# Cortical Responses Time-Locked to Continuous Speech in the High-Gamma Band Depend on Selective Attention

**Vrishab Commuri** [1], **Joshua P. Kulasingham** [2] **and Jonathan Z. Simon** [1,3,4,*]

[1] *Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, United States*
[2] *Department of Electrical Engineering, Linköping University, Linköping, Sweden*
[3] *Department of Biology, University of Maryland, College Park, MD, United States*
[4] *Institute for Systems Research, University of Maryland, College Park, MD, United States*

Correspondence*:
Jonathan Z. Simon
jzsimon@umd.edu

**ABSTRACT**

Auditory cortical responses to speech obtained by magnetoencephalography (MEG) show robust speech tracking in the high-gamma band (70-200 Hz), but little is currently known about whether such responses depend on the focus of selective attention. In this study 22 human subjects listened to concurrent, fixed-rate, speech from male and female speakers, and were asked to selectively attend to one speaker at a time, while their neural responses were recorded with MEG. The male speaker's pitch range coincided with the lower range of the high-gamma band, whereas the female speaker's higher pitch range had much less overlap, and only at the upper end of high-gamma band. Neural responses were analyzed using the temporal response function (TRF) framework. As expected, the responses demonstrate robust speech tracking in the high gamma band, but only to the male's speech, with a peak latency of approximately 40 ms. The response magnitude also depends on selective attention: the response to the male speaker is significantly greater when male speech is attended than when it is not attended. This is a clear demonstration that even very early cortical auditory responses are influenced by top-down, cognitive, neural processing mechanisms.

**Keywords: cortical FFR, cocktail party, speech tracking, primary auditory cortex, phase-locked response**

## 1 INTRODUCTION

Time-locked auditory responses are one mechanism by which the auditory system preserves temporal information about sounds. For example, subcortical responses to voiced sections of speech time-lock to the speaker's fundamental frequency (F0), whether $\gtrsim$100 Hz for a typical male voice (Skoe and Kraus, 2010) or $\gtrsim$200 Hz for a typical female voice (Lehmann and Schönwiesner, 2014), and can be measured via the frequency following response (FFR)(Kraus et al., 2017). As neural responses propagate up the auditory pathway, characteristic time-locking frequencies are generally observed to decrease. For example, cortical responses time-lock to the envelope of the speech most strongly below $\sim$10 Hz

(Ahissar et al., 2001; Luo and Poeppel, 2007). Nevertheless, recent FFR studies have observed cortical time-locked responses at rates often associated with subcortical processing, $\gtrsim$100 Hz, using responses measured from magnetoencephalography (MEG) (Coffey et al., 2016; Gorina-Careta et al., 2021), and electroencephalography (EEG) (Bidelman, 2018).

The FFR is obtained from the average of many (e.g., thousands of) responses to a repeated auditory stimulus, and has been used to provide insight into the representation of speech in the auditory periphery and the fidelity of sound encoding in the brain (Basu et al., 2010; Kraus et al., 2017). Modulations of the FFR strength and consistency can be used to study cognitive processes such as learning (Skoe et al., 2013), selective attention (Lehmann and Schönwiesner, 2014; Holmes et al., 2017), level of attention (Price and Bidelman, 2021), intermodal (auditory vs. visual) attention (Hartmann and Weisz, 2019), and the effect of familiar vs. unfamiliar background language (Presacco et al., 2016; Zan et al., 2019). These studies demonstrate that FFRs can be affected by top-down auditory processes, though it is not clear how much of the FFR modulation is due to subcortical vs. cortical sources (Gnanateja et al., 2021; Gorina-Careta et al., 2021). In the present study, listeners heard two concurrent speakers (both speaking the listeners' native language), with the task of attending to one speaker and ignoring the other, in order to directly investigate the the role of auditory cortex in selective attention.

The FFR, in order to be averaged over so many trials, uses many repetitions of a short stimulus (e.g., a single speech syllable). In contrast, temporal response functions (TRFs) characterize neuronal responses to speech using single long-duration trials of continuous speech (Lalor et al., 2009; Ding and Simon, 2012). While TRF analysis is most often applied to low frequency cortical responses (Brodbeck and Simon, 2020), TRFs obtained with MEG have recently also been used to investigate cortical responses to speech in the the high gamma range (70-200 Hz)(Kulasingham et al., 2020; Schüller et al., 2023a), i.e., for frequencies similar to those investigated using cortical FFR, showing a single response peak with latency $\sim$40 ms, indicating a focal neural origin in primary auditory cortex (see also Kegler et al. (2022) for EEG). The present study extends the work of Kulasingham et al. (2020) by applying high-gamma TRF analysis of MEG responses to subjects listening to speech from male and female speakers in single-speaker and "cocktail-party" (competing speaker) paradigms.

The single-speaker conditions used here allow comparison of subjects' responses to both male (F0 $\gtrsim$ 100 Hz) and female speech (F0 $\gtrsim$ 200 Hz) in isolation. Many recent studies on high-gamma cortical responses to speech only use stimuli from male speakers in their experimental design (Kulasingham et al., 2020; Canneyt et al., 2021a; Gnanateja et al., 2021; Guo et al., 2021; Kegler et al., 2022). This may be because typical male speakers have a lower F0 than typical female speakers, and stronger responses are evoked by speech with a lower F0. Indeed, Canneyt et al. (2021b) investigated responses to stimuli from both male and female speakers and observed that high-gamma cortical response strength was inversely related to F0.

The competing speakers conditions used here allow the investigation of how these fast cortical responses change depending on top-down influences such as task specificity and selective attention. In humans it is seen widely that auditory low frequency ($\lesssim$10 Hz) time-locked cortical responses depend on selective attention, whether for simple sounds (Hillyard et al., 1973; Elhilali et al., 2009; Holmes et al., 2017) or speech (Lalor et al., 2009; Ding and Simon, 2012). To what extent selective attention changes response properties in early latency primary auditory cortex, as opposed to secondary auditory areas and beyond, is not yet well understood. Using invasive intracranial EEG (iEEG) recordings, effects of selective attention has been observed for simple stimuli (Bidet-Caulet et al., 2007) but not for competing speakers (O'Sullivan et al., 2019). From MEG studies there is recent evidence for selective attention affecting the low frequency

response properties of very early auditory cortex during a competing speaker task (Brodbeck et al., 2020), but the effect is small and occurs only under limited conditions.

Thus, the main focus of the present study concerns two primary research questions. Firstly, what differences are there in high-gamma cortical responses between the cases of male (F0 $\gtrsim$ 100 Hz) vs female (F0 $\gtrsim$ 200 Hz) speech? Secondly, do early ( $\sim$40 ms latency) high-gamma cortical responses to speech, putatively arising only from primary auditory cortex (Simon et al., 2022), depend on selective attention? Both these questions are addressed by analyzing MEG recordings of subjects listening to single male and female voices, and to the same voices presented simultaneously but with the task of selectively attending to only one or the other.

