peaks and their time lags were compared between Ms and NMs. Furthermore, intragroup differences between distractor and target response magnitudes characterized by the *AI* were quantified. The results for all three frequency bands are depicted in Figure 5.

A tendency of increased $M100_{TRF}$ amplitude in NMs compared with Ms was still observable for both, target and distractor TRFs. However, these differences were only slightly significant for the distractor delta $M100_{TRF}$ ($t = -2.46$, $p = .049$, corrected for three comparisons). All other disparities were not significant. Regarding the latencies, only the one of the $M200_{TRF}$ response in the delta band in the distractor condition was found to be significantly longer in Ms than NMs ($t = 2.60$, $p = .03$, corrected for three comparisons). In turn, attentional modulation was highest at $M100_{TRF}$, with Ms tending to show larger *AI* values than NMs in the broadband and the delta-band response. This difference was most pronounced in the delta band yielding an average $AI = 0.27$ for Ms and $AI = 0.20$ for NMs. In addition, in the broadband $M200_{TRF}$, Ms showed a stronger representation of the target feature yielding $AI = 0.1$, on average, whereas NMs had a modulation value of $AI = -0.01$, suggesting no modulation of target and distractor cortical representations. However, none of these differences in the attentional modulation coefficient were statistically significant. Furthermore, in the theta band, the mean *AI* values of both groups were negligible ($AI = 0.01$).

## Relationship of Musicality Features and Cortical Responses

As a more fine-grained assessment of the influence of musical training on cortical speech tracking, we investigated whether there were relations between the participants' different aspects of musical training and the TRF peak amplitudes.

The estimated contributions of each of the aspects of musical training to predicting the neural responses are depicted in Figure 6. No considerable contributions were found for the weekly practice feature. Feature weights for the number of instruments indicate a positive contribution of this property to the peak amplitudes throughout all frequency bands and listening scenarios, but did not reach statistical significance. Only two features, namely, the starting age of learning an instrument and the playing duration in years, were significant predictors.

The former significantly contributed to the target $M200_{TRF}$ in the delta band ($p = .01$, corrected for 24 comparisons), with an assigned feature weight of 0.54, whereas the latter exhibited a negative relation with the TRF peak amplitudes. This observation became significant for the distractor $M100_{TRF}$ in the delta band, where the smallest model coefficient was estimated ($-0.55$, $p = .04$, corrected for 24 comparisons).

