Unfolding of a noise-invariant neural representation of attended speech

Lorenz Fiedler,
Malte Wöstmann,
Sophie Herbst,
& Jonas Obleser

Department of Psychology, University of Lübeck, Lübeck, Germany

Running title: Unfolding of a noise-invariant speech representation

Author correspondence:
Lorenz Fiedler & Jonas Obleser
Department of Psychology, University of Lübeck
Maria-Goeppert Straße 9a
23562 Lübeck, Germany
lorenz.fiedler@uni-luebeck.de; jonas.obleser@uni-luebeck.de

Number of figures: 4
Number of tables: 0
Number of words: 10,838
Abstract (245), Significance statement (119), Introduction (652), Discussion (1,527)

Acknowledgments: Research was supported by the European Research Council (ERC-CoG-2014 646696 to JO) and the Oticon Foundation (NEURO-CHAT).

Abstract

In real-life multi-talker listening environments, the auditory system needs to isolate attended from distracting sound sources and to compensate for non-stationary acoustical conditions. How and at which stages of the central auditory pathway this is achieved is unclear. Here we used electroencephalography (EEG) to investigate the effect of continuously varying signal-to-noise ratio (SNR) on the neural response to speech while listeners (N=18) attended to one of two simultaneously presented, spatially non-segregated talkers. We show that the differential impact of attentional set (i.e., which talker to attend to) and SNR (i.e., which talker is louder) on successive components of neural phase-locking reflects the unfolding of an SNR-invariant representation of the target talker in time and cortical topography. Using a forward encoding-model approach, neural responses to the temporal envelopes of individual talkers and their respective modulation by both, attentional set and SNR were estimated. The model response yielded a clear succession of P1–N1–P2-like components and attention detection accuracies of ~80% in sensor and source space. The earlier component were driven almost exclusively by SNR, while the latest P2 component reflected only attentional set. Under the most adverse SNR, the modeled response yielded an additional, late component and enhanced low-frequency phase coherence to the ignored talker, which indicate contributions of a fronto-parietal attention network in suppressing irrelevant acoustic input. Modeling the neurocortical response can thus provide us with a comprehensive spatio-temporal view on how attentional filters for successful suppression of distracting sensory information are implemented neurally.

Significance statement

Listening requires neural means of tracking an attended sound source (e.g., an attended talker) and of identifying and inhibiting processing of concurrent, distracting sound sources. Here, we investigate the neural response in a highly distracting listening scenario by training forward encoding models, which are linear mappings from the broad-band envelope of the speech signal towards the recorded neural response in the electroencephalogram. Over the initial 400 ms in response to concurrent speech, the neural representation becomes gradually more biased towards the attended source and increasingly invariant to adverse acoustic conditions. These results fill a gap in our understanding of how auditory attentional filters are implemented neurally, that is, when and where attentional control succeeds at suppressing distracting sensory information.

Introduction

Human listeners understand speech even in the presence of distracting sound sources (Cherry, 1953). An emerging question is, how competing acoustic events capture bottom-up attention due to their saliency (e.g., by being louder than the background), and how top-down attention shapes neural responses in order to overcome these adverse listening conditions (Kaya and Elhilali, 2017).

In recent years, the use of encoding and decoding models (Paninski et al., 2007) to investigate the neural responses to continuous speech has opened new paths to study the neural implementation of auditory attention (Lalor et al., 2009). It is by now well-established that the auditory cortical system differentially phase-locks to the temporal envelope of attended vs. ignored speech (Magnetoencephalography: Ding and Simon, 2012a; Electroencephalography: Power et al., 2012). Accordingly, auditory cortical responses allow for a reconstruction of the spectrogram of speech and to detect the attended speaker (e.g., Mesgarani and Chang, 2012; Zion Golumbic et al., 2013).

It is also known that adverse listening conditions in general attenuate the neural tracking of attended speech. Manipulations have included temporal fine structure (Ding et al., 2014), rhythmicity (Kayser et al., 2015), reverberation (Fuglsang et al., 2017) or signal-to-noise ratio (SNR; Kong et al., 2014; Ding and Simon, 2012b; Giordano et al., 2017). Not least, neural selection of speech appears weakened in people with hearing loss (Petersen et al., 2016). Selective neural processing of speech thus compensates for adverse conditions to some degree by filtering the acoustic signal to obtain a more robust neural representation of the attended talker, which reflects in the modulation of the neural response to speech by a listener's attentional set (e.g., which talker to attend to?).

However, the often strictly controlled speech materials of previous studies (e.g. matched sound intensity, dichotic presentation) allow only limited inference to which extent this neural prioritization of attended vs. ignored sound is robust to dynamically varying and more adverse acoustic conditions (i.e. a negative SNR and deficient spatial cues). Commonly, a certain acoustic condition (e.g. SNR) is presented blocked for the duration of minutes (e.g. Ding and Simon, 2012b), whereas in daily-life listening scenarios, acoustic conditions vary more rapidly and more unpredictably. In the present study, to avoid possible effects of a listener's adaptation to longer and highly predictable acoustic conditions, we thus continuously varied the SNR of two concurrent talkers.

The neural response to broad-band continuous speech can be obtained from EEG by estimating the (delayed) covariance of the temporal speech envelope and the EEG, which results in a linear model of the cortical response; a temporal response function (TRF; Lalor et al., 2009; Crosse et al., 2016). Analogous to the event-related potential (ERP), the components of the TRF can be interpreted as

reflecting a sequence of processing stages where later components reflect higher order processes within the hierarchy of the auditory system (Davis and Johnsrude, 2003; Picton et al., 2013; Di Liberto et al., 2015). Typically, this model response or TRF to speech consists of three successive components: Positive weights at around 50 ms ($P1_{TRF}$), followed by negative weights between 100 and 200 ms ($N1_{TRF}$) and another half-wave of positive weights between 200 and 300 ms ($P2_{TRF}$). Although mainly the $N1_{TRF}$ and $P2_{TRF}$ components of the TRF have been found enhanced for attended vs. ignored speech (Power et al., 2012; Kong et al., 2014; Hambrook and Tata, 2014; Fiedler et al., 2017; Horton et al., 2014), a comprehensive understanding of the functional roles of this cascade of neural response components and a more mechanistic link to their underlying neural generators is still missing.

Here, we use a listening scenario with two concurrent talkers undergoing continuous SNR variation. Our results demonstrate differential effects of bottom-up acoustics vs. top-down attentional set on earlier vs. later model response components, respectively. These findings reveal the temporal organization (early vs. late selection) and the underlying neural generators (auditory sensory regions vs. attentional-control parietal and cingular regions) for successful attention to speech.

Methods

Participants

Eighteen native speakers of German (9 females) were invited from the participant database of the Department of Psychology, University of Lübeck, Germany. Six participants were aged between 23 and 33 (M = 27); four participants were aged between 46 and 54 (M = 49); eight participants were aged between 60 and 68 years (M = 64). All reported normal hearing and no histories of neurological disorders. Incomplete data due to recording hardware failure was obtained in four more, initially invited participants. All participants gave informed consent and received payment of 8 €/hour. The study was approved by the local ethics committee of the University of Lübeck.

Experimental Design

The goal of this study was to investigate the selective neural processing of one of two talkers under a continuously varying signal-to-noise ratio (SNR). Here, the signal is a to-be-attended talker and the noise is a to-be-ignored talker. Our study was conducted in a within subject 2 by 3 design (attention by SNR).

The identical mixture of the attended and ignored talker was presented on both ears, resulting in a concurrent listening scenario without any spatial cue (i.e. diotic, Fig. 1A), such that the only cues available for talker segregation consisted in the spectro-temporal features of the talkers, such as pitch, formants, and amplitude modulation. The SNR was stochastically varied between three levels of –6, 0 and +6 dB (Fig. 1B; see *Stimuli* below). The particular dB range was chosen to create a challenging but at the same time solvable listening task. Even if an SNR of –6 dB is rare in real-life listening scenarios (Smeds et al., 2015), the neural tracking of attended speech has been reported as intact at SNRs as low as –6 dB (Ding and Simon, 2013). However, speech perception (number of words repeated correctly) of normal hearing subjects starts to suffer around an SNR < 0 dB and the speech-reception threshold (i.e. 50% correct) usually lies between –5 and 0 dB (Pichora-Fuller et al., 1995, Bentler et al., 2004).

Stimuli

We selected two audiobooks read by native German speakers, one female (Elke Heidenreich, 'Nero Corleone kehrt zurück', read by Elke Heidenreich) and one male (Yuval Noah Harari, 'Eine kurze Geschichte der Menschheit', read by Jürgen Holdorf). The following steps of stimulus preparation were done using custom code written in MATLAB (*Mathwoks Inc.*). Sequences of silence longer than 500 ms were truncated to 500 ms in order to avoid long periods of silence (O'Sullivan et al., 2014). The first hour of each audiobook was selected for further preparation. The first 30 minutes of each audiobook served as the to-be-attended and the rest served as the to-be-ignored speech, such that

all subjects could attend both stories from the beginning and attended (and ignored) both the male and the female voice the same amount of time.