## 2 MATERIALS AND METHODS

### 2.1 Data

The data set analyzed here was previously obtained and analyzed in an earlier study that investigated differing cortical responses between spoken language and arithmetic using two different speakers (Kulasingham et al., 2021). The data are available at https://doi.org/10.13016/xd2i-vyke and the code is available at https://github.com/vrishabcommuri/mathlang-highgamma.

### 2.2 Participants

The data set comprises MEG responses recorded from 22 individuals (average age 22.6 years, 10 female, 21 right handed) who were native English speakers. The participants provided written informed consent and received monetary compensation. The experimental procedure was approved by the Internal Review Board of the University of Maryland, College Park.

### 2.3 Data Acquisition and Preprocessing

The data were collected from subjects using a 157 axial gradiometer whole head KIT (Kanazawa Institute of Technology) MEG system with subjects resting in the supine position in a magnetically shielded room (Vacuumschmelze GmbH & Co. KG, Hanau, Germany). The data were recorded at a sampling rate of 1 kHz with an online 200 Hz low pass filter with a wide transition band above 200 Hz and a 60 Hz notch filter. Data were preprocessed in MATLAB by first automatically excluding saturating channels and then applying time-shift principal component analysis (TSPCA) (de Cheveigné and Simon, 2007) to remove external noise, and sensor noise suppression (SNS) (de Cheveigné and Simon, 2008) to suppress channel artifacts. Two of the sensor channels were excluded during the preprocessing stage.

The denoised MEG data were filtered from 70 to 200 Hz using an FIR bandpass filter with 5 Hz transition bands and were subsequently downsampled to 500 Hz. Independent component analysis (ICA) was then applied to remove artifacts such as heartbeats, head movements, and eye blinks.

The subsequent analyses were performed in Python using the mne-python (1.3.1) (Gramfort et al., 2013) and eelbrain (0.38.4) (Brodbeck et al., 2019) libraries, and in R using the lme4 (1.1-21) (Bates et al., 2015) and buildmer (2.8) (Voeten, 2023) package.

### 2.4 Neural Source Localization

Prior to the data collection, the head shape of each subject was digitized using the a Polhemus 3SPACE FASTRAK system, and subject head position in the MEG scanner was measured before and after the

experiment using five marker coils. The marker coil locations and the digitized head shape were used to co-register the template FreeSurfer "fsaverage" brain (Fischl, 2012) using rotation, translation and uniform scaling.

Source localization was performed using the mne-python software package. First, a volume source space was composed from a grid of 7-mm sized voxels. Then, an inverse operator was computed, mapping the sensor space to the source space using minimum norm estimation (MNE) (Hämäläinen and Ilmoniemi, 1994) and dynamic statistical parametric mapping (dSPM) Dale et al. (2000) with a depth weighting parameter of 0.8 and a noise covariance matrix estimated from empty room data. The result of the localization procedure was a single 3-dimensional current dipole centered within each voxel.

The Freesurfer 'aparc+aseg' parcellation was used to define a cortical region of interest (ROI). The ROI consisted of voxels in the gray and white matter of the brain that were closest to the temporal lobe – Freesurfer 'aparc' parcellations with labels 'transversetemporal', 'superiortemporal', 'inferiortemporal', and 'bankssts'. All analyses were constrained to this ROI to conserve computational resources.

## 2.5 Stimuli

Subjects listened to isochronous (fixed-rate) speech from two synthesized voices – one male and one female. Speech was generated using the ReadSpeaker synthesizer with the "James" and "Kate" voices (https://www.readspeaker.com). Two kinds of speech stimuli were created: 'language' stimuli that consisted of four-word sentences, and 'arithmetic' stimuli that consisted of five-word equations. The word rate of the arithmetic stimuli was faster than that of the sentence stimuli so that neural responses to each could be separated in the frequency domain and so that each stimulus was 18 seconds in duration. The stimulus files are available in the same repository as the data: https://doi.org/10.13016/xd2i-vyke.

## 2.6 Experimental Design

The experiment was divided into two conditions. In the first condition ('single-speaker'), subjects listened to speech from either the male or the female speaker; and in the second condition ('cocktail-party'), subjects listened to both speakers concurrently and were instructed to attend to only one. Each condition was conducted in blocks: four single speaker blocks ($2 \times 2$: male and female, sentences and equations) followed by eight cocktail party blocks. At the start of each cocktail party block, the subject was instructed as to which stimulus to attend to, and was asked to press a button at the end of each trial to indicate whether a deviant was detected. The subjects were generally able to attend to the instructed speaker (Kulasingham et al., 2020). The order in which blocks were presented was counterbalanced across subjects.

## 2.7 Stimulus Representations

In accordance with the methods of Kulasingham et al. (2020), two predictors (i.e., stimulus representations) were used, one capturing the high-frequency envelope and another capturing the stimulus carrier. Figure 1 illustrates the procedure for extracting both predictors from a stimulus waveform.

### 2.7.1 Carrier Predictor

To create the carrier predictor, each stimulus was first resampled to a frequency of 500 Hz to reduce the ensuing computation required. Prior to downsampling, an anti-aliasing FIR prefilter with 500 Hz cutoff and 5 Hz transition band was applied to the data. This resampled signal was then bandpass filtered in the high-gamma range of 70-200 Hz using an FIR filterwith 5 Hz transition band. Finally, the signal was standardized (i.e., mean subtracted and normalized by the standard deviation) to produce the carrier

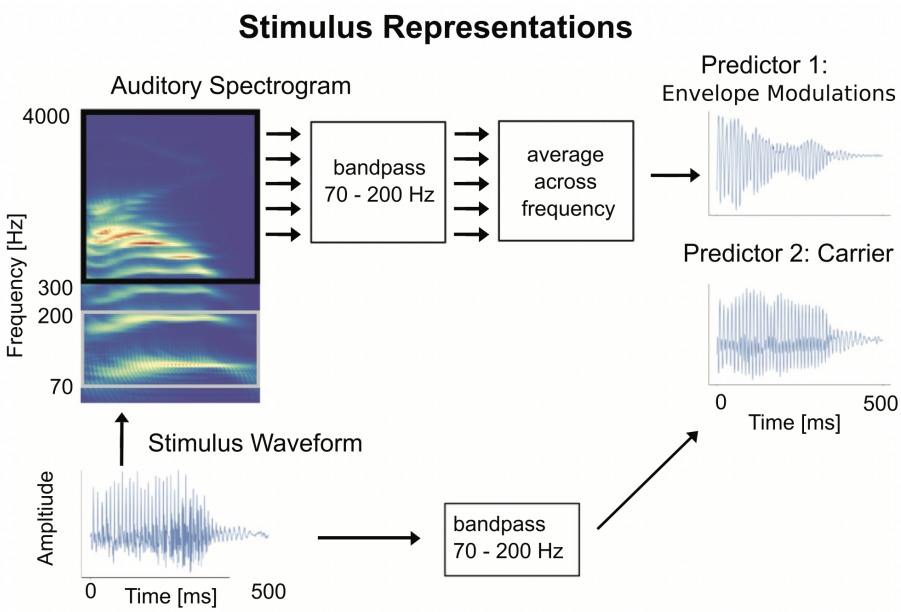


**Figure 1.** Illustration of how the carrier and envelope predictors are extracted from an auditory stimulus. The raw stimulus waveform is shown in the bottom-left corner. **Envelope predictor:** to generate the envelope predictor, starting with the raw waveform and following the arrows up and to the right, first an auditory spectrogram is generated using a model of the auditory periphery (Yang et al., 1992). Then, output of the frequency channels corresponding to 300 to 2000 Hz are selected, bandpassed in the high-gamma range (70-200 Hz), and averaged across the channels. The result is a single time-series signal. **Carrier predictor:** to generate the carrier predictor, following the arrows to the right, the raw stimulus waveform is simply bandpass filtered the in the high gamma range. The result is a second single time-series signal. (Figure reproduced with permission from Kulasingham et al. (2020).)

predictor. Standardized carrier predictors for each stimulus in the condition were concatenated to form one long-form carrier predictor per condition.