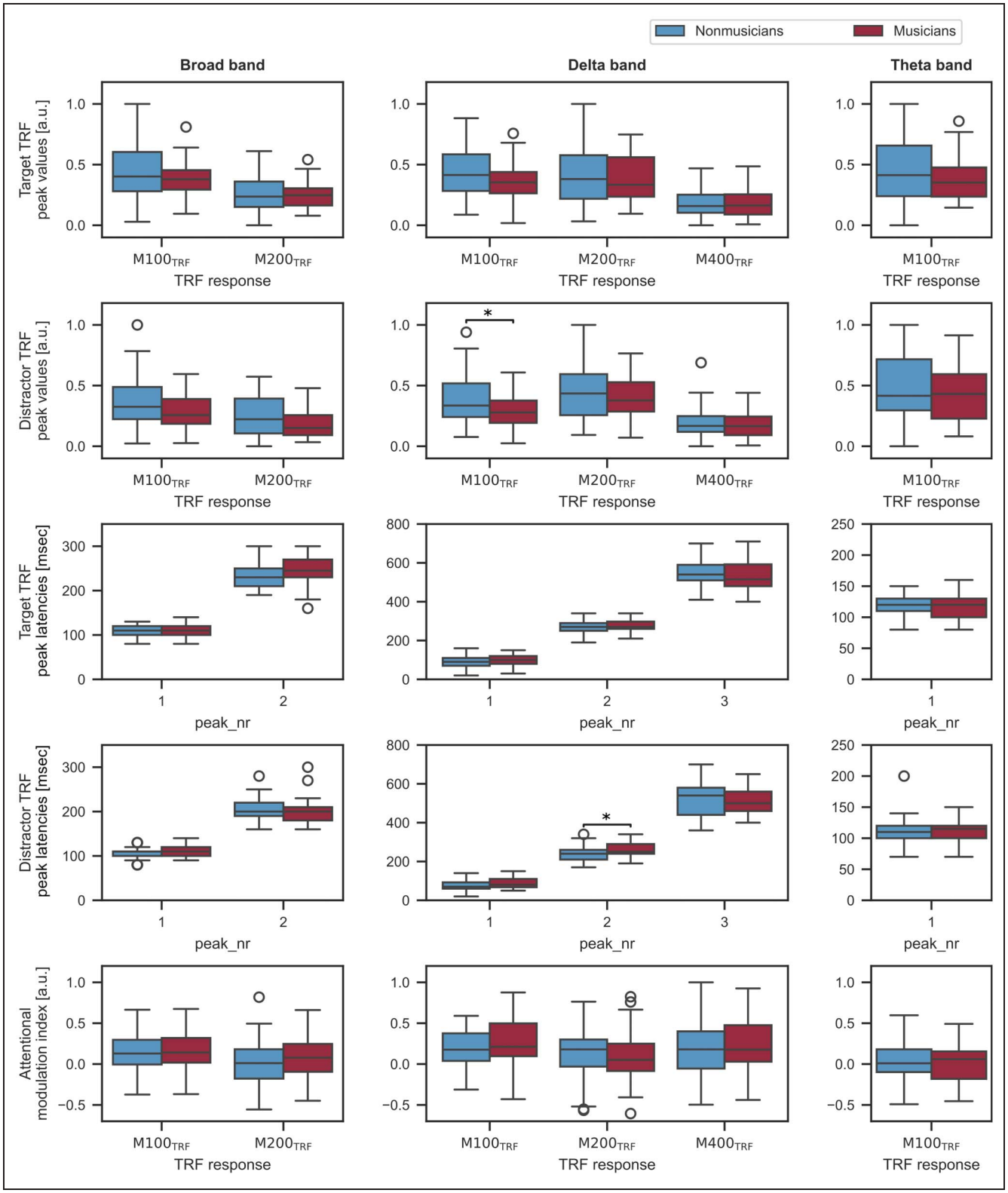


**Figure 5.** Comparison of subject-level TRF peak amplitudes (upper rows), latencies (middle rows), and attentional modulation indices (lower row) between Ms and NMs. Compared with Ms, NMs showed slight tendencies toward higher peak values in target and distractor $M100_{TRF}$ and $M200_{TRF}$, but only the distractor delta band $M100_{TRF}$ is slightly significant (upper two rows). Ms had a significantly longer latency for the $M200_{TRF}$ in the distractor delta band compared with NMs (fourth row). In turn, Ms show marginally larger attentional modulation indices for these response types compared with NMs (lower row). Peaks with significant differences in amplitude between the target and distractor condition are marked with asterisks indicating their significance level (*$0.01 \leq p < .05$).

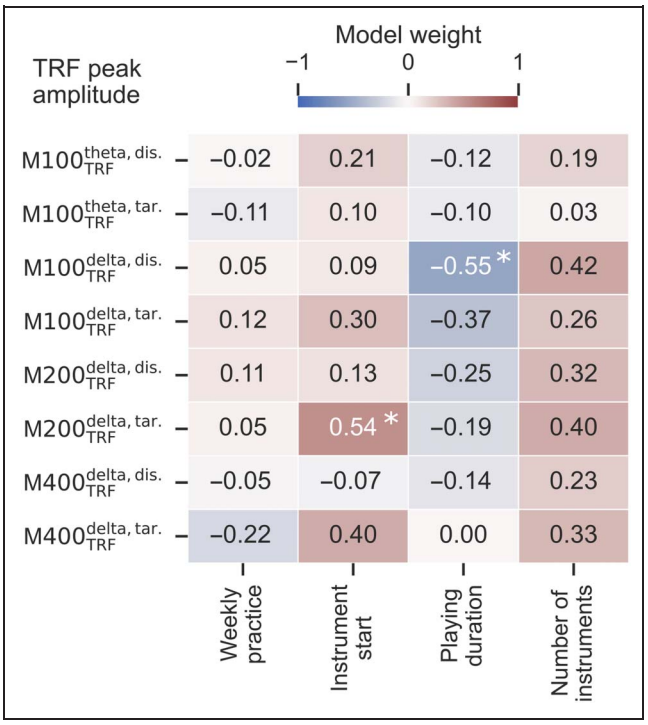


**Figure 6.** Influence of the individual aspects of musical training on the cortical responses. For each response type, frequency band, and attention mode, a multiple linear regression model was fitted to predict the response amplitude from the four aspects of musical training (hours of current weekly practice, starting age of playing an instrument, total duration of playing instruments, and the number of played instruments). The resulting weights of the linear model are displayed and assessed for statistical significance (*$0.01 \leq p < .05$). Two correlated aspects of musical training were found to be significant, the starting age and the duration of playing. Whereas the former is positively contributing to the $M200_{TRF}^{delta,tar.}$ response to the target signal (0.54), the latter is negatively affecting the $M100_{TRF}^{delta,dis.}$ response in the delta band to the distractor speaker (−.55).