The SNR was modulated symmetrically around 0 dB. An SNR of 0 dB refers to concurrent talker signals with a matched long-term root-mean-square (rms) amplitude as used previously in numerous studies (e.g. Power et al., 2012; O'Sullivan et al., 2015; Mirkovic et al., 2015). Coming from an SNR of 0dB, the SNR was either increased to +6 dB by raising the sound pressure level (SPL) of the to-be-attended talker by 6 dB or decreased to –6 dB by raising the SPL of the to-be-ignored talker by 6 dB (Fig. 1B). As building blocks for SNR modulation, we created a sample of plateaus (i.e., constant SNR of –6, 0 or +6 dB) and ramps (i.e., transition between plateaus). The length of plateaus was uniformly distributed between 5 and 9 seconds in discrete steps of one second. The ramps were linear interpolations between SNRs with the length uniformly distributed between 1 and 5 seconds in discrete steps of one second. The length distributions of plateaus and ramps were kept uniform within each talker and within their assignments as being attended or ignored. By concatenating a 0 dB plateau, a ramp towards +6 or –6 dB, a respective plateau of +6 or –6 dB, and another ramp back to 0 dB, we created sequences with an average length of 20 seconds.

In total, 180 sequences containing either a +6 dB or a –6 dB plateau were created, resulting in a total length of one hour. By randomly concatenating those sequences, we created randomly varying SNR time courses for every subject individually in order to avoid systematic overlap between the SNR modulation and the audiobooks. Stimulus material was cut into twelve blocks, each consisting of 15 sequences, which resulted in an average block length of five minutes. Sound files were created with a sampling rate of 44.1 kHz and a 16-bit resolution. The experiment was implemented in the software *Presentation (Neurobehavioural Systems)*. Stimuli were presented via headphones (Sennheiser HD25).

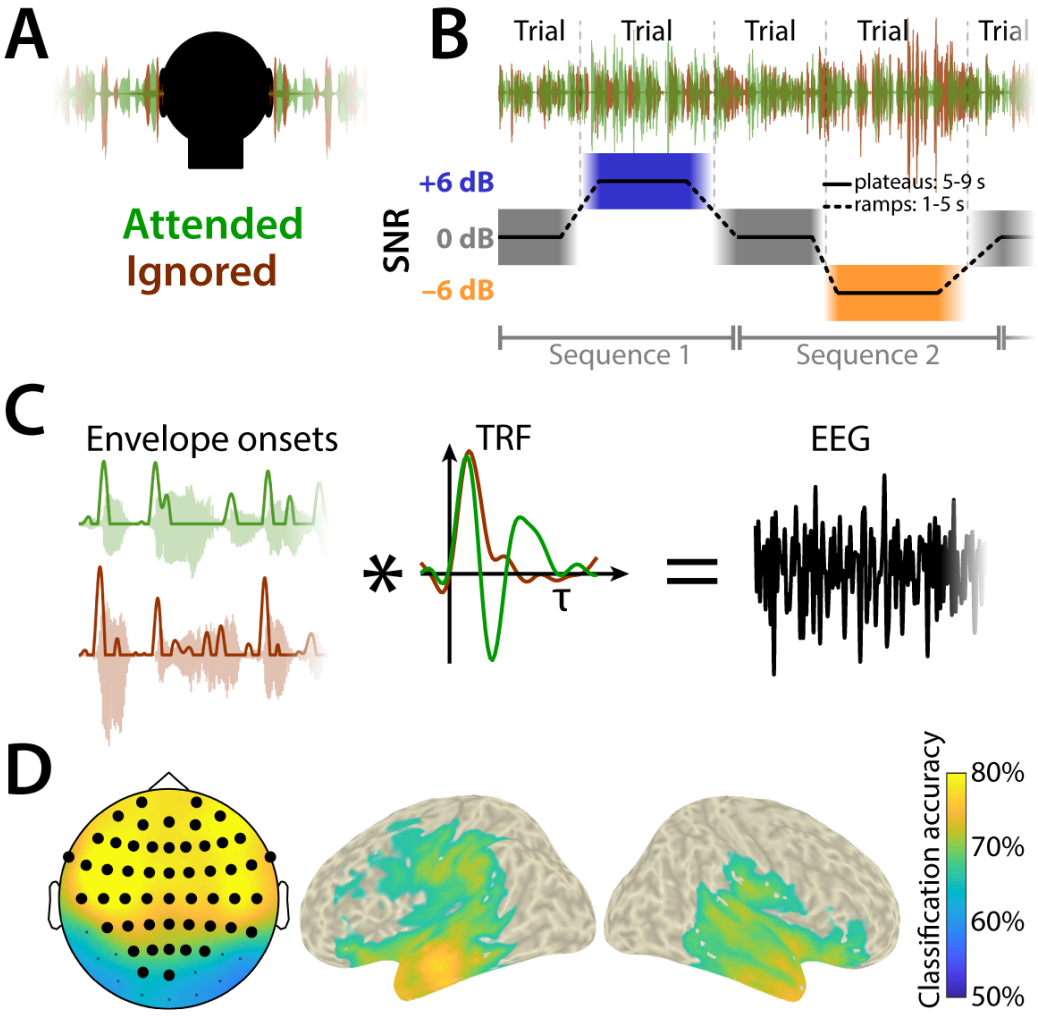

Figure 1: Experimental design, forward model, and classification accuracy. A) Two mixed talkers (female & male) were presented on both ears without spatial segregation (diotic). **B)** The signal-to-noise ratio (SNR) between attended (signal) and ignored (noise) talker was varied between −6, 0 and +6 dB. Length of ramps and plateaus were drawn from uniform distributions. Trials were extracted by cutting the data in the middle of the ramps (i.e. at −3 dB or +3 dB). **C)** Temporal response functions (TRF) to the attended and ignored talker were extracted by a forward (encoding) regression model based on the assumption that the measured EEG signal is the superposition (convolution) of the envelope onsets (of the attended and ignored talkers) and the TRFs, respectively. TRFs reflect the neural response evoked by a single envelope onset. **D)** Accuracy in classification of the attended talker averaged across subjects, obtained by prediction of EEG signals by a forward model (Fiedler et al., 2017) at single EEG channels and single voxels in source space, respectively. Highlighted channels of topographic maps indicate that classification accuracy on the group level was significantly above chance level ($chance_{95\%}$ = 60%). Source maps show voxels exceeding an average classification accuracy of 65%.

Task

The twelve blocks were presented such that subjects were instructed to attend to the female or to the male talker in an alternating fashion. After instruction before each block (i.e. attend to female or attend to male), subjects were asked to start the stimulus presentation by a button press, which

enabled the subjects to take a break between blocks. During listening, subjects were asked to fixate a cross presented on the screen to reduce eye movement.

Every other block, the story picked up at the point it ended two blocks before. After each block, subjects were asked to rate the difficulty of maintaining attention on a continuous color bar ranging from red (difficult) to green (easy) by a mouse click. For later analysis, the continuous color bar was discretized into ten segments (1 = hard, 10 = easy). Subsequently, participants were asked to answer four multiple-choice questions concerning the content of the to-be-attended audiobook. The average rating of difficulty was neither significantly correlated with the number of questions correctly answered (Pearson's r = 0.1, p = 0.7), nor with participants' age (Pearson's r = –0.1, p = 0.5). Furthermore, we found no significant correlation of the number of correctly answered questions with age (Pearson's r = –0.1, p = 0.65).

Data acquisition and preprocessing

EEG was recorded with 64 electrodes *Acticap* (*Easy Cap*) connected to an *ActiChamp* (*Brain Products*) amplifier. EEG signals were recorded with the software *Brain Recorder* (*Brain Products*) at a sampling rate of 1 kHz. Impedances were kept below 10 kΩ. Electrode TP9 (left mastoid) served as reference during recording.

The EEG data were pre-processed in *MATLAB (2017a)* using both the *Fieldtrip*-toolbox (version: 20170321; Oostenveld et al., 2011) and custom written code. The EEG data were re-referenced to the average of the electrodes TP9 and TP10 (left and right mastoids) and resampled to f_s = 125 Hz. The continuous EEG data were highpass-filtered at f_c = 1Hz and lowpass-filtered at f_c = 30 Hz (two-pass Hamming window FIR, filter order: $3f_s/f_c$).

From the continuous EEG data, we extracted the parts during which the twelve blocks of audiobooks were presented (see above). We applied independent component analysis (ICA; Makeig et al., 2004) in order to reject components that were clearly related to eye movements, eye blinks, muscle artifacts as well as heartbeat. On average, 26 components (SD: 7.3) were rejected.

For further analysis, we lowpass-filtered the data again at f_c = 10 Hz (two-pass Hamming window FIR, filter order: $3f_s/f_c$), which assured that the amplitudes at all frequencies up to 8 Hz were not reduced. Previously, neural activity phase-locked to the envelope was only found up to a frequency of approximately 8 Hz (Zion Golumbic et al., 2013; Ding et al., 2014). We could confirm this finding by incrementally raising the cutoff frequency, which didn't change the morphology of the TRFs (see below) but only decreased the prediction accuracy due to the interference of non-phase-locked noise.