### 2.7.2 Envelope Predictor

To create the envelope predictor, the speech was transformed into an auditory spectrogram representation at millisecond time resolution using a model of the auditory periphery (Yang et al., 1992) (http://nsl.isr.umd.edu/downloads.html).

The spectrogram frequency bins in the range 300-4000 Hz were selected, resulting in a time-series for each frequency bin: the time course of the acoustic power in the signal in that band. The range 300-4000 Hz was chosen in order to effect a clear separation between the lowest frequency in the predictor the upper end of the high-gamma range (200 Hz) and because the stimulus was presented through air tubes which attenuate frequencies above 4000 Hz (Kulasingham et al., 2021). Each time-series was filtered to the high-gamma range in the same method as the carrier predictor, using an FIR filter with a 70-200 Hz passband. The time-series signals were then averaged across frequency bins, and the resulting signal standardized, to produce a single time-series – the envelope predictor.

## 2.8 TRF Estimation

The simplest model of a single temporal response function (TRF) is given by

$$y(t) = \sum_{\tau} x(t-\tau)h(\tau) + n(t) \tag{1}$$

where $x(t-\tau)$ is the time-shifted predictor signal (e.g., high-frequency envelope or carrier) at time lag $\tau$; $h(\tau)$ is the TRF at time $\tau$; $y(t)$ is the MEG measured response signal; and $n(t)$ is the residual noise (i.e., everything not captured by convolving the predictor and TRF).

From equation 1, we see that the TRF $h$ is simply the impulse response of the neural system with predictor input $x$ and with MEG measured response output $y$. The TRF can be interpreted as the average time-locked neural response to continuous stimuli (Lalor and Foxe, 2010).

### 2.8.1 Single-Speaker Model

In the present study, a more complex model with two predictors was used for the single-speaker condition

$$y(t) = \sum_{\tau} (x_c(t-\tau)h_c(\tau) + x_e(t-\tau)h_e(\tau)) + n(t) \tag{2}$$

where $x_c$ and $x_e$ are, respectively, the carrier and envelope predictors derived from the single-speaker stimulus, and $h_c$ and $h_e$ are the corresponding TRFs.

### 2.8.2 Cocktail-Party Model

A TRF model with four predictors – the carrier and envelope for the attended speaker and the carrier and envelope for the unattended speaker – was used for the cocktail-party conditions.

$$y(t) = \sum_{\tau} \sum_{s=\{\text{attend,ignore}\}} (x_{c,s}(t-\tau)h_{c,s}(\tau) + x_{e,s}(t-\tau)h_{e,s}(\tau)) + n(t) \tag{3}$$

where predictors and TRFs are similar to the the single-speaker model in 2, with the additional subscript $s$ indicating the attended and unattended speaker. TRFs corresponding to the male and female speakers were analyzed separately, but TRFs corresponding to 'attend language' and 'attend arithmetic' were pooled together within each speaker.

### 2.8.3 Estimation Procedure

The parameters for each TRF model were estimated jointly such that the ordering of the predictors did not affect the estimates, enabling predictors to compete to explain the variance in the data. Predictors that contributed more to the neural response had larger TRFs. TRFs were estimated for time lags from -40 to 210 ms using boosting with cross-validation via the 'boosting' routine from the eelbrain library (Brodbeck et al., 2019). Overlapping bases of 4 ms Hamming windows with 1 ms spacing were employed to promote smoothly varying responses.

Since the source space MEG responses are 3-dimensional current vectors, the estimated TRFs also comprise vectors that span 3 spatial dimensions. The L2 norm of each vector in the TRF was taken at each time instance, resulting in a 1-dimensional time-series for each TRF—one TRF per source space voxel—thereby simplifying the interpretation and visualization of the results.

## 2.9 F0 Analysis

Prior work has posited that high-gamma cortical responses may reflect the processing of F0 and related features in a speech stimulus (Guo et al., 2021). Additionally, Kulasingham et al. (2020) found that high-gamma cortical responses were driven mainly by the segments of the speech with F0 below 100 Hz and that responses to F0 above 100 Hz were not easily detected. This suggests that responses to speech from a typical female speaker (average F0 $\gg 100$ Hz) may be reduced in comparison to responses to a male speaker (average F0 $\sim$100 Hz). To investigate the extent to which the responses in our study were affected by speaker F0, a simple comparison was conducted whereby the time-averaged F0 of each speaker was extracted using Praat (Boersma and Weenink, 2022) and then compared to the amplitude of the TRFs.

## 2.10 Statistical Tests

To determine whether peaks in the estimated TRFs were induced by time-locked neural responses to the predictors and not simply obtained by chance, a noise (floor) model was created by circularly time-shifting the predictors and recomputing TRFs using the shifted predictors. This procedure enables us to disentangle responses to the typical temporal structure of the predictor from responses that time-lock to the predictor. Three shifted versions of each predictor were produced by shifting in increments of one-fourth of the total duration of the original predictor, resulting in three noise-model TRFs for each original TRF. Cluster-based permutation tests (Nichols and Holmes, 2001) with Threshold Free Cluster Enhancement (TFCE) (Smith and Nichols, 2009) were used to test for significance across the TRF peak regions over the average of the three noise models and to account for multiple comparisons. Significance for all tests was set at the 0.05 level.

To test that the TRFs were better than chance at predicting the MEG responses, we compared the prediction accuracy of the TRF model to the average prediction accuracies of the three noise models. Since all predictors were fit jointly, this results in one prediction accuracy per voxel per model. Because each subject was mapped individually to the 'fsaverage' brain, individual variation was mitigated by smoothing the voxel prediction accuracies over the source space using a Gaussian window with 5 mm standard deviation. Cluster-based permutation tests with TFCE were used to test for significance across the cortical region of interest.

In order to test for differences in the prediction accuracies between the two hemispheres, a lateralization test was conducted. This test consisted of subtracting the averaged noise model TRF from the original TRF and then conducting a two-tailed test with paired sample $t$-values and TFCE between the left and right hemisphere prediction accuracies. Care was taken to ensure that only voxels with a symmetric pair in the opposite hemisphere had their prediction accuracies compared.