## DISCUSSION

In this work, we utilized a large data set of MEG recordings from 52 participants, each of whom listened to 37 min of continuous speech. We thereby found that the attentional modulation of the cortical speech tracking occurred in the delta band but not in the theta band. We further found that musical training had only a slight influence on the neural responses, with most aspects of the latter remaining unaffected by musicianship.

### Different Types of Cortical Responses Masked in Broadband Become Apparent in the Delta and Theta Band

The speech envelope is a basic auditory feature that conveys various acoustic and linguistic information (Gillis et al., 2022; Brodbeck & Simon, 2020). The low-frequency range, below 20, is of particular interest, as it contains the rates of words and syllables.

Previous work has not only shown strong cortical envelope tracking in this frequency range (Ding & Simon, 2013) but also hypothesized that distinct functions can be attributed to delta- and theta-band tracking (Mai & Wang, 2023). In particular, theta-band tracking is assumed to reflect lower level syllable processing (Etard & Reichenbach, 2019; Di Liberto et al., 2015; Hyafil, Fontolan, Kabdebon, Gutkin, & Giraud, 2015), whereas higher level word processing is associated with delta-band responses (Etard & Reichenbach, 2019; Brodbeck, Hong, & Simon, 2018; Broderick, Anderson, Di Liberto, Crosse, & Lalor, 2018).

Our results show noticeable differences between the cortical responses in the delta and theta bands, which remain obscured in the broadband because of the superposition of the individual components. Whereas in the theta band the most prominent response occurs around 100 msec, which is also visible in the delta band and broadband, additional response peaks are present in the delta band around 250 msec and 500 msec. Although the latter is considerably weaker than the other response peaks, it is still clearly recognizable. In turn, in the broadband signal, the $M200_{TRF}$ is discernible but has a much lower amplitude than in the delta band. No later response peak emerges in the broadband response.

### Low-frequency Speech Envelope Tracking Is Mainly Right-Lateralized

To obtain a better understanding of the underlying mechanisms of the observed cortical responses, we also looked at their spatial distribution and lateralization. Except for the late $M400_{TRF}$, all cortical responses were significantly right-lateralized, with the highest activations centered in Heschl's gyrus, that is, primary auditory cortex. This is in line with previous findings that also found a right-lateralization of low-frequency envelope tracking (Assaneo et al., 2019). Furthermore, bottom–up processes on the sublexical level are dominated by the right hemisphere (Brodbeck et al., 2022; Luo & Poeppel, 2007). For top–down semantic processing starting from the word level, however, a lateralization shift toward the left hemisphere has been observed in previous studies, which might cause the lack of lateralization of the $M400_{TRF}$ observed here (Brodbeck, Hong, et al., 2018; Brodbeck, Presacco, et al., 2018).

### Delta and Theta Band Responses Are Modulated Differently by Selective Attention

It is well researched that in a competing speaker scenario, the envelopes of both speech streams are tracked individually in auditory cortex, with the cortical response to the distractor speaker being weaker than to the target speaker (Ding & Simon, 2012). We recover this behavior in the broadband for the $M100_{TRF}$ but not for the $M200_{TRF}$. Because the broadband response contains both the delta

and the theta bands, the lack of attentional modulation of its $M200_{TRF}$ peak is presumably because of the conflation of these two individual frequency bands.

Therefore, we focused on the separate assessment of the cortical tracking in the delta and theta bands. In the delta band, target speech leads to significantly stronger tracking than distractor speech, across a large range of delays and encompassing the $M100_{TRF}$, $M200_{TRF}$, and $M400_{TRF}$. In contrast, the theta band did not exhibit attentional modulation. This finding is also visible in the TRF model performance scores of the target and distractor models in all three frequency bands. The significant difference in prediction scores between target and distractor conditions for the broadband and delta band suggests that these frequency bands are more sensitive to attentional modulation. The higher prediction accuracy for target speech in these bands aligns with the notion that more neural resources are devoted to the processing of target stimuli, reflecting enhanced neural tracking of attended speech. In contrast, the absence of a significant difference in prediction scores between target and distractor conditions in the theta band suggests that neural tracking in this frequency range is not influenced or influenced only a little by attention. The significant difference in the AI between the delta and the theta frequency bands furthermore suggests that the attentional modulation is particularly robust in the delta band.