Extraction of envelope onsets

A temporal representation of the syllable onsets, further called envelope onsets, was extracted from the presented speech signals (Fiedler et al., 2017). Those representations later served as regressors to model neural responses to the talkers (see below). First, we extracted an auditory spectrogram containing 128 spectrally resolved sub-band envelopes of the speech signals logarithmically spaced between approximately 90 and 4000 Hz using the *NSL* toolbox (Chi et al., 2005). Second, the auditory spectrogram was summed up across frequencies, which resulted in broad-band temporal envelopes of the audiobooks. Taking the derivative of the envelope and zeroing all values smaller than zero (Hertrich et al., 2012) returned the envelope onsets, which only contain positive values at time periods of an increasing envelope, as can be found at syllable onsets (Fig. 1C). Using the envelope onsets as regressor does not imply that we only modeled the encoding of syllable onsets. Every syllable onset is followed by a peak in the speech envelope (Fig. 1B), which is then again followed by an offset and the next onset and so forth, resulting in a high autocorrelation between those features. Nevertheless, onsets are the earliest feature that could possibly evoke a neural response (Picton, 2013). The latency of modelled responses to envelope onsets (compared to envelopes) was found to be most similar to conventional ERPs (Fiedler et al., 2017, supplemental material).

Estimation of temporal response functions

We applied an established method to estimate a linear forward (encoding) model (Lalor et al., 2009; Crosse et al., 2016). The model contains temporal response functions (TRF), which are estimations of the neural response to a certain continuously varying stimulus feature. In our case, this stimulus feature is the envelope onsets (see above) of both, the attended and the ignored talker. Based on the assumption that every sample in the EEG signal *r(t)* is the superposition of neural responses to past onsets and thus can be expressed for one talker by a convolution operation:

$$r(t) = s * TRF = \sum_{\tau} [s(t-\tau) \cdot TRF(\tau)] \quad (1)$$

where *s(t)* is the envelope onsets, TRF is the temporal response function that describes the relationship between s and r over a range of time lags τ (Fig. 1C). The TRF contains a weight for each time lag τ. We investigated time lags in the range from –100 to 500 ms. In order to obtain the weights of the TRF to both talkers contained in the matrix G_{TRF}, ridge regression (Hoerl and Kennard, 1970) was applied, which can be expressed in the linear algebraic form:

$$G_{TRF} = (S^T S + \lambda m I)^{-1} S^T R \quad (2)$$

where *S* is matrix containing the onset envelopes of both the attended and ignored talker and its sample-wise time lagged replications, *R* contains the measured EEG signal, λ is the ridge parameter for regularization, the scalar *m* is the mean of the trace of S^TS (Biesmans et al., 2016) and *I* is the identity matrix. The optimal ridge parameter λ was estimated according to Fiedler et al. (2017) and was set to $\lambda = 10^2$.

TRFs were estimated on a trial-by-trial basis, where trial refers to a part (e.g. a plateau of +6 dB) of certain length cut from the continuous stimulus and the respective EEG data. For the subsequent analysis, we subdivided the data in two ways: First, to get a general estimate of the model's ability to dissociate between attended and ignored talkers, we cut the data into one-minute trials, resulting in trial lengths comparable to previous studies (O'Sullivan et al., 2014; Mirkovic et al., 2015; Biesmans et al., 2016; Fiedler et al., 2017). This resulted in 60 trials per subject. Second, we cut the data based on the applied SNR modulation, which resulted in three groups of trials: –6 dB, 0 dB and +6 dB. To use the entire recording, the data were cut at the time points where ramps of the SNR time courses either crossed –3 dB or + 3 dB (Fig. 1B). This resulted in 180 trials of 0 dB and 90 trials of –6 and +6 dB, respectively. The average length of those trials was 10 seconds (i.e. average length of a plateau (7 seconds) and average length of two halves of a ramp (2x1.5 seconds)). In order to balance the number of trials across SNRs, 90 trials from 0 dB were randomly drawn from the 180 trials of every subject.

Forward model classification accuracy

Besides the statistical analysis of the TRFs (see below), we evaluated the TRFs regarding their ability to detect the attended talker expressed in classification accuracy. In order to obtain classification accuracy, we followed the forward method of predicting two EEG signals and comparing those to the measured EEG signal, as described in detail by in Fiedler et al. (2017). We used the data cut into trials of one-minute length, independent of the applied SNR modulation (see above). This resulted in 60 epochs per subject, which we trained TRFs on. In a leave-one-out fashion, we predicted to be expected EEG signals of a single trial contained in $\hat{R}$ following the equation:

$$\hat{R} = SG_{TRF}, \tag{3}$$

where S is the matrix containing the onset envelopes and G_{TRF} is the matrix containing the TRFs. Two different EEG signals were predicted per trial, the first representing the one and the second the other talker being attended. To obtain a classification decision of which talker was attended, we compared the Pearson correlations from both predicted EEG signals with the measured EEG signal and chose the one that produced the stronger correlation (Fiedler et al., 2017). Per subject, 60 decisions were made.

Classification accuracy was defined as the percentage of trials in which the to-be-attended talker was detected correctly as such by the model prediction. Since this is a forward model approach, classification accuracy is obtained at every single EEG channel (Crosse et al., 2016). Likewise, classification accuracy was obtained at the source level at every single voxel.

Statistical analysis on temporal response functions

To extract significant spatio-temporal deflections in the TRFs at an SNR of 0 dB, we applied a two-level statistical analysis (two-level cluster-test; e.g. Obleser et al., 2012).

On the single-subject level, we used independent sample t-tests to test the TRF to the attended, the ignored as well as the attended–ignored difference against zero. Resulting t-values were transformed to z-scores. Since the weights obtained from Eq. 2 are arbitrary (i.e., depend on λ), we decided to show these normalized (i.e. z-scored) TRFs. These z-scores have the advantage of expressing the deviation from zero relative to the standard deviation of TRFs across trials (as a measure of how consistent TRFs are across the 90 trials of each subject under a certain SNR).

On the group level, the deflection of z-scores from zero was tested by a cluster-based permutation one-sample t-test (Maris and Oostenveld, 2007), which clusters t-values of adjacent time lags in time-electrode space (with a minimum of 4 neighboring EEG channels). The extracted cluster is compared to 4,000 clusters drawn randomly from the data by permuting condition labels. The resulting cluster p-value reflects the relative number of Monte Carlo iterations in which the summed t-statistic of the observed cluster is exceeded.

In a second step, the identical cluster-based permutation test was applied to obtain significant differences between the extreme SNRs –6 vs. +6 dB in the TRF based on the attended and the ignored signal, as well as on the attended–ignored difference.

For illustration of the neural responses, we averaged single-subject z-scores obtained from the first level test across channels of interest. Channels of interest were defined as the channels being part of both significant clusters found in the attended–ignored difference between TRFs under a balanced SNR of 0 dB (Fig. 2C). The 95%-confidence-bands were obtained by bootstrapping (Efron, 1979) across the averaged responses of all subjects, using 4,000 iterations.

Extraction of individual amplitudes and instantaneous phase

Because we observed latency shifts in the TRFs between SNRs, which could be explained by varying degrees of energetic masking (see results), a time-lag wise comparison of TRFs across SNRs wasn't suitable. In order to disentangle amplitude- and latency-effects, we treated the TRF as a band-limited

signal and extracted the amplitude and instantaneous phase of the prominent components ($P1_{TRF}$, $N1_{TRF}$ and $P2_{TRF}$) from the single-trial TRFs in every subject, only averaged across channels of interest.

Amplitude was defined as the maximum or minimum within a certain time interval ($P1_{TRF}$: 0–100 ms; $N1_{TRF}$: 100–200 ms; $P2_{TRF}$: 200–300 ms) of the subject- and SNR-specific TRF. This individual extraction of amplitudes compensated for the observed latency shifts of components.

The instantaneous phase was extracted from the TRF averaged across channels of interest. Here, the instantaneous phase is an appropriate measure, since the time-locked response to continuous speech is band-limited below 8 Hz (Zion Golumbic et al., 2013; Ding et al., 2014) and the TRF was found to pass through three successive and comparably low frequent components. Thus, we avoided to split up the EEG signal into frequency bands of arbitrary edges, but rather looked at the response phase as a sequential process going through several stages (i.e. components). The instantaneous phase φ(τ) was extracted from the complex analytic signal $TRF_a(\tau)$ of the z-scored TRFs of the attended and ignored talker:

$$\varphi(\tau) = \arg\{TRF_a(\tau)\} = \arg\{TRF(\tau) + \mathrm{j}\hat{TRF}(\tau)\}\,, \tag{4}$$

where $\hat{TRF}$ is the Hilbert transform of the TRF.

TRF phase coherence

As in ERPs, a reduced amplitude in the averaged TRF can either originate from reduced amplitude or reduced phase coherence between individual trials or subjects. In order to assure that the observed effects do not base on differences in phase coherence between trials alone, we calculated the phase coherence for every single subject and SNR. The phase coherence was calculated by obtaining the TRFs' analytic signal of every single trial, setting the magnitude of the complex phasor to one, adding up all single trial complex phasors within each SNR and dividing by the number of trials (Lachaux et al., 1999). Analogous to the TRFs, we tested the difference between SNRs against zero in a cluster-based permutation one-sample t-test (see above).