TRFs were computed for each source voxel as a time-varying, three-dimensional current dipole that varies over time lags. For each TRF vector, its amplitude was compared to the average of three noise models across subjects at each time lag. Each noise model was obtained by circularly shifting the predictors and refitting the TRF model. Time lags for which the true model amplitude was significantly greater than the average noise model was determined using a one-tailed test with paired sample $t$-values and TFCE.

To assess differences in TRF peak amplitude across conditions (single-speaker and cocktail-party) two linear mixed effects models were used. Prior to fitting, the average of the three noise models was subtracted from each TRF; this had the effect of subtracting off the noise floor of each TRF, thereby facilitating a more direct comparison of peak amplitudes. From the result, the peak amplitudes in the range 20-50 ms were extracted. For each condition, two models were developed: one maximal model that attempts to

account for as many fixed (population-level) and random (subject-level) effects as possible in the data, and a reduced model that was obtained by pruning effects from the maximal model that failed to significantly explain variance in the data. Maximal models are the largest possible models that will still converge and were obtained using the R package buildmer. Reduced models were then obtained for each condition using buildmer's backward elimination protocol.

## 3 RESULTS

### 3.1 F0 Analysis

The average F0 for each speaker was computed for voiced regions of speech over all trials:

- Male speaker: average F0 of 95 Hz (std. dev. 8 Hz)
- Female speaker: average F0 of 168 Hz (std. dev. 10 Hz)

Because that the male speaker's average F0 is below 100 Hz, but that the female speaker's average is well above 100 Hz, we anticipate stronger high-gamma cortical responses for the male speaker than the the female speaker.

### 3.2 TRF Response Estimation and Lateralization

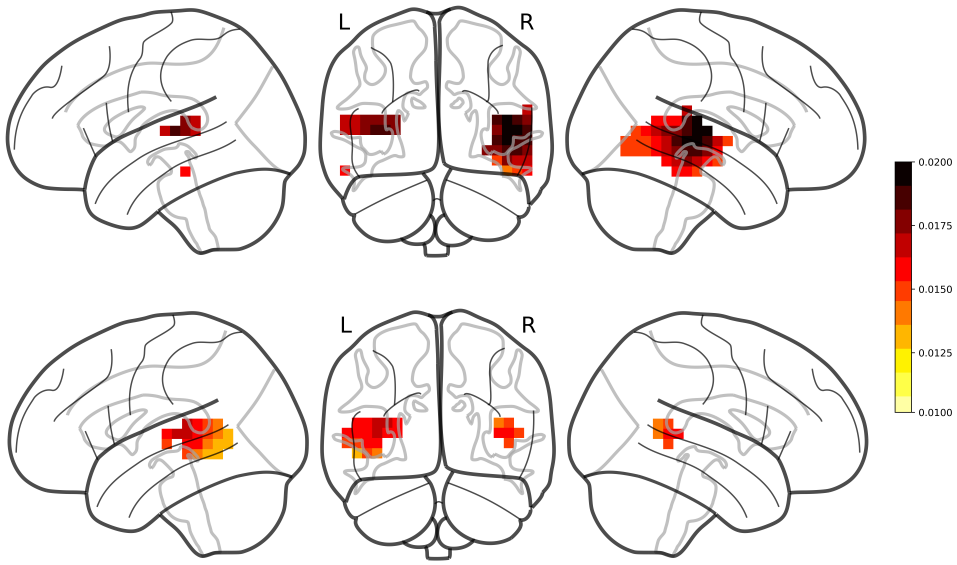


**Figure 2.** Prediction accuracies for male single-speaker (**Top**) and cocktail-party (**Bottom**) models. Red regions denote voxels where the TRF model produced a prediction accuracy that was significantly greater than that of the noise within the ROI. TRFs to female speech (not shown) did not produce significant responses in any voxels.

To validate the extent to which the estimated TRFs can predict the neural responses from the predictor signals, a prediction accuracy is computed for each TRF. The prediction accuracy is the correlation coefficient between the normalized predicted and true neural responses for each TRF. Since a TRF was estimated for each voxel in the source space, this assesses which cortical regions were best predicted by the TRF model.

Prediction accuracies were computed for the single-speaker and cocktail-party models (single-speaker: mean = 0.0149, std = 0.0065; cocktail-party: mean = 0.0132, std = 0.0061). The prediction accuracies for the average of the three noise models (single-speaker Noise: mean = 0.0123, std = 0.0057; cocktail-party Noise: mean = 0.0115, std = 0.0059) were compared to the original models by means of a one-tailed test with paired sample $t$-values and TFCE for each of the two conditions. A large portion of the voxels showed a significant increase in prediction accuracy over the noise model (single-speaker: $t_{\text{max}} = 7.372$, $p < 0.001$, cocktail-party: $t_{\text{max}} = 5.055$, $p < 0.001$; see Figure 2).

No significant voxels were identified in TRFs to the female single speaker (single-speaker: $t_{\text{max}} = 3.436$, $p = 0.055$).

No significant lateralization was found, indicating no evidence that responses were better predicted in one hemisphere versus the other ($t_{\text{max}} = 2.164$, $t_{\text{min}} = -2.422$, $p = 0.459$).