However, we found a significant positive relationship between the theta band's $M100_{TRF}$ and the behavioral performance of the participants. This finding suggests that the theta activity may play a key role in auditory stream segregation and aligns with previous research indicating the importance of theta band tracking for speech-in-noise intelligibility (Ding & Simon, 2014). However, because both the theta-band tracking of the attended and of the ignored speech stream correlate positively with speech comprehension, the theta activity may be involved in processing both auditory objects rather than selectively enhancing the target stream. This is in line with our finding that the theta-band tracking is not modulated by selective attention. Theta activity could hence be responsible for certain aspects of speech processing, such as syllable parsing, of both speech streams, whereas further cognitive processes, such as selective attention and task-specific neural modulation, may be required for selective attention to the target stream.

Whether the cortical tracking of speech rhythms in the delta and theta bands reflects evoked neural activity or entrainment of ongoing cortical oscillations remains an ongoing debate (Oganian et al., 2023; Van Bree, Sohoglu, Davis, & Zoefel, 2021; Ding & Simon, 2014; Giraud & Poeppel, 2012). Our finding that an $M100_{TRF}$ emerges both in the delta- and in the theta-band tracking, but that only the delta-band response is affected by attention, suggests that this neural response may not, or at least not only, be evoked by acoustic activity like an ERP. In fact, already the earliest evoked potentials in the cortex are affected by attention (Poghosyan & Ioannides, 2008). Although it remains possible that the theta-band $M100_{TRF}$ is modulated by attention, the effect would need to be much smaller than that of the delta-band response so that we were not able to observe it in our data. Our results therefore lead us to the hypothesis that delta- and theta-band tracking of the envelope may be continuously executed in parallel, whereby theta tracking provides bottom–up information about the unfiltered auditory input, whereas in the delta band, background information is attenuated and further processed for linguistic information.

In our study, we observed a latency difference between the TRFs for target and distractor speech, with the distractor stream showing faster processing than the target stream. This could indicate that this speech stream is processed less thoroughly than the target one, causing an earlier response. However, this finding contrasts with previous studies, which typically do not report faster processing for distractor speech (O'Sullivan et al., 2019; Puschmann et al., 2019). In contrast, one previous study reported that the peak of the neural response to the distractor speech occurred around 10 msec after that to the attended speaker (Ding & Simon, 2012).

This discrepancy may be attributed to differences in experimental design and task difficulty. For example, Ding and Simon (Ding & Simon, 2012) and O'Sullivan and colleagues (O'Sullivan et al., 2019) used shorter speech stimuli (1 min and 30 sec, respectively), whereas our study used longer audio book chapters (3–5 min). The longer stimuli likely placed greater demands on working memory for the target stream, requiring more cognitive resources for maintaining and integrating information. In contrast, the distractor stream may rely more on automatic, bottom–up processing, leading to faster neural responses. In addition, the demanding posttrial tasks in our study, where participants answered three challenging multiple-choice questions, likely further increased the cognitive load for the target stream, potentially delaying its response. This contrasts with studies like Puschmann and colleagues (Puschmann et al., 2019), where participants freely recapped the target speech, and O'Sullivan and colleagues (O'Sullivan et al., 2019), where participants only needed to repeat the last sentence. These factors suggest that latency differences in neural processing may be sensitive to the specific conditions of the experiment, and further research is needed to clarify these effects.

### Few Significant Differences between Ms and NMs

When assessing the influence of musical training on cortical speech tracking, we only found minor significant effects on the $M100_{TRF}$ and $M200_{TRF}$ in the delta band. For the distractor speech signal, the delta band $M100_{TRF}$ peak was slightly greater in NMs than in Ms. Moreover, its magnitude was negatively impacted by the duration over which participants had played an instrument. Furthermore, the distractor $M200_{TRF}$ peak was slightly

delayed in Ms than in NMs. In addition, its magnitude resulting from the target speech signal had a positive correlation with the starting age of practicing an instrument. The other components of the neural speech tracking were not significantly affected by musicianship, although we observed a general trend of lower amplitudes for Ms as compared with NMs.

To interpret these findings, we note that, in contrast to intrasubject modulation effects, where stronger tracking of a speech signal in background noise is indicative of better intelligibility, this does not necessarily apply when comparing cortical responses between participants. Previous studies investigating speech perception in the elderly in fact demonstrated that stronger cortical speech tracking correlated with lower intelligibility scores (Presacco, Simon, & Anderson, 2016; Karunathilake et al., 2023). The strength of the neural tracking might thus not indicate improved comprehension, but rather increased listening effort.

In line with these deliberations, Jantzen and colleagues (2014) observed that NMs showed increased cortical activity, especially for early responses (Jantzen et al., 2014). This finding matches with the tendencies toward higher response magnitudes in NMs, which we observed here (despite a lack of statistical significance). At the same time, this might explain the slightly higher attentional modulation indices for early delta responses in Ms.