Source localization

To further trace the origin of effects observed in sensor space, we applied LCMV-beamforming (Drongelen et al., 1994; Van Veen et al., 1997) to obtain source-activity time courses in single voxels of the brain. Using a standard template brain from Fieldtrip/SPM (Montreal Neurological Institute) together with the *Acticap* electrode layout, leadfields were calculated with a grid resolution of 10 mm. Individual LCMV-filter weights were obtained using 5% regularization. The continuous time-domain EEG data were projected to source space, resulting in three source activity time courses (X-

Y-Z) per voxel. In order to obtain a single time course in each voxel, the direction of highest variance was determined by principal component analysis and used for further analysis. All further processing steps in source space were done analogously to sensor space EEG data.

Statistical analysis

Statistical analysis was performed according to respective data type and its underlying distribution. We performed cluster-based permutation tests correcting for multiple comparisons on the sensor level (see above). Confidence intervals (95%) were calculated by bootstrapping the mean across the z-scored TRFs of the single subjects (Efron, 1979). The amplitude peak differences were tested using a two-sided t-test. Arithmetic and circular statistical analysis on phase angles were performed using the toolbox *circstat2012a* (Berens, 2009), including the Hotelling paired-samples test (Zar, 1999). Since we observed that pre-requisites for a Hotelling paired-sample test weren't always met, we performed a non-parametric permutation test by shuffling the labels (20,000 permutations) of the to-be-tested data and obtaining a Hotelling paired-sample test statistic for every permutation. Reported p-values (p_{perm}) refer to the relative number of permutations that the test returned a higher F statistic than the empirical data. Confidence slices (i.e. circular confidence intervals; 95%) were calculated by bootstrapping the circular mean across phase angles of the single subjects (Efron, 1979).

Results

After each five-minute block, subjects were asked to rate the difficulty of listening to the to-be-attended talker on a color bar ranging from red (difficult = 1) to green (easy = 10). The average difficulty ratings strongly varied between subjects (mean: 5.2, SD: 2.2, range: 2.3–8.9). No difference in difficulty ratings for of listening to the female versus male talker was found (paired t-test, $t_{17} = 1.17$, $p = 0.26$). With respect to successful attending, participants were asked to answer four multiple-choice questions on the content of the to-be-attended audiobook after each five-minute block. The percentage of correctly answered questions was far above chance (25%) for all participants (mean: 81%, SD: 9%, range: 60–96%). All participants were thus able to follow the to-be-attended talker.

Classification accuracy

To get a general estimate of which EEG channels and which voxels show signatures of selective neural processing, we detected the attended talker by forward prediction of EEG signals (see methods). We obtained highest classification accuracy of approximately 80% at fronto-central channels, slightly lateralized towards temporal channels (Fig. 1D). The source localization revealed that high classification accuracy was mainly driven by temporal channels, where we found classification accuracy within single voxels of up to 78% (Fig. 1D).

Attention-modulated neural responses to concurrent speech

Next, we assessed in greater detail the unfolding of attentional selection of to-be-attended speech in time. To this end, we assessed the most prominent response components and their modulation by attention independent of our SNR manipulation (i.e., we estimated the TRFs from the balanced SNR trials of 0 dB). We inspected both the TRFs to the attended and ignored talker individually (Fig. 2A&B), as well as the difference between the TRFs to the attended and ignored talker (Fig. 3C) to examine signatures of selective neural processing.

Three prominent components ($P1_{TRF}$, $N1_{TRF}$, $P2_{TRF}$; Fig. 2A) were identifiable with notable consistency across individual subjects. The latter two components were absent in the TRF to the ignored talker and thus indicated selective neural processing. All three components ($P1_{TRF}$, $N1_{TRF}$, $P2_{TRF}$) mainly localized to superior and inferior temporal regions (Fig. 2A). Note that the source localizations of the two latter components ($N1_{TRF}$, $P2_{TRF}$) compared well to the source localization of enhanced classification accuracy (Fig. 1D).

First, an early positive component (termed $P1_{TRF}$) appeared in the TRFs to the attended (Fig. 2A, 24–88 ms, $p = 2\times10^{-4}$) and ignored (Fig. 2B, 24–112 ms, $p = 2\times10^{-4}$) talkers, but without any attention-

related difference (Fig. 2C). Latency, polarity, and topography of this component compared well to a P1 as found in auditory evoked potentials (AEPs).

Second, a later negative deflection (termed $N1_{TRF}$) was only present in the TRF to the attended talker (Fig. 2A; 112–176 ms, $p = 5\times10^{-4}$). This component was significantly increased in magnitude (i.e., more negative) for the attended versus the ignored talker (Fig. 2C, 80–176 ms, $p = 5\times10^{-4}$). Noteworthy, the significant attentional modulation of this component (attended–ignored) started already at a time lag of 80 ms, when both the TRF to the attended and to the ignored talkers were still in positive deflection.

Third, a positive deflection between 200 and 300 ms (termed $P2_{TRF}$; Fig. 2A, 216–304 ms, $p = 5\times10^{-4}$), was again only present in the TRF to the attended talker. This component mainly drove the significant difference between the responses to the attended and ignored talker (Fig. 2C, $p = 2\times10^{-4}$). In the same time interval, a comparably long negative deflection was found in the TRF to the ignored talker (Fig. 2B, 248–424 ms, $p = 2\times10^{-4}$). This component was in anti-phase to the $P2_{TRF}$ found in the TRF to the attended talker (Fig. 2A). Effectively, this also enhanced the late, attended–ignored difference in the $P2_{TRF}$ time range (Fig. 2C).

Lastly, a late negative deflection of the response to the attended talker (Fig. 2A, 360–384 ms, $p = 0.03$) was found, but no equivalent cluster occurred in the difference between the TRFs to the attended and ignored talker (Fig. 2C). Hence, this cluster was excluded from further inspections.

Sustained phase-locking of TRFs for attended speech

To further investigate how attention modulates the TRF components, we inspected TRF phase coherence (1–8 Hz) across individual trials. TRF phase coherence to both the attended and the ignored talker peaked at around 100 ms (Fig. 2D), before decaying back to baseline at around 300 ms. This decay of phase coherence, however, was more pronounced in the TRF of the ignored talker (Fig. 2D, $p = 0.01$ at left fronto-central EEG channels). On the source level, this attention-related difference in TRF phase coherence was strongest in the left anterior temporal lobe, where we also found maximal classification accuracy (Fig. 1D).