### 3.3 Single-Speaker TRFs

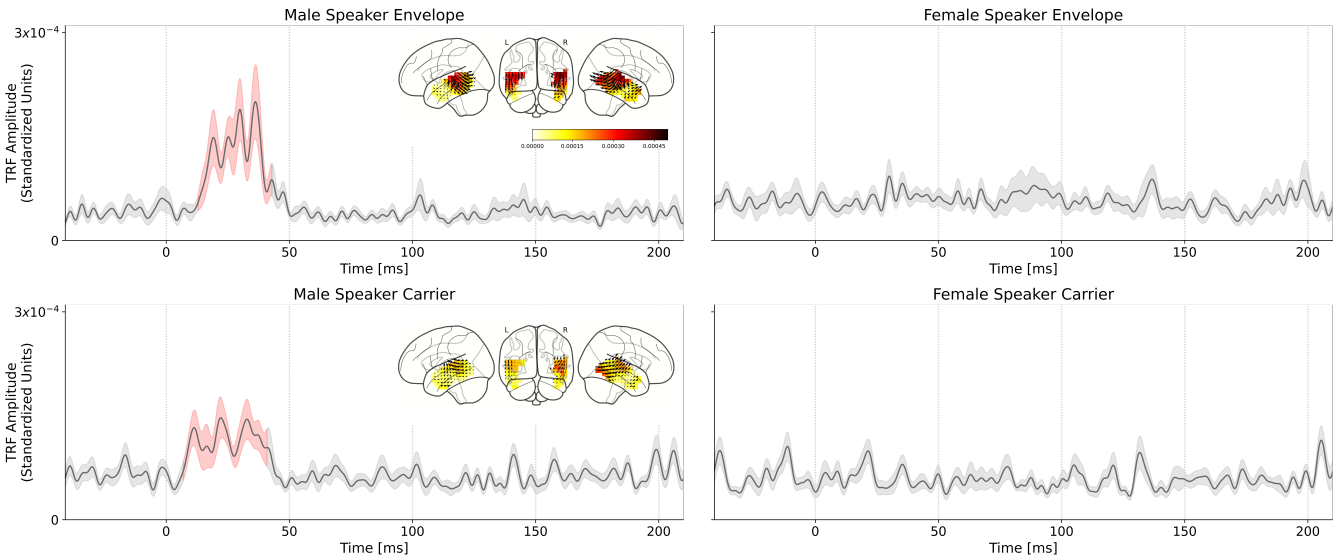


**Figure 3.** Comparison of male speech and female speech TRFs for the single speaker conditions. Solid black lines indicates TRF grand average; shaded regions indicate values within one standard error of the mean. Red shading indicates TRF values significantly above the noise floor. The distribution of TRF vectors in the brain at the time with the maximum significant response are plotted as an inset for each TRF. **Top Left:** average TRF of the envelope predictor derived from the male speaker stimulus. Note the large significant response at ∼30-50 ms in the TRF which indicates a consistent, time-locked neural response to the speech envelope at a 30-50 ms latency. **Top Right:** average TRF of the envelope predictor derived from the female speaker stimulus. Notice the lack of a significant response in the average TRF or a region of significance over the noise model. Similar results were observed for the carrier stimuli: **Bottom Left:** average TRF of the carrier predictor derived from the male speaker stimulus. Note the significant response in the TRF at the same latency observed for the corresponding envelope TRF. **Bottom Right:** average TRF of the carrier predictor derived from the female speaker stimulus. As in the case of the corresponding envelope TRF, there is no significant response observed for this TRF.

Figure 3 shows the various TRFs, averaged across voxels and subjects, and latency ranges for which the TRFs were significantly greater than the noise floor. In total, four TRFs were computed for the single-speaker scenario: carrier and envelope TRFs for male and female speakers.

The envelope TRFs for the male speaker exhibited a significant response over the noise models driven by an effect from 13 to 43 ms ($t_{\text{max}} = 4.643$, $p < 0.001$). Similarly, the significant response of the carrier TRFs to the male speaker was driven by an effect from 19 to 37 ms ($t_{\text{max}} = 3.393$, $p < 0.001$). These results corroborate those obtained in Kulasingham et al. (2020). No significant responses were found for the TRFs for the female speaker.

We used a linear mixed effects model to test the differences between the male and female speaker TRFs. The model was fit to the maximum TRF amplitude for each subject in the range 20-50 ms. The model that best captured the variability in the data (as determined by backward elimination from a maximal model; see section 2.10) was given by:

$$\text{peak amplitude} \sim \text{speaker gender}$$

i.e., a single fixed effect of speaker gender and no random effects. The effect of speaker gender was significant ($F = 18.28$, $p < 0.01$), indicating that speaker gender was the only meaningful predictor of peak height. Additionally, since the reduced model did not contain any effects of predictor type, we conclude that there is no substantive difference between the envelope and carrier predictors in the single-speaker condition – both contribute significantly to predicting the neural response.

### 3.4 Cocktail-Party TRFs

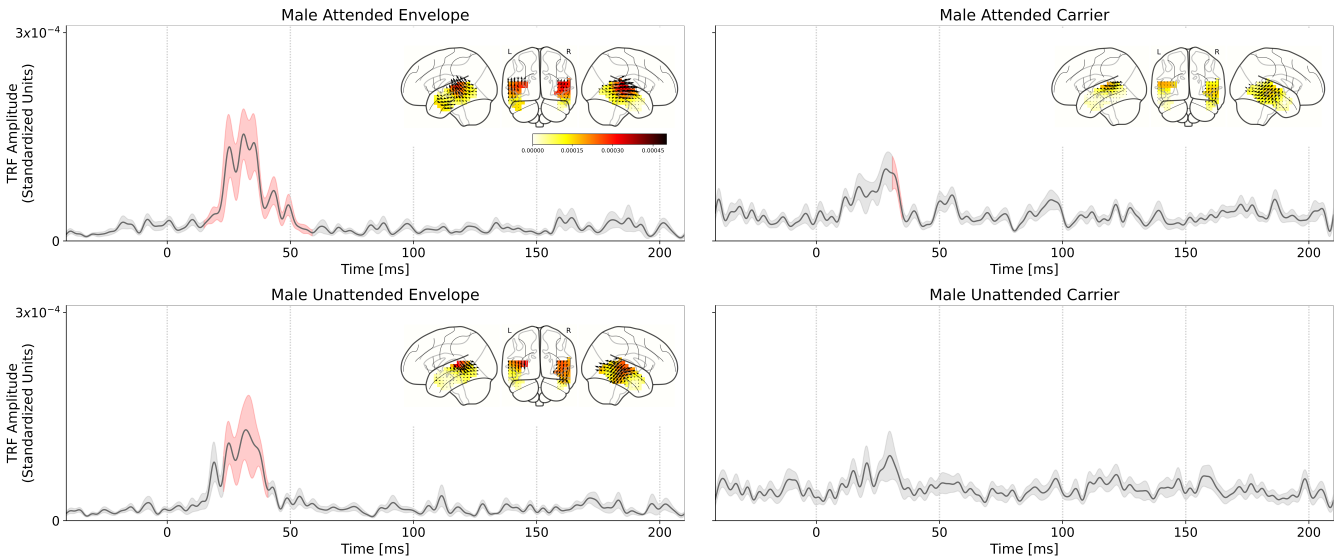


**Figure 4.** Comparison of attended and unattended TRFs for the male speech stimuli, in the cocktail-party setting. Solid black lines indicates TRF grand average; shaded regions indicate values within one standard error of the mean. Red shading indicates TRF values significantly above the noise floor. The distribution of TRF vectors in the brain at the time with the maximum significant response are plotted as an inset for each TRF. **Top Left:** Male speech envelope TRF for subjects attending to the male speech (female speech is background). A large significant response in the TRF is observed between 30-50 ms which indicates a consistent, time-locked neural response to the speech envelope. **Top Right:** Male speech envelope TRF for subjects attending to the female speech (male speech is background). **Bottom Left:** Male speech carrier TRF for subjects attending to the male speech (female speech is background). **Bottom Right:** Male speech carrier TRF for subjects attending to the female speech (male speech is background). Linear mixed effects model and post-hoc test results indicate that the attended speech TRF peak amplitude is significantly greater than the unattended speech TRF peak amplitude.

In the single-speaker case, we reported significant differences in the TRFs of subjects listening to male and female speakers, which differ strongly in their acoustics. In contrast, the cocktail party conditions do not strongly differ in their acoustics but rather only in the subjects' task and state of selective attention; additionally, TRFs are simultaneously obtained for the male speech and female speech for the same stimulus. We repeated the TRF estimation procedure from the single-speaker analysis, with the result being four average TRFs for the cocktail-party scenario: carrier and envelope TRFs for male attended and unattended speech. TRFs for female speech were estimated but not analyzed further due to lack of a significant response. The grand average TRFs are presented in Figure 4.