This "inverse" relationship between musicianship and the strength of the cortical tracking is also reflected in the relation between the aspects of musical training and the neural responses that we determined. The earlier the participants started playing an instrument, the weaker the delta-band responses were. This aspect of musical training hence seems to have a strong influence on speech-in-noise perception, which is consistent with the results of other studies (Puschmann et al., 2021; Kraus & Chandrasekaran, 2010).

Our analysis revealed only minimal differences between Ms and NMs in the behavioral task. However, our study was not specifically designed to detect behavioral differences between these groups. The primary goal of the comprehension questions was to assure that the participants did attend the target speaker rather than revealing behavioral differences. As a result, the task employed in this study may not have been challenging enough to elicit detectable differences in behavioral performance, as both groups achieved similar comprehension scores. This may also explain the limited differences in cortical tracking observed between Ms and NMs. Previous studies, such as Puschmann and colleagues (Puschmann et al., 2019), used more complex paradigms or more demanding tasks, potentially leading to stronger differences between groups.

We expect the $M400_{TRF}$ to be most closely associated with semantic integration and contextualization of the speech signals. This response was remarkably similar between Ms and NMs. Together with the insignificance of most of the other aspects of the neural speech tracking that we examined, we conclude that cortical speech tracking is essentially not affected by musical training. The behavioral improvements in speech-in-noise perception observed in Ms appear to originate in other neural mechanisms, probably further downstream of the cortical speech tracking because the latter already involves early auditory responses.

## Conclusions

In summary, our study showed that the attentional modulation of cortical speech tracking results from the delta band but not from the theta band. The theta band presumably reflects lower level acoustic processing such as syllable parsing, which appears to be equally executed for the target and the distractor speech stream. Neural activity in the delta band, in contrast, is responsible for higher level linguistic processing, and our results show that attentional effects emerge at this stage. Musical training leaves both types of responses largely unchanged, although we observed a tendency toward stronger cortical speech tracking in people with less musical training.

Corresponding author: Tobias Reichenbach, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Werner-von-Siemens-Straße 61, Erlangen, 91052, Germany, e-mail: tobias.j.reichenbach@fau.de.

### Data Availability Statement

The MEG data are available at zenodo.org (https://zenodo.org/records/12793944).

### Author Contributions

Alina Schüller: Conceptualization; Data curation; Formal analysis; Investigation; Writing—Original draft. Annika Mücke: Conceptualization; Data curation; Formal analysis; Investigation; Writing—Original draft. Jasmin Riegel: Conceptualization; Funding acquisition, Writing—Review & editing. Tobias Reichenbach: Conceptualization; Data curation; Supervision; Writing—Review & editing.

### Funding Information

This project was supported by the German Federal Ministry of Education and Research (Cluster4Future SEMECO), grant number: 03ZU1210FB, and the German Science Foundation (Deutsche Forschungsgemeinschaft), grant number: 523344822.

### Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent

pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance.

## REFERENCES


Assaneo, M. F., Rimmele, J. M., Orpella, J., Ripollés, P., de Diego-Balaguer, R., & Poeppel, D. (2019). The lateralization of speech-brain coupling is differentially modulated by intrinsic auditory and top–down mechanisms. *Frontiers in Integrative Neuroscience*, *13*, 28. https://doi.org/10.3389/fnint.2019.00028, PubMed: 31379527

Biesmans, W., Das, N., Francart, T., & Bertrand, A. (2017). Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*, 402–412. https://doi.org/10.1109/TNSRE.2016.2571900, PubMed: 27244743

Brodbeck, C., Bhattasali, S., Cruz Heredia, A. A., Resnik, P., Simon, J. Z., & Lau, E. (2022). Parallel processing in speech perception with local and global representations of linguistic context. *eLife*, *11*, e72056. https://doi.org/10.7554/eLife.72056, PubMed: 35060904

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, *28*, 3976–3983. https://doi.org/10.1016/j.cub.2018.10.042, PubMed: 30503620

Brodbeck, C., Jiao, A., Hong, L. E., & Simon, J. Z. (2020). Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers. *PLoS Biology*, *18*, e3000883. https://doi.org/10.1371/journal.pbio.3000883, PubMed: 33091003

Brodbeck, C., Presacco, A., & Simon, J. Z. (2018). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *Neuroimage*, *172*, 162–174. https://doi.org/10.1016/j.neuroimage.2018.01.042, PubMed: 29366698