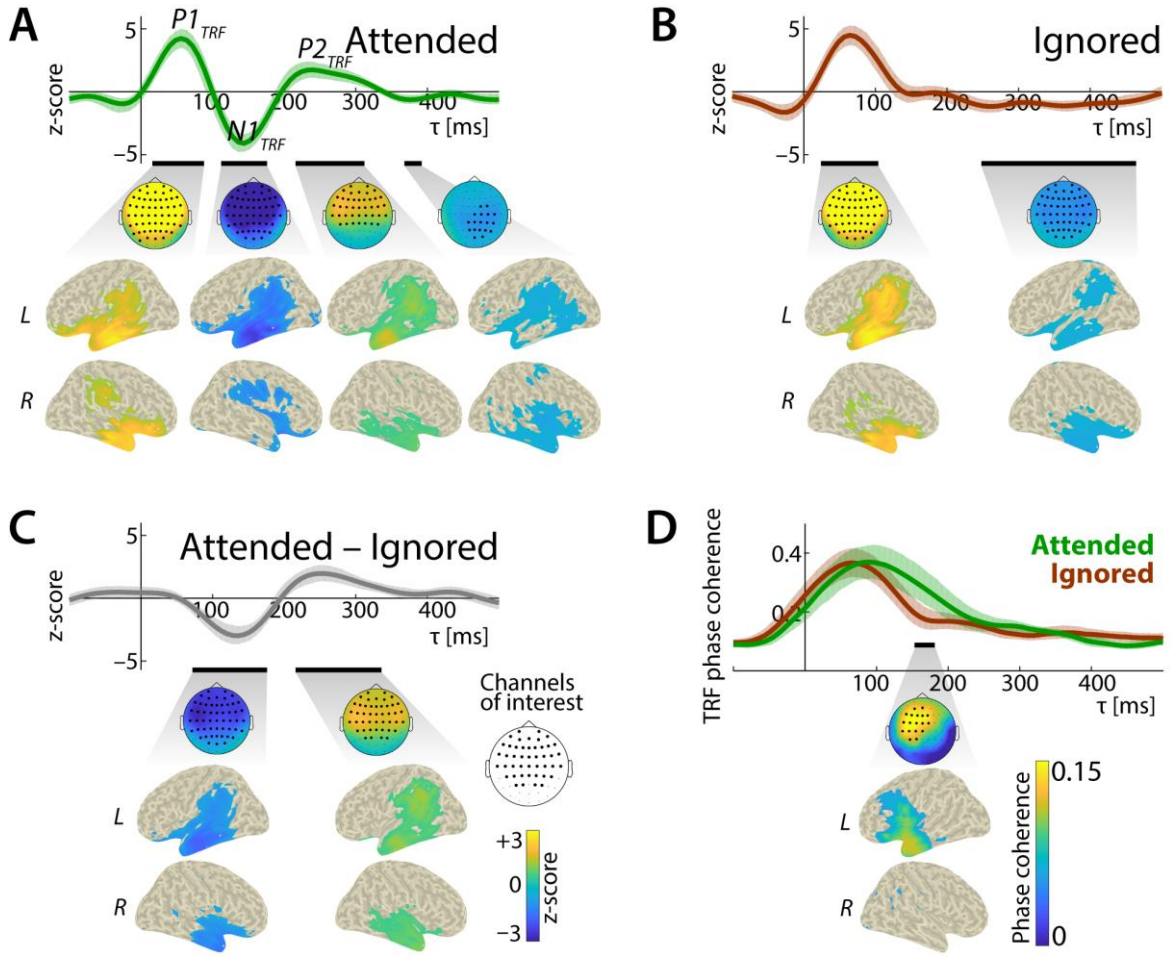

Figure 2: Temporal response functions (TRF) to continuous speech of concurrent talkers under balanced SNR (0 dB). Z-scores of TRFs depict average across subjects. Z-scores were obtained by testing the distribution of single-trial TRF weightings against zero at single EEG channels. TRFs shown are averaged across channels of interest. Confidence bands (95%) were obtained by bootstrapping. Black horizontal lines indicate time ranges of significant difference from zero obtained from a cluster-based permutation test at the group level. Topographic maps show z-scores of clusters averaged across the cluster time range. Highlighted channels are part of the significant clusters. Source localizations show the 35% most strongly contributing voxels. **A)** Responses to the attended talker clearly show a cascade of three components ($P1_{TRF}$–$N1_{TRF}$–$P2_{TRF}$). **B)** Responses to the ignored talker only show a $P1_{TRF}$, whereas the $N1_{TRF}$ and $P2_{TRF}$ are suppressed. **C)** Significant differences between neural responses to the attended and ignored talker are present in the $N1_{TRF}$- and $P2_{TRF}$-timerange. **D)** TRF phase coherence (1–8 Hz) shows sustained phase coherence in the TRFs to the attended vs. ignored talker.

Varying signal-to-noise ratio differentially affects neural responses to attended versus ignored speech

Next, we analyzed the impact of a varying SNR on the TRFs identified in response to SNR = 0 dB. To this end, we contrasted the TRFs of the two extreme conditions, –6 vs. +6 dB (Fig. 3).

The TRFs to the attended talker (Fig. 3A) showed significant SNR differences during two time intervals, first at around 100 ms (72–136 ms, $p = 2\times10^{-4}$) and second around 200 ms (176–232 ms, $p = 2\times10^{-4}$). These differences occurred in the transition between components ($P1_{TRF}$ to $N1_{TRF}$, and $N1_{TRF}$ to $P2_{TRF}$). This was consistent with the visual impression of the TRFs being similar in morphology, yet delayed under an SNR of –6 dB compared to +6dB.

The TRFs to the ignored talker (Fig. 3B) also showed such an SNR-related delay, captured by a negative cluster (96–104 ms, $p = 0.04$). Two later additional components appeared under an SNR of –6 dB compared to +6 dB selectively for ignored speech: the first (160–176 ms, $p = 0.02$) localized to temporal regions, and the second localized to parietal regions (240–280 ms, $p = 0.004$). This parietal localization clearly differentiated this detrimental-SNR, ignored-speech component from all others.

Exploratory inspection of TRF phase coherence (1–8 Hz; Fig. 3D) across SNRs gave further evidence for a superimposed neural mechanism being involved in the suppression of the ignored talker. The TRF to the ignored talker exhibited significantly enhanced phase coherence under an SNR of –6 dB (vs. +6 dB), again in the time range of the late $P2_{TRF}$ and again at parietal EEG channels (232–248, $p = 0.005$). No such change was observable in the TRF to the attended talker (Fig. 3D inset). In source space, this enhanced phase coherence localized to the dorsal anterior cingulate (dACC), spreading into parietal regions. This is further, if exploratory, evidence for an additional neural mechanism originating from non-auditory, supra-modal regions in the suppression of the ignored talker.

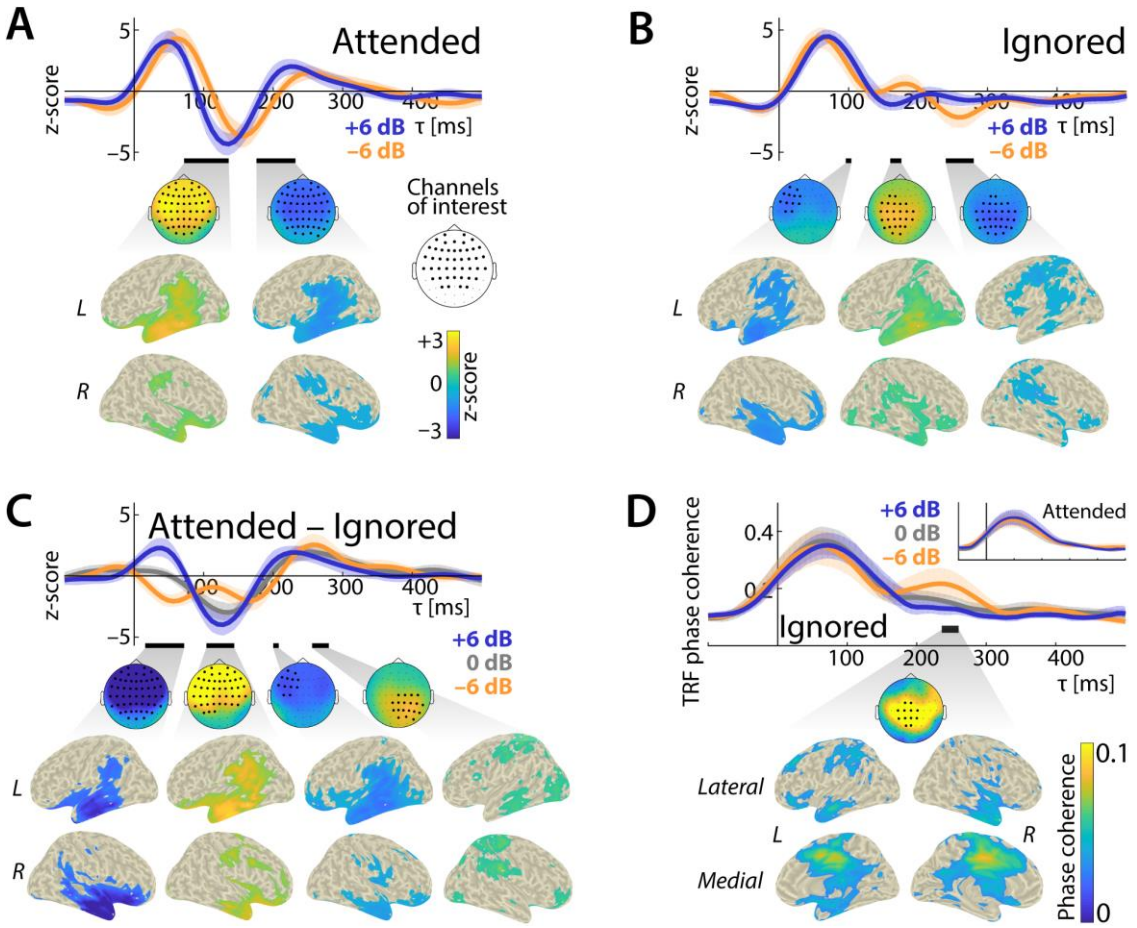

Figure 3: Temporal response functions (TRF) to continuous speech for signal-to-noise ratios of –6 vs. +6 dB. Z-scores of TRFs depict average across subjects. Z-scores were obtained by testing distribution of single-trial TRF weightings against zero at single EEG channels. TRFs shown are averaged across channels of interest. Confidence bands (95%) were obtained by bootstrapping. Black horizontal lines indicate time ranges of significant difference between –6 vs. +6 dB obtained from a cluster-based permutation test at the group level. Topographic maps show z-scores of clusters averaged across the cluster time range. Highlighted channels are part of the significant clusters. Source localizations show the 35% most strongly contributing voxels. **A)** Responses to the attended talker are delayed under an SNR of –6 dB compared to +6 dB. **B)** Under an SNR of –6 dB, a late component appeared, which was localized in parietal and frontal regions. **C)** The components of the difference between the responses to the attended and ignored talker are differently affected by a varying signal-to-noise ratio. Note that TRF under 0 dB is only shown in C for better overview. **D)** TRF phase coherence (1–8 Hz) of TRFs to the attended (inset) and ignored talker under the different SNRs.

Unfolding of a noise-invariant representation of the attended talker: TRF magnitude

A central question was how the SNR (i.e., bottom-up acoustic conditions) affects the difference in attending versus ignoring speech (i.e., the top-down attentional set), which is shown in detail in Figure 3C. Crucially, in order to account for observed latency shifts in the TRFs, we here also inspected the amplitude differences (attended–ignored) at individual participants' peaks of the TRF components (shown in Figure 4 A&B).