As in the single-speaker scenario, we compared TRF amplitudes to those of the average noise model to determine the significance of the TRF peaks. Statistical tests revealed that the envelope TRFs for the attend male speaker condition exhibited a significant peak over the noise models driven by an effect lasting from 15 to 59 ms ($t_{\text{max}} = 4.230$, $p < 0.001$). Similarly, the significant regions of the carrier TRFs to the attend male speaker condition were driven by an effect lasting from 31 to 35 ms ($t_{\text{max}} = 2.755$, $p = 0.04$). In the case of the unattended male speaker condition, only the envelope TRF was significant over the noise model, driven by an effect lasting from 23 to 31 ms ($t_{\text{max}} = 3.651$, $p < 0.01$).

Next, the effect of selective attention on TRF peak amplitude was analyzed. A linear mixed effects model was fit to the maximum TRF amplitude for each subject in the range 20-50 ms. The model that best captured the variability in the data (as determined by backward elimination from a maximal model; see section 2.10) was given by:

$$\text{peak amplitude} \sim \text{predictor type} \times \text{attention focus} + (\text{predictor type} \mid \text{subject})$$

The model indicates that the fixed effects and interaction of predictor type and the focus of attention (attend male or attend female) significantly contributes to its prediction of the TRF peak amplitudes, even when controlling for variation in predictor response strength at the subject level. A statistical summary for each model is presented in Table 1. The presence of a signifcant interaction ($t = -2.499$, $p = 0.012$) between predictor type and attention focus suggests that TRF response strength is modulated by attention to different degrees between the envelope and carrier predictors.

A post-hoc Wilcoxon signed-rank test was conducted to test attentional modulation of peak TRF amplitudes between attended and unattended condiditons. Two tests were conducted: one for the envelope TRFs and once for the carrier TRFs. The results showed a significant difference for the envelope TRFs ($W = 29.0$, $p < 0.001$) and no significant difference for the carrier ($W = 122.0$, $p = 0.899$). Figure 5 shows individual subjects' maximum TRF amplitudes in the attend and ignore conditions (male speaker only).

## 4 DISCUSSION

In this study, we investigated time-locked high gamma cortical responses to continuous speech measured using MEG in a cocktail-party paradigm consisting of concurrent male and female speech. Such responses were found, and their volume source localized TRFs provided evidence that these responses are modulated by the focus of attention.

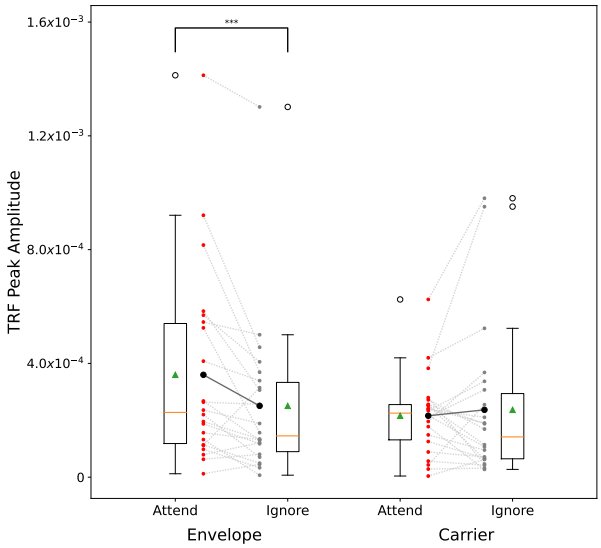


**Figure 5.** Cocktail-party male speech TRF peak amplitude comparison across subjects. Male speech TRF peak amplitudes in the latency range 20-50 ms are presented for attend male (red) and ignore male (grey) conditions. Dashed lines show each individual subject's change in peak height between attend and ignore conditions. Solid lines show the change in the mean between the conditions. For the envelope TRFs, note the significant decrease of the mean value, and for most subjects, between the conditions. No such trend is observed in the carrier TRFs.

### 4.1 The Effect of Selective Attention on High-Gamma Cortical Responses to Continuous Speech

In this work, we assessed the effects of selective attention on time-locked high gamma cortical responses to continuous speech. When subjects were instructed to attend to the male speaker, their time-locked responses to the speech envelope and carrier were significantly larger than when subjects attended to the female speaker. No significant time-locked high gamma responses were seen for the the female speaker, either as a single speaker or concurrently with the male speaker.

The MEG-measured TRFs estimated here indicate a cortical origin of the responses with a ∼40 ms peak latency, in line with the findings from earlier studies on time-locked high gamma auditory cortical responses (Kulasingham et al., 2020; Hertrich et al., 2012) and support the idea that these cortical high gamma responses are due to time-locked responses to the fast (∼100 Hz) oscillations prevalent in vowels produced in the continuous speech of a typical male speaker.

Effects of selective attention on high gamma EEG FFR have also been observed previously, for non-speech sounds (concurrent amplitude modulated tones) by Holmes et al. (2017), and for simple speech sounds (concurrent vowels) but only when already segregated at the periphery (presented dichotically) (Lehmann and Schönwiesner, 2014). The FFR frequencies for which these selective attention effects were observed (∼100 Hz, and 170 Hz, respectively) are consistent with a nerual source of primary auditory cortex, but the FFR paradigm does not lend itself to latency analysis. Recently, using TRF analysis of EEG responses to continuous speech, Kegler et al. (2022) demonstrated that a high gamma TRF (with a latency profile consistent with a contributing source of primary auditory cortex), was modulated by the presence or absence of word-boundaries, i.e., a higher order cognitive (linguistic) cue.

### 4.2 Effect of F0 on High-Gamma Cortical Responses

Most prior studies on high-gamma cortical responses to speech, whether FFR or continuous speech TRFs, employ male speech (e.g., Hertrich et al., 2012; Kulasingham et al., 2020; Canneyt et al., 2021a; Gnanateja et al., 2021; Guo et al., 2021; Kegler et al., 2022). Male speakers typically have lower F0 ($\gtrsim$100 Hz) than typical female speakers with a higher F0 ($\gtrsim$200 Hz). Indeed Kulasingham et al. (2020) observed that even for speech stimuli restricted to a single male speaker, the lower pitch segments of voiced speech contributed more to the cortical response than segments with higher pitch. Furthermore Canneyt et al. (2021b) compared responses to male and female speech, observing that cortical response strength was inversely related to F0 and rate of F0 change throughout continuous speech.

In the present study, we have replicated the findings of these previous works by demonstrating a significant difference in the strength of high-gamma cortical responses to male and female speech. Our findings show a strong, time-locked response to male speech, whether presented in isolation or concurrently with female speech, at a latency of 30-50 ms, and no significant response to female speech, even in isolation. This is consistent with the idea that high gamma time-locked MEG responses can act as a unique window into primary auditory cortex, without interference from subcortical or other cortical areas (Simon et al., 2022).