Brodbeck, C., & Simon, J. Z. (2020). Continuous speech processing. *Current Opinion in Physiology*, *18*, 25–31. https://doi.org/10.1016/j.cophys.2020.07.014, PubMed: 33225119

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*, 803–809. https://doi.org/10.1016/j.cub.2018.01.080, PubMed: 29478856

Chen, Y.-P., Schmidt, F., Keitel, A., Rösch, S., Hauswald, A., & Weisz, N. (2023). Speech intelligibility changes the temporal evolution of neural speech tracking. *Neuroimage*, *268*, 119894. https://doi.org/10.1016/j.neuroimage.2023.119894, PubMed: 36693596

Clayton, K. K., Swaminathan, J., Yazdanbakhsh, A., Zuk, J., Patel, A. D., & Kidd, G., Jr. (2016). Executive function, visual attention and the cocktail party problem in musicians and non-musicians. *PLoS One*, *11*, e0157638. https://doi.org/10.1371/journal.pone.0157638, PubMed: 27384330

Coffey, E. B., Mogilever, N. B., & Zatorre, R. J. (2017). Speech-in-noise perception in musicians: A review. *Hearing Research*, *352*, 49–69. https://doi.org/10.1016/j.heares.2017.02.006, PubMed: 28213134

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in Human Neuroscience*, *10*, 604. https://doi.org/10.3389/fnhum.2016.00604, PubMed: 27965557

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*, 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030, PubMed: 26412129

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences, U.S.A.*, *109*, 11854–11859. https://doi.org/10.1073/pnas.1205381109, PubMed: 22753470

Ding, N., & Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience*, *33*, 5728. https://doi.org/10.1523/JNEUROSCI.5297-12.2013, PubMed: 23536086

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in Human Neuroscience*, *8*, 311. https://doi.org/10.3389/fnhum.2014.00311, PubMed: 24904354

Du, Y., & Zatorre, R. J. (2017). Musical training sharpens and bonds ears and tongue to hear speech better. *Proceedings of the National Academy of Sciences, U.S.A.*, *114*, 13579–13584. https://doi.org/10.1073/pnas.1712223114, PubMed: 29203648

Etard, O., Kegler, M., Braiman, C., Forte, A. E., & Reichenbach, T. (2019). Decoding of selective attention to continuous speech from the human auditory brainstem response. *Neuroimage*, *200*, 1–11. https://doi.org/10.1016/j.neuroimage.2019.06.029, PubMed: 31212098

Etard, O., & Reichenbach, T. (2019). Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *Journal of Neuroscience*, *39*, 5750–5759. https://doi.org/10.1523/JNEUROSCI.1828-18.2019, PubMed: 31109963

Fischl, B. (2012). FreeSurfer. *Neuroimage*, *62*, 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021, PubMed: 22248573

Gillis, M., Van Canneyt, J., Francart, T., & Vanthornhout, J. (2022). Neural tracking as a diagnostic tool to assess the auditory pathway. *Hearing Research*, *426*, 108607. https://doi.org/10.1016/j.heares.2022.108607, PubMed: 36137861

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*, 511–517. https://doi.org/10.1038/nn.3063, PubMed: 22426255

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*, 1–13. https://doi.org/10.3389/fnins.2013.00267, PubMed: 24431986

Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., & Giraud, A.-L. (2015). Speech encoding by coupled cortical theta and gamma oscillations. *eLife*, *4*, e06213. https://doi.org/10.7554/eLife.06213, PubMed: 26023831

Jantzen, M., Howe, B., & Jantzen, K. (2014). Neurophysiological evidence that musical training influences the recruitment of right hemispheric homologues for speech perception. *Frontiers in Psychology*, *5*, 171. https://doi.org/10.3389/fpsyg.2014.00171, PubMed: 24624107

Kadir, S., Kaza, C., Weissbart, H., & Reichenbach, T. (2019). Modulation of speech-in-noise comprehension through transcranial current stimulation with the phase-shifted speech envelope. *IEEE Transactions on Neural Systems and*