During the early time interval of the $P1_{TRF}$, the TRF difference (attended–ignored) indicated that a higher relative sound intensity evoked a more positive $P1_{TRF}$ amplitude, independent of being attended or ignored (16–72 ms, $p = 2\times10^{-4}$). The $P1_{TRF}$ peak amplitude difference (attended vs. ignored, Fig. 4B) showed no significant difference from zero under an SNR of 0 dB (one-sample t-test, $t_{17} = -0.01$, $p = 0.99$), whereas under an SNR of –6 dB, all but two subjects showed a negative difference ($t_{17} = -6.2$, $p = 1\times10^{-5}$) and all subjects showed a positive difference under an SNR of +6 dB ($t_{17} = 5.8$, $p = 2\times10^{-5}$). This linear trend was reflected in the highly significant differences between SNRs (–6 vs. 0 dB: $t_{17} = 6.0$, $p = 1.5\times10^{-5}$, 0 vs. +6 dB: $t_{17} = 5.8$, $p = 2\times10^{-5}$, –6 dB vs. +6 dB: $t_{17} = 7.3$, $p = 1.3\times10^{-6}$). This contrast of the three SNRs centered around zero, which indicates that the $P1_{TRF}$ amplitude is purely driven by the varying SNR, with the absence of any attention-related influence.

The $N1_{TRF}$ was more negative to the attended vs. ignored talker under all SNRs. Critically, the TRF difference (attended–ignored) of this attentional modulation of the TRF further increased (i.e., larger negativity in the neural response; Fig. 3C, 104–144 ms, $p = 5\times10^{-4}$) under a more favorable SNR (–6 vs. +6 dB). The $N1_{TRF}$ peak amplitude difference (attended vs. ignored, Fig. 4B) turned out to be generally affected by attention, since across all SNRs, a more negative TRF (negative offset) could be detected in the response to the attended talker, resulting in significant differences from zero under all SNRs (one-sample t-test, –6 dB: $t_{17} = -3.6$, $p = 0.002$; 0 dB: $t_{17} = -8.1$, $p = 3\times10^{-7}$; +6 dB: $t_{17} = 8.8$, $p = 10^{-9}$). Interestingly, the negativity of the $N1_{TRF}$ increased with a more favorable SNR (negative slope across SNRs in Fig. 4B), which was reflected in a significant difference between SNRs (paired sample t–test, –6 dB vs. 0 dB: $t_{17} = -2.5$, $p = 0.02$; 0 dB vs. +6 dB: $t_{17} = -4.4$, $p = 4\times10^{-4}$, –6 dB vs. +6 dB, $t_{17} = -4.7$, $p = 2\times10^{-4}$). Thus, the $N1_{TRF}$ peak amplitude was always more negative in the TRF to the attended talker, which indicates a consistent signature of selective neural processing. However, the $N1_{TRF}$ magnitude was not entirely robust against varying acoustic conditions.

Lastly, the magnitude of the TRF difference (attended–ignored) in the $P2_{TRF}$ interval was remarkably constant across SNRs (Fig. 3C). However, the delay (for –6 vs. 6 dB) and the additional component in the response to the ignored talker at an adverse SNR of –6 dB (Fig. 3B) might indicate an additional mechanism being involved during that comparably late interval of the responses to the concurrent talkers. Finally, the $P2_{TRF}$ peak amplitude difference (attended–ignored, Fig. 4B) showed an increased response to the attended talker under all SNRs (one-sample t-tests, –6 dB: $t_{17} = 4.8$, $p = 2\times10^{-4}$; 0 dB: $t_{17} = 7.6$, $p = 8\times10^{-7}$; +6 dB: $t_{17} = 8.8$, $p = 10^{-7}$). In contrast to the $N1_{TRF}$ amplitudes, the $P2_{TRF}$ amplitudes were not modulated by SNR (paired sample t-test, –6 dB vs. 0 dB: $t_{17} = 0.3$, $p = 0.77$; 0 dB vs. +6 dB: $t_{17} = 0.8$, $p = 0.4$, –6 dB vs. +6 dB, $t_{17} = 0.95$, $p = 0.35$). This indicates the $P2_{TRF}$ amplitude to be robust against a varying SNR and solely driven by attention.

In sum, whether a talker was attended or ignored did not affect early TRF components (Fig. 4B; $P1_{TRF}$) but only the later components (Fig. 4B; $N1_{TRF}$ & $P2_{TRF}$). In contrast, the impact of SNR was large for early ($P1_{TRF}$ & $N1_{TRF}$) but absent for late neural response components ($P2_{TRF}$). Thus, the peak amplitudes of the TRFs indicate that the neural representation of concurrent speech within the first ~400 ms becomes gradually more biased towards the attended talker and becomes gradually more SNR-invariant

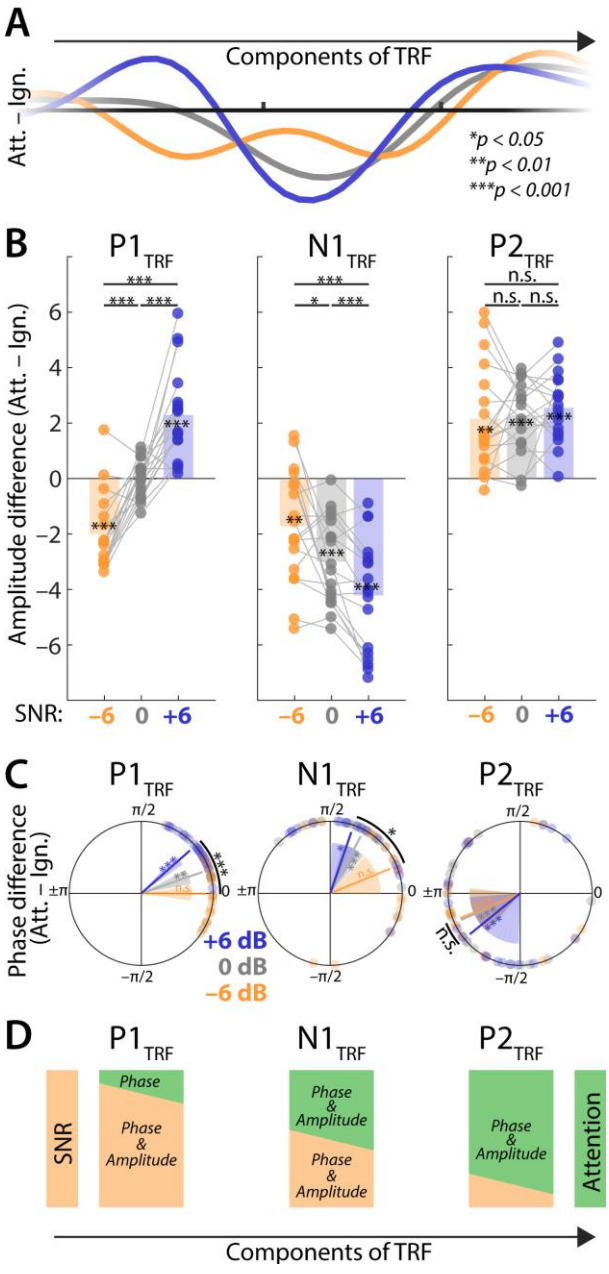

Figure 4: Effects of attentional set and SNR on response components. A) The attended–ignored amplitude difference of the TRFs for the three SNR levels (adopted from Fig. 3C) shows three major components ($P1_{TRF}$, $N1_{TRF}$ & $P2_{TRF}$). **B)** Individual peak amplitude difference. Within each panel, the slope across the three bars indicates the effect of increasing SNR, whereas an offset indicates the influence of attending (versus ignoring). **C)** Individual differences in phase angles (attended–ignored) for TRF components $P1_{TRF}$, $N1_{TRF}$ & $P2_{TRF}$ under different SNRs (color coded). Dots show phase angle differences for individual participants, colored lines and shades indicate the circular mean phase angle difference and the 95%-confidence-slices obtained by bootstrap across subjects. **D)** The schematic illustration depicts the successively decreasing impact of SNR (orange) and the increasing impact of attention (green) on the TRF components ($P1_{TRF}$, $N1_{TRF}$ & $P2_{TRF}$). The labels *amplitude* and *phase* describe the measures affected by attention and SNR, respectively.

Unfolding of a noise-invariant representation of the attended talker: TRF phase

In the section above we controlled for latency shifts of TRF components in order to investigate effects of SNR and attention on TRF amplitude. Here, to also investigate latency-specific effects of attention and SNR on neural response components, we determined the individual instantaneous phase of the TRFs (see *Methods*). Specifically, we investigated the SNR-dependent phase difference (attended–ignored) by extracting phase angles at the time lags of the three prominent TRF components for every single subject. Fig. 4C shows the attended–ignored phase difference of the three prominent components ($P1_{TRF}$, $N1_{TRF}$, $P2_{TRF}$) under the three different SNRs (–6, 0, +6 dB). Analogous to the analysis of the components' amplitude, we can assume that a phase difference under an SNR of 0 dB is purely attention-related, whereas the change of the phase difference across SNR levels is due to the varying acoustics.