Although no significant responses to female speech were observed, this does not imply that such responses are not present. Recent studies have shown that response strength greatly improves for stimuli with strong higher harmonic content. For instance, Guo et al. (2021) recorded strong cortical responses in subjects listening to speech-like harmonic stimuli with a missing fundamental. Canneyt et al. (2021b) also observed that stimuli with strong harmonic content evoke stronger cortical responses.

### 4.3 Other Considerations

While the speech stimuli were complete sentences and equations, they were not naturalistic continuous speech: the text was spoken at a fixed rate, the sentences were unrelated to each other, and the syntactic/mathematical form of the stimuli were strongly stereotypical. The results seen here would be strengthened by similar findings using continuous speech.

Previous studies have found that high gamma cortical speech reponses were right-lateralized (Coffey et al., 2016; Kulasingham et al., 2020); we were not able to replicate those findings here, possibly owing to our reduced sample size (22 subjects vs the 40 from Kulasingham et al. (2020)).

Results related to those of this study have also been recently obtained independently, using competing male speakers (Schüller et al., 2023b).

In summary, we have shown that time-locked high-gamma cortical responses to speech are modulated by selective attention in a cocktail-party setting. We have previously argued that time-locked high-gamma MEG cortical responses to speech constitute a valuable physiological window into human primary auditory cortex, with minimal interference from subcortical auditory areas, due to MEG's relative insensitivity to subcortical structures, and minimal interference from higher order cortical areas, due to the high-frequency/low-latency of the responses. In this way we provide new evidence for (top-down) selective attentional processing of competing speakers as early as primary auditory cortex.

## FUNDING

This work was supported by grants from the National Institute of Deafness and Other Communication Disorders (R01-DC019394), the National Science Foundation (SMA-1734892), and the William Demant Foundation (20-0480).

## REFERENCES

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., and Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences* 98, 13367–13372. doi:10.1073/pnas.201400998

Basu, M., Krishnan, A., and Weber-Fox, C. (2010). Brainstem correlates of temporal auditory processing in children with specific language impairment. *Developmental Science* 13, 77–91. doi:10.1111/j.1467-7687.2009.00849.x

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67, 1–48. doi:10.18637/jss.v067.i01

Bidelman, G. M. (2018). Subcortical sources dominate the neuroelectric auditory frequency-following response to speech. *NeuroImage* 175, 56–69. doi:10.1016/j.neuroimage.2018.03.060

Bidet-Caulet, A., Fischer, C., Besle, J., Aguera, P.-E., Giard, M.-H., and Bertrand, O. (2007). Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex. *The Journal of Neuroscience* 27, 9252–9261. doi:10.1523/jneurosci.1402-07.2007

[Dataset] Boersma, P. and Weenink, D. (2022). Praat: doing phonetics by computer

Brodbeck, C., Jiao, A., Hong, L. E., and 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. doi:10.1371/journal.pbio.3000883

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

[Dataset] Brodbeck, C., Teon L Brooks, Proloy Das, and Reddigari, S. (2019). christianbrodbeck/eelbrain: 0.30. doi:10.5281/ZENODO.2653785

Canneyt, J. V., Wouters, J., and Francart, T. (2021a). Enhanced neural tracking of the fundamental frequency of the voice. *IEEE Transactions on Biomedical Engineering* 68, 3612–3619. doi:10.1109/TBME.2021.3080123

Canneyt, J. V., Wouters, J., and Francart, T. (2021b). Neural tracking of the fundamental frequency of the voice: The effect of voice characteristics. *European Journal of Neuroscience* 53, 3640–3653. doi:10.1111/ejn.15229

Coffey, E. B. J., Herholz, S. C., Chepesiuk, A. M. P., Baillet, S., and Zatorre, R. J. (2016). Cortical contributions to the auditory frequency-following response revealed by MEG. *Nature Communications* 7. doi:10.1038/ncomms11070

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., et al. (2000). Dynamic statistical parametric mapping. *Neuron* 26, 55–67. doi:10.1016/s0896-6273(00)81138-1

de Cheveigné, A. and Simon, J. Z. (2007). Denoising based on time-shift PCA. *Journal of Neuroscience Methods* 165, 297–305. doi:10.1016/j.jneumeth.2007.06.003

de Cheveigné, A. and Simon, J. Z. (2008). Sensor noise suppression. *Journal of Neuroscience Methods* 168, 195–202. doi:10.1016/j.jneumeth.2007.09.012

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

Elhilali, M., Xiang, J., Shamma, S. A., and Simon, J. Z. (2009). Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biology* 7, e1000129. doi:10.1371/journal.pbio.1000129

Fischl, B. (2012). FreeSurfer. *NeuroImage* 62, 774–781. doi:10.1016/j.neuroimage.2012.01.021

Gnanateja, G. N., Rupp, K., Llanos, F., Remick, M., Pernia, M., Sadagopan, S., et al. (2021). Frequency-following responses to speech sounds are highly conserved across species and contain cortical contributions. *eneuro* 8, ENEURO.0451–21.2021. doi:10.1523/eneuro.0451-21.2021

Gorina-Careta, N., Kurkela, J. L., Hämäläinen, J., Astikainen, P., and Escera, C. (2021). Neural generators of the frequency-following response elicited to stimuli of low and high frequency: A magnetoencephalographic (MEG) study. *NeuroImage* 231, 117866. doi:10.1016/j.neuroimage.2021.117866

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. doi:10.3389/fnins.2013.00267

Guo, N., Si, X., Zhang, Y., Ding, Y., Zhou, W., Zhang, D., et al. (2021). Speech frequency-following response in human auditory cortex is more than a simple tracking. *NeuroImage* 226, 117545. doi:10.1016/j.neuroimage.2020.117545

Hämäläinen, M. S. and Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & Biological Engineering & Computing* 32, 35–42. doi:10.1007/bf02512476

Hartmann, T. and Weisz, N. (2019). Auditory cortical generators of the frequency following response are modulated by intermodal attention. *NeuroImage* 203, 116185. doi:10.1016/j.neuroimage.2019.116185

Hertrich, I., Dietrich, S., Trouvain, J., Moos, A., and Ackermann, H. (2012). Magnetic brain activity phase-locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech signal. *Psychophysiology* 49, 322–334. doi:10.1111/j.1469-8986.2011.01314.x

Hillyard, S. A., Hink, R. F., Schwent, V. L., and Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science* 182, 177–180. doi:10.1126/science.182.4108.177

Holmes, E., Purcell, D. W., Carlyon, R. P., Gockel, H. E., and Johnsrude, I. S. (2017). Attentional modulation of envelope-following responses at lower (93–109 hz) but not higher (217–233 hz) modulation rates. *Journal of the Association for Research in Otolaryngology* 19, 83–97. doi:10.1007/s10162-017-0641-9

Kegler, M., Weissbart, H., and Reichenbach, T. (2022). The neural response at the fundamental frequency of speech is modulated by word-level acoustic and linguistic information. *bioRxiv* doi:10.1101/2022.04.08.487621