*Rehabilitation Engineering*, *28*, 23–31. https://doi.org/10.1109/TNSRE.2019.2939671, PubMed: 31751277
Karunathilake, I. M. D., Dunlap, J. L., Perera, J., Presacco, A., Decruy, L., Anderson, S., et al. (2023). Effects of aging on cortical representations of continuous speech. *Journal of Neurophysiology*, *129*, 1359–1377. https://doi.org/10.1152/jn.00356.2022, PubMed: 37096924
Kegler, M., Weissbart, H., & Reichenbach, T. (2022). The neural response at the fundamental frequency of speech is modulated by word-level acoustic and linguistic information. *Frontiers in Neuroscience*, *16*, 915744. https://doi.org/10.3389/fnins.2022.915744, PubMed: 35942153
Keshavarzi, M., Kegler, M., Kadir, S., & Reichenbach, T. (2020). Transcranial alternating current stimulation in the theta band but not in the delta band modulates the comprehension of naturalistic speech in noise. *Neuroimage*, *210*, 116557. https://doi.org/10.1016/j.neuroimage.2020.116557, PubMed: 31968233
Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience*, *11*, 599–605. https://doi.org/10.1038/nrn2882, PubMed: 20648064
Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123, PubMed: 20809790
Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, *54*, 1001–1010. https://doi.org/10.1016/j.neuron.2007.06.004, PubMed: 17582338
Mai, G., & Wang, W. S.-Y. (2023). Distinct roles of delta- and theta-band neural tracking for sharpening and predictive coding of multi-level speech features during spoken language processing. *Human Brain Mapping*, *44*, 6149–6172. https://doi.org/10.1002/hbm.26503, PubMed: 37818940
Musacchia, G., Sams, M., Skoe, E., & Kraus, N. (2007). Musicians have enhanced subcortical auditory and audiovisual processing of speech and music. *Proceedings of the National Academy of Sciences, U.S.A.*, *104*, 15894–15898. https://doi.org/10.1073/pnas.0701498104, PubMed: 17898180
O'Sullivan, J., Herrero, J., Smith, E., Schevon, C., McKhann, G. M., Sheth, S. A., et al. (2019). Hierarchical encoding of attended auditory objects in multi-talker speech perception. *Neuron*, *104*, 1195–1209. https://doi.org/10.1016/j.neuron.2019.09.007, PubMed: 31648900
O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., et al. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, *25*, 1697–1706. https://doi.org/10.1093/cercor/bht355, PubMed: 24429136
Oganian, Y., Kojima, K., Breska, A., Cai, C., Findlay, A., Chang, E. F., et al. (2023). Phase alignment of low-frequency neural activity to the amplitude envelope of speech reflects evoked responses to acoustic edges, not oscillatory entrainment. *Journal of Neuroscience*, *43*, 3909–3921. https://doi.org/10.1523/JNEUROSCI.1663-22.2023, PubMed: 37185238
Parbery-Clark, A., Anderson, S., Hittner, E., & Kraus, N. (2012). Musical experience strengthens the neural representation of sounds important for communication in middle-aged adults. *Frontiers in Aging Neuroscience*, *4*, 30. https://doi.org/10.3389/fnagi.2012.00030, PubMed: 23189051
Parbery-Clark, A., Skoe, E., & Kraus, N. (2009). Musical experience limits the degradative effects of background noise on the neural processing of sound. *Journal of Neuroscience*, *29*, 14100–14107. https://doi.org/10.1523/JNEUROSCI.3256-09.2009, PubMed: 19906958
Parbery-Clark, A., Strait, D. L., Anderson, S., Hittner, E., & Kraus, N. (2011). Musical experience and the aging auditory system: Implications for cognitive abilities and hearing speech in noise. *PLoS One*, *6*, e18082. https://doi.org/10.1371/journal.pone.0018082, PubMed: 21589653
Parbery-Clark, A., Tierney, A., Strait, D. L., & Kraus, N. (2012). Musicians have fine-tuned neural distinction of speech syllables. *Neuroscience*, *219*, 111–119. https://doi.org/10.1016/j.neuroscience.2012.05.042, PubMed: 22634507
Poghosyan, V., & Ioannides, A. A. (2008). Attention modulates earliest responses in the primary auditory and visual cortices. *Neuron*, *58*, 802–813. https://doi.org/10.1016/j.neuron.2008.04.013, PubMed: 18549790
Presacco, A., Simon, J. Z., & Anderson, S. (2016). Evidence of degraded representation of speech in noise, in the aging midbrain and cortex. *Journal of Neurophysiology*, *116*, 2346–2355. https://doi.org/10.1152/jn.00372.2016, PubMed: 27535374
Puschmann, S., Baillet, S., & Zatorre, R. J. (2019). Musicians at the cocktail party: Neural substrates of musical training during selective listening in multispeaker situations. *Cerebral Cortex*, *29*, 3253–3265. https://doi.org/10.1093/cercor/bhy193, PubMed: 30137239
Puschmann, S., Regev, M., Baillet, S., & Zatorre, R. J. (2021). MEG intersubject phase locking of stimulus-driven activity during naturalistic speech listening correlates with musical training. *Journal of Neuroscience*, *41*, 2713–2722. https://doi.org/10.1523/JNEUROSCI.0932-20.2020, PubMed: 33536196
Riecke, L., Formisano, E., Sorger, B., Başkent, D., & Gaudrain, E. (2018). Neural entrainment to speech modulates speech intelligibility. *Current Biology*, *28*, 161–169. https://doi.org/10.1016/j.cub.2017.11.033, PubMed: 29290557
Riegel, J., Schüller, A., & Reichenbach, T. (2024). No evidence of musical training influencing the cortical contribution to the speech-FFR and its modulation through selective attention. *eNeuro*, *11*. ENEURO.0127-24.2024. https://doi.org/10.1523/ENEURO.0127-24.2024, PubMed: 39160069
Schilling, A., Tomasello, R., Henningsen-Schomers, M. R., Zankl, A., Surendra, K., Haller, M., et al. (2021). Analysis of continuous neuronal activity evoked by natural speech with computational corpus linguistics methods. *Language, Cognition and Neuroscience*, *36*, 167–186. https://doi.org/10.1080/23273798.2020.1803375
Schüller, A., Schilling, A., Krauss, P., Rampp, S., & Reichenbach, T. (2023). Attentional modulation of the cortical contribution to the frequency-following response evoked by continuous speech. *Journal of Neuroscience*, *43*, 7429–7440. https://doi.org/10.1523/JNEUROSCI.1247-23.2023, PubMed: 37793908
Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*. https://doi.org/10.25080/Majora-92bf1922-011
Seghier, M. L. (2008). Laterality index in functional MRI: Methodological issues. *Magnetic Resonance Imaging*, *26*, 594. https://doi.org/10.1016/j.mri.2007.10.010, PubMed: 18158224
Souza, P. E., Boike, K. T., Witherell, K., & Tremblay, K. (2007). Prediction of speech recognition from audibility in older listeners with hearing loss: Effects of age, amplification, and background noise. *Journal of the American Academy of Audiology*, *18*, 54–65. https://doi.org/10.3766/jaaa.18.1.5, PubMed: 17252958
Strait, D., & Kraus, N. (2011). Can you hear me now? Musical training shapes functional brain networks for selective auditory attention and hearing speech in noise. *Frontiers in Psychology*, *2*, 113. https://doi.org/10.3389/fpsyg.2011.00113, PubMed: 21716636