Interestingly, in the $P1_{TRF}$ we found a phase difference (attended–ignored) under an SNR of 0 dB (Hotelling paired-sample test, mean: 0.34 rad, $F_{2,16} = 7.3$, $p_{perm} = 0.002$). Since the sound intensity was balanced, this delay indicates that the response to the attended talker is leading already at the early stage of the $P1_{TRF}$. This phase difference was also modulated by SNR. Under the favorable SNR of +6 dB, this early phase difference (attended–ignored) further increased (Hotelling paired-sample test, mean: 0.72 rad, $F_{2,16} = 59.2$, $p_{perm} = 4\times10^{-6}$), whereas under the adverse SNR of –6 dB, it diminishes (Hotelling paired-sample test, mean: –0.01 rad, $F_{2,16} = 0.016$, $p_{perm} = 0.99$). Contrasting the phase difference under an SNR of –6 dB against +6 dB confirmed a significant increase of phase difference due to a more favorable SNR (Hotelling paired sample test, mean: –0.72 rad, $F_{2,16} = 12.1$, $p_{perm} = 5.5\times10^{-4}$). The early attended–ignored phase difference in $P1_{TRF}$ indicates an early attentional selection.

In the $N1_{TRF}$, an even stronger phase difference (attended–ignored) was found. Under the balanced SNR of 0 dB, we found a significant phase difference (attended–ignored) under the balanced SNR of 0 dB (Hotelling paired-sample test, mean: 1.1 rad, $F_{2,16} = 554.3$, $p_{perm} = 3.8\times10^{-6}$, Fig. 4C, center). Comparable to the $P1_{TRF}$, the phase difference was modulated by SNR. Under the favorable SNR of +6 dB, a further increase of the phase difference (attended–ignored) was present (Hotelling paired-sample test, mean: 1.2 rad, $F_{2,16} = 27.2$, $p_{perm} = 0.029$). Under the adverse SNR of –6 dB, the phase difference (attended–ignored) was not significant (Hotelling paired-sample test, mean: 0.39 rad, $F_{2,16} = 66.8$, $p_{perm} = 0.44$), but the confidence slice indicated an evolving phase difference even under the adverse SNR of –6 dB. Contrasting the phase difference under an SNR of –6 dB against +6 dB revealed a significant increase of phase difference due to a more favorable SNR (Hotelling paired-sample test, mean: –0.94 rad, $F_{2,16} = 4.3$, $p = 0.03$). Note that even though we found a phase difference in the $N1_{TRF}$,

this phase difference was not exceeding π/2 (i.e. 90°). Thus, we cannot speak of a counter-phasic relationship at this stage.

Strikingly, the later $P2_{TRF}$ showed an attended–ignored phase difference under all SNRs (Hotelling paired-sample test, –6 dB: mean: –2.8 rad, $F_{2,16} = 62.5$, $p_{perm} = 4\times10^{-6}$, 0dB: mean: –2.7 rad, $F_{2,16} = 37.6$, $p_{perm} = 4\times10^{-6}$, –6 dB: mean: –2.5 rad, $F_{2,16} = 42.7$, $p_{perm} = 5\times10^{-5}$, Fig. 4C, right). In contrast to the preceding components, the phase difference under an SNR of –6 dB against +6 dB revealed no significant increase of phase difference due to a more favorable SNR (Hotelling paired-sample test, mean: –0.13 rad, $F_{2,16} = 0.64$, $p_{perm} = 0.55$). Comparable to the amplitude differences, the almost counter phasic relationship between TRFs to the attended and ignored talker (which reflects in phase angles of the attended–ignored difference close to ±π) present under all SNRs indicates an SNR-invariant selective neural processing of the concurrent talkers.

Discussion

In the present study, human listeners attended to one of two concurrent talkers under continuously varying signal-to-noise ratio (SNR). Forward modeling revealed neural responses to the temporal envelopes of individual talkers and their modulation by both, top-down attentional set, and bottom-up SNR. The model response yielded a clear succession of P1–N1–P2-like components, localized in auditory temporal regions, with attention-classification accuracies around 80%. While a distinction between different SNR levels occurred for earlier components (P1 and N1), separation between attended and ignored talkers unfolded later in time (N1 and P2), establishing an SNR-invariant representation of attended speech. Critically, under the most adverse SNR, distinct late components in the modeled response to ignored speech originating in a supra-modal attentional network indicate suppression of irrelevant acoustic input.

Neural responses reflect unfolding of a noise-invariant representation of attended speech

In accordance with previous studies on neural tracking of auditory stimuli in EEG (e.g. Power et al., 2012), three prominent components from the modeled TRFs to the attended talker were selected for further investigation (Fig. 2A; $P1_{TRF}$, $N1_{TRF}$ and $P2_{TRF}$). Akin to the more classically studied auditory evoked potential (AEP), we interpret the TRF as a sequence of components reflecting consecutive stages of (selective) neural processing along the auditory pathway (for review see: Picton, 2013).

The $P1_{TRF}$ peak amplitude was strongly dominated by the saliency of the talkers (i.e. the SNR variation), independently from being attended or ignored. This agrees with the supposed role of the P1, described as reflecting mostly bottom-up processing (Herrmann et al., 2013). At this relatively early stage of auditory processing, the relevant spectro-temporal features of the acoustic input might be extracted. Attentional modulations at this component have rarely been described (Giuliano et al. 2014; cf. Picton & Hillyard 1974; Ding and Simon, 2012b). In the present data, the only signature of selective processing during the $P1_{TRF}$ was the slight forward-shift in phase for the $P1_{TRF}$ to the attended compared to the ignored talker (Fig 4C), which most likely results from the more negative $N1_{TRF}$ to the attended talker. This early phase shift is suggesting that as soon as relevant features of the attended talker are identified at the stage of the $P1_{TRF}$, the $N1_{TRF}$ is evoked (Fig 2A), whereas no $N1_{TRF}$ is elicited by irrelevant features leading to a longer sustain (and phase shift) of the $P1_{TRF}$ (Fig 2B), which, in line with Chait et al. (2010), also leads to a relatively early TRF difference (attended–ignored) emerging at 80 ms (Fig 2C).

The $N1_{TRF}$ was strongly selective towards the attended talker, both in terms of magnitude and phase differences between the modeled response to the attended and the ignored talker. In AEPs, comparable effects have been observed (e.g. Hillyard et al., 1973; Nätäänen et al., 1981). The N1 can

be regarded as the pivotal stage of attentional selection, decisive upon the 'perceptual fate' of concurrent speech signals. First, this negative-going deflection of the modeled response function in the 100–150 ms time window is the most robustly replicated TRF component (Ding & Simon, 2012a, Ding & Simon, 2012b). Second, the most likely generators of the auditory N1 are located in superior temporal gyrus (STG; Obleser et al., 2004b; Scherg et al., 1989; Tavabi et al., 2007; see also Figs. 2,3), a region shown to hold strongly attentionally biased representations of speech (e.g. Obleser et al., 2004a). Recent attempts to directly reconstruct attended and ignored speech from the gamma-band electrocorticographic response to mixed speech from STG revealed a representation of the attended speech signal with a fidelity that approaches clean speech (Mesgarani and Chang, 2012). Thus, at the level of the STG, with a delay of about 100 to 150 ms, a speech signal that is being successfully ignored is virtually absent. Here, however, we demonstrate that this attentional selectivity of the $N1_{TRF}$ is not SNR-invariant, but benefits from better SNR.

The modulation of magnitude, phase and likely generators of the ensuing $P2_{TRF}$ component demonstrate how such a robust neural representation of attended speech might be brought about: The $P2_{TRF}$ was found to be strongly selective towards the attended talker. In contrast to the $N1_{TRF}$, the strength of this selectivity was not affected by adverse SNRs (Fig. 4B,C). A robust representation of the attended signal at the $P2_{TRF}$ is in line with previous findings: Fuglsang et al. (2017) exposed participants to a cocktail-party scenario of varying reverberation and found the $P2_{TRF}$ most robust. Di Liberto et al. (2015) also suggested the $P2_{TRF}$ to reflect an enhanced, post-categorical stage of speech processing along the auditory pathway. Taken together, those findings suggest that the $P2_{TRF}$ reflects a neural stage by which the representation of the attended signal has been largely isolated from distracting sources.

Using the forward encoding model including all its components, the detection of the attended talker revealed enhanced classification accuracy (i.e., attentional selectivity) at fronto-temporal sites, which is in line with previous findings in backward models (O'Sullivan et al., 2014; Mirkovic et al., 2015; Fuglsang et al., 2017). Crucially, we show that the enhanced classification accuracy mainly emerges from auditory brain areas, namely superior and middle temporal cortex (Fig. 1D) and show which components ($N1_{TRF}$, $P2_{TRF}$) are driving this attentional selectivity and how those components are affected by varying acoustical conditions.

Notably, previous studies had remained incongruent with respect to what might be the earliest cortical signature of selective attention in such a concurrent-speech setup (Power et al., 2012; Kong et al., 2014; O'Sullivan et al., 2014). By comparison, our findings indicate that the attentional effort of selective neural processing is discernable within the $N1_{TRF}$ window only if the concurring stimuli are spatially non-segregated. In an AEP study, Lange (2012) also related the N1 to temporal (but not

spatial) attention. However, Forte et al. (2017) found mechanisms of auditory selective attention already at the brainstem level. Upon further experimentation, effects of selective neural processing on the components of the TRF seem to strongly depend on the cues available for talker segregation.

To our knowledge, the instantaneous phase of TRFs has not been analyzed before. We argue here that the unfolding anti-phasic TRF for attended vs. ignored speech not only indicates amplification of relevant but also active inhibition of irrelevant acoustic input (Fig. 4C). In the $P2_{TRF}$, the attentional selection not only suppresses the response to the ignored talker (which would result in an amplitude but not a phase difference), but rather responds in an anti-polar fashion. Phase-related sensitivity and attention effects have been found both in the visual and in the auditory domain (Spaak et al., 2014; Lakatos et al., 2008; Henry and Obleser, 2012). Here we show that attention establishes such an anti-phasic relationship, which reflects active inhibition of the ignored talker by (Fig. 2A,B).

In general, it is likely that further decreases in SNR beyond –6 dB will prevent the neural extraction (~$P1_{TRF}$) and amplification (~$N1_{TRF}$) of the relevant features of the attended talker due to extensive energetic masking by the ignored talker. Ding and Simon (2013) estimated such a breakdown of the neural tracking of attended speech in noise to occur between –6 and –9 dB in MEG. Critically, the present data show that this selectivity might stay intact down to an SNR of –6 dB due to the support by additional neural mechanisms, as will be discussed in the next section.

Late distractor suppression by a non-auditory, supra-modal attention network

Under the adverse SNR of –6 dB compared to +6 dB, our analysis revealed an enhanced response to the ignored talker in the $P2_{TRF}$ time range consisting of a positive and a negative component (Fig. 3B). Together with the late increase in phase coherence (Fig. 3D), we interpret this additional component as a signature of active suppression of the ignored talker emerging from non-auditory, supra-modal regions, which are part of the fronto-parietal attentional or global-demand network (Woolgar et al. 2016).

Under the assumption that such active suppression is costly to the cognitive system, it has been suggested that it is only deployed if necessary (Chait et al., 2010). Neural signatures for active suppression of irrelevant signals during late (~200 ms) AEPs have been examined before (Melara et al., 2002; Chait et al., 2010). Pomper and Chait (2017) related enhanced centro-parietal activity in the theta band (4–7 Hz) to enhanced top-down control. None of these studies, however, had reported these signatures of suppression to originate from the dorsal anterior cingulate (dACC), a key region in adaptive control of effortful listening (e.g., Vaden et al., 2016; Erb et al., 2013) and showing here increased TRF phase coherence for hard-to-ignore speech (Fig. 3D).

Conclusions

The present data show how components of the unfolding temporal response function as identified in a forward encoding model reflect distinct neural stages of attentional filtering. These stages contain the initial, attention-independent encoding of acoustic signals; the extraction and amplification of relevant features; and lastly a robust, purely attention-driven response to attended acoustic signals. A phase-locked, active-suppression response to ignored acoustic signals originates from supra-modal attentional networks. In sum, with a design closer to real-life listening scenarios, our study provides insight into how selective neural processing of attended speech unfolds and is upheld under varying degrees of listening demand.

References

Bentler RA, Palmer C, Dittberner AB (2004) Hearing-in-Noise: Comparison of Listeners with Normal and (Aided) Impaired Hearing. 225:216–225.

Berens P (2009) CircStat: A MATLAB Toolbox for Circular Statistics. J Stat Softw 31.

Biesmans W, Das N, Francart T, Bertrand A (2016) Auditory-inspired speech envelope extraction methods for improved {EEG}-based auditory attention detection in a cocktail party scenario. {IEEE} Trans Neural Syst Rehabil Eng 4320:1--1.

Billings CJ, McMillan GP, Penman TM, Gille SM (2013) Predicting Perception in Noise Using Cortical Auditory Evoked Potentials. J Assoc Res Otolaryngol 14:891–903.

Cervantes Constantino F, Villafañe-Delgado M, Camenga E, Dombrowski K, Walsh B, Simon JZ (2017) Functional significance of spectrotemporal response functions obtained using magnetoencephalography. bioRxiv.

Chait M, Cheveigné A De, Poeppel D, Simon JZ (2010) Neuropsychologia Neural dynamics of attending and ignoring in human auditory cortex. Neuropsychologia 48:3262–3271.

Cherry EC (1953) Some Experiments on the Recognition of Speech, with One and with Two Ears. J Acoust Soc Am 25:975–979.

Chi T, Ru P, Shamma SA (2005) Multiresolution Spectrotemporal Analysis of Complex Sounds. J Acoust Soc Am 118:887–906.

Crosse MJ, Di Liberto GM, Bednar A, Lalor EC (2016) The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. Front Hum Neurosci 10:604.

Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. J Neurosci 23:3423–3431.

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr Biol 25:2457–2465.

Ding N, Chatterjee M, Simon JZ (2014) Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. Neuroimage 88:41–46.

Ding N, Simon JZ (2012a) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 107:78–89.

Ding N, Simon JZ (2012b) Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci 109:11854–11859.

Ding N, Simon JZ (2013) Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J Neurosci 33:5728–5735.

Drongelen W Van, Yuchtman M, Veen BD Van, Huffelen AC Van (1994) A Spatial Filtering Technique to Detect and Localize Multiple Sources in the Brain. 9:39–49.

Efron B (1979) Bootstrap methods: Another look at the jackknife. Ann Stat 7:1–26.

Erb J, Henry MJ, Eisner F, Obleser J (2013) The Brain Dynamics of Rapid Perceptual Adaptation to Adverse Listening Conditions. 33:10688–10697.

Fiedler L, Wöstmann M, Graversen C, Brandmeyer A, Lunner T, Obleser J (2017) Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. J Neural Eng 14.

Forte AE, Etard O, Reichenbach T (2017) The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. Elife:1–12.

Fuglsang SA, Dau T, Hjortkjær J (2017) NeuroImage Noise-robust cortical tracking of attended speech in real-world acoustic scenes. Neuroimage 156:435–444.

Giordano BL, Ince RAA, Gross J, Schyns PG, Panzeri S, Kayser C (2017) Contributions of local speech encoding and functional connectivity to audio-visual speech perception. :1–27.

Giuliano RJ, Karns CM, Neville HJ, Hillyard SA (2015) NIH Public Access. J Cogn Neurosci 26:2682–2690.

Hambrook DA, Tata MS (2014) Theta-band phase tracking in the two-talker problem. Brain Lang 135:52–56.

Henry MJ, Obleser J (2012) Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. PNAS 109:20095–20100.

Herrmann B, Henry MJ, Obleser J (2013) Frequency-specific adaptation in human auditory cortex depends on the spectral variance in the acoustic stimulation. J Neurophysiol 109:2086–2096.

Hertrich I, Dietrich S, Trouvain J, Moos A, 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.

Hoerl AE, Kennard RW (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12:55–67.

Horton C, Srinivasan R, Zmura MD (2014) Envelope responses in single-trial EEG indicate attended speaker in a " cocktail party ." J Neural Eng 11:12pp.

Kaya EM, Elhilali M (2017) Modelling auditory attention.

Kayser SJ, Ince RAA, Gross J, Kayser C (2015) Irregular Speech Rate Dissociates Auditory Cortical Entrainment, Evoked Responses, and Frontal Alpha. J Neurosci 35:14691–14701.

Kong YY, Mullangi A, Ding N (2014) Differential modulation of auditory responses to attended and unattended speech in different listening conditions. Hear Res 316:73–81.

Lachaux J, Rodriguez E, Martinerie J, Varela FJ (1999) Measuring Phase Synchrony in Brain Signals. 208:194–208.

Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE (2008) Entrainment of Neuronal Attentional Selection. Science (80-) 320:110–113.

Lalor EC, Power AJ, Reilly RB, Foxe JJ (2009) Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli. J Neurophysiol 102:349–359.

Lange K (2012) The N1 effect of temporal attention is independent of sound location and intensity : Implications for possible mechanisms of temporal attention. 49:1468–1480.

Makeig S, Debener S, Onton J, Delorme A (2004) Mining event-related brain dynamics. Trends Cogn Sci 8:204–210.

Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods 164:177–190.

Melara RD, Rao A, Tong Y (2002) The duality of selection: Excitatory and inhibitory processes in auditory selective attention. J Exp Psychol 28:279–306.

Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485:233–236.

Mirkovic B, Debener S, Jaeger M, Vos M De (2015) Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. J Neural Eng 12:46007.

Näätänen R, Gaillard AWK, Varey CA (1981) Attention effects on auditory EPs as a function of inter-stimulus interval. Biol Psychol 13:173–187.

O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC (2015) Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. Cereb Cortex 25:1697–1706.

Obleser J, Elbert T, Eulitz C (2004a) Attentional influences on functional mapping of speech sounds in human auditory cortex. BMC Neursoscience 9:1–9.

Obleser J, Lahiri A, Eulitz C (2004b) Magnetic Brain Response Mirrors Extraction of Phonological Features from Spoken Vowels. J Cogn Neurosci 16:31–39.

Obleser J, Wöstmann M, Hellbernd N, Wilsch A, Maess B (2012) Adverse Listening Conditions and Memory Load Drive a Common Alpha Oscillatory Network. 32:12376–12383.

Oostenveld R, Fries P, Maris E, Schoffelen J (2011) FieldTrip : Open Source Software for Advanced Analysis of MEG , EEG , and Invasive Electrophysiological Data. Comput Intell Neurosci 2011.

Paninski L, Pillow J, Lewi J (2007) Statistical models for neural encoding, decoding, and optimal stimulus design. In: Computational Neuroscience: Theoretical Insights into Brain Function (Cisek P, Drew