Kraus, N., Anderson, S., White-Schwoch, T., Fay, R. R., and Popper, A. N. (eds.) (2017). *The Frequency-Following Response: A Window into Human Communication* (Springer International Publishing). doi:10.1007/978-3-319-47944-6

Kulasingham, J. P., Brodbeck, C., Presacco, A., Kuchinsky, S. E., Anderson, S., and Simon, J. Z. (2020). High gamma cortical processing of continuous speech in younger and older listeners. *NeuroImage* 222, 117–291. doi:10.1016/j.neuroimage.2020.117291

Kulasingham, J. P., Joshi, N. H., Rezaeizadeh, M., and Simon, J. Z. (2021). Cortical processing of arithmetic and simple sentences in an auditory attention task. *The Journal of Neuroscience* 41, 8023–8039. doi:10.1523/jneurosci.0269-21.2021

Lalor, E. C. and Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience* 31, 189–193. doi:10.1111/j.1460-9568.2009.07055.x

Lalor, E. C., Power, A. J., Reilly, R. B., and Foxe, J. J. (2009). Resolving precise temporal processing properties of the auditory system using continuous stimuli. *Journal of Neurophysiology* 102, 349–359. doi:10.1152/jn.90896.2008

Lehmann, A. and Schönwiesner, M. (2014). Selective attention modulates human auditory brainstem responses: Relative contributions of frequency and spatial cues. *PLoS ONE* 9, e85442. doi:10.1371/journal.pone.0085442

Luo, H. and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54, 1001–1010. doi:10.1016/j.neuron.2007.06.004

Nichols, T. E. and Holmes, A. P. (2001). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping* 15, 1–25. doi:10.1002/hbm.1058

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.e3. doi:10.1016/j.neuron.2019.09.007

Presacco, A., Simon, J. Z., and Anderson, S. (2016). Effect of informational content of noise on speech representation in the aging midbrain and cortex. *Journal of Neurophysiology* 116, 2356–2367. doi:10.1152/jn.00373.2016

Price, C. N. and Bidelman, G. M. (2021). Attention reinforces human corticofugal system to aid speech perception in noise. *NeuroImage* 235, 118014. doi:10.1016/j.neuroimage.2021.118014

Schüller, A., Schilling, A., Krauss, P., Rampp, S., and Reichenbach, T. (2023a). Attentional modulation of the cortical contribution to the frequency-following response evoked by continuous speech. *bioRxiv* doi:10.1101/2023.07.03.547608

Schüller, A., Schilling, A., Krauss, P., and Reichenbach, T. (2023b). Early subcortical response at the fundamental frequency of continuous speech measured with meg. *bioRxiv* doi:10.1101/2023.06.23.546296

Simon, J. Z., Commuri, V., and Kulasingham, J. P. (2022). Time-locked auditory cortical responses in the high-gamma band: A window into primary auditory cortex. *Frontiers in Neuroscience* 16. doi:10.3389/fnins.2022.1075369

Skoe, E. and Kraus, N. (2010). Auditory brain stem response to complex sounds: A tutorial. *Ear & Hearing* 31, 302–324. doi:10.1097/aud.0b013e3181cdb272

Skoe, E., Krizman, J., Spitzer, E., and Kraus, N. (2013). The auditory brainstem is a barometer of rapid auditory learning. *Neuroscience* 243, 104–114. doi:10.1016/j.neuroscience.2013.03.009

Smith, S. and Nichols, T. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage* 44, 83–98. doi:10.1016/j.neuroimage.2008.03.061

Voeten, C. C. (2023). *buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression*. R package version 2.8

Yang, X., Wang, K., and Shamma, S. (1992). Auditory representations of acoustic signals. *IEEE Transactions on Information Theory* 38, 824–839. doi:10.1109/18.119739

Zan, P., Presacco, A., Anderson, S., and Simon, J. Z. (2019). Mutual information analysis of neural representations of speech in noise in the aging midbrain. *Journal of Neurophysiology* 122, 2372–2387. doi:10.1152/jn.00270.2019

**Table 1.** Statistical summary for single-speaker and cocktail-party models.

| | Speaker Gender | Predictor | Significant Lags | Statistics |
|---|---|---|---|---|
| Single-Speaker | male | envelope | 13-43 ms | $t_{\text{max}} = 4.643$, $\mathbf{p < 0.001}$ |
| | | carrier | 19-37 ms | $t_{\text{max}} = 3.393$, $\mathbf{p < 0.001}$ |
| | female | envelope | N.S. | $t_{\text{max}} = 3.790, p = 0.062$ |
| | | carrier | N.S. | $t_{\text{max}} = 2.935, p = 0.188$ |
| | Attended Speaker | | | |
| Cocktail-Party | attend male (ignore female) | male envelope | 15-59 ms | $t_{\text{max}} = 4.230$, $\mathbf{p < 0.001}$ |
| | | male carrier | 31-35 ms | $t_{\text{max}} = 2.755$, **p = 0.04** |
| | | female envelope | N.S. | $t_{\text{max}} = 3.236, p = 0.074$ |
| | | female carrier | N.S | $t_{\text{max}} = 2.466, p = 0.476$ |
| | attend female (ignore male) | male envelope | 23-41 ms | $t_{\text{max}} = 3.651$, $\mathbf{p < 0.01}$ |
| | | male carrier | N.S. | $t_{\text{max}} = 2.758, p = 0.097$ |
| | | female envelope | N.S. | $t_{\text{max}} = 2.979, p = 0.433$ |
| | | female carrier | N.S. | $t_{\text{max}} = 2.421, p = 0.949$ |

**Table 2.** Linear Mixed Effects Model summary, Single-Speaker model.

| Fixed Effects | Estimate | Std. Err. | $t$-Value | $p$-Value |
|---|---|---|---|---|
| Intercept | $4.702 \times 10^{-4}$ | $4.245 \times 10^{-5}$ | 11.075 | **<0.001** |
| Speaker Gender | $-2.567 \times 10^{-4}$ | $6.004 \times 10^{-5}$ | -4.276 | **<0.001** |

**Table 3.** Linear Mixed Effects Model summary, Cocktail-Party model.

| Fixed Effects | Estimate | Std. Err. | $t$-Value | $p$-Value |
|---|---|---|---|---|
| Intercept | $2.159 \times 10^{-4}$ | $4.518 \times 10^{-5}$ | 4.780 | **<0.001** |
| Predictor Type | $1.440 \times 10^{-4}$ | $4.815 \times 10^{-5}$ | 2.990 | **<0.01** |
| Attended Speaker | $2.079 \times 10^{-5}$ | $3.678 \times 10^{-5}$ | 0.565 | 0.57 |
| Predictor Type : Attended Speaker | $-1.300 \times 10^{-4}$ | $5.202 \times 10^{-5}$ | -2.499 | **<0.05** |
| Random Effects | Variance | Std. Dev. | | |
| Intercept — Subject | $3.003 \times 10^{-8}$ | $1.733 \times 10^{-4}$ | | |
| Predictor Type — Subject | $2.125 \times 10^{-8}$ | $1.458 \times 10^{-4}$ | | |