Van Bree, S., Sohoglu, E., Davis, M. H., & Zoefel, B. (2021). Sustained neural rhythms reveal endogenous oscillations supporting speech perception. *PLoS Biology*, *19*, e3001142. https://doi.org/10.1371/journal.pbio.3001142, PubMed: 33635855

Van Hirtum, T., Somers, B., Verschueren, E., Dieudonné, B., & Francart, T. (2023). Delta-band neural envelope tracking predicts speech intelligibility in noise in preschoolers. *Hearing Research*, *434*, 108785. https://doi.org/10.1016/j.heares.2023.108785

Van Veen, B., Van Drongelen, W., Yuchtman, M., & Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on Biomedical Engineering*, *44*, 867–880. https://doi.org/10.1109/10.623056, PubMed: 9282479

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0 contributors, Scipy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2, PubMed: 32015543

Wilsch, A., Neuling, T., Obleser, J., & Herrmann, C. S. (2018). Transcranial alternating current stimulation with speech envelopes modulates speech comprehension. *Neuroimage*, *172*, 766–774. https://doi.org/10.1016/j.neuroimage.2018.01.038, PubMed: 29355765

Zendel, B. R., & Alain, C. (2009). Concurrent sound segregation is enhanced in musicians. *Journal of Cognitive Neuroscience*, *21*, 1488–1498. https://doi.org/10.1162/jocn.2009.21140, PubMed: 18823227

Zendel, B. R., Tremblay, C.-D., Belleville, S., & Peretz, I. (2015). The impact of musicianship on the cortical mechanisms related to separating speech from background noise. *Journal of Cognitive Neuroscience*, *27*, 1044–1059. https://doi.org/10.1162/jocn_a_00758, PubMed: 25390195

Zoefel, B., Archer-Boyd, A., & Davis, M. H. (2018). Phase entrainment of brain oscillations causally modulates neural responses to intelligible speech. *Current Biology*, *28*, 401–408. https://doi.org/10.1016/j.cub.2017.11.071, PubMed: 29358073

Zuk, J., Ozernov-Palchik, O., Kim, H., Lakshminarayanan, K., Gabrieli, J. D. E., Tallal, P., et al. (2013). Enhanced syllable discrimination thresholds in musicians. *PLoS One*, *8*, e80546. https://doi.org/10.1371/journal.pone.0080546, PubMed: 24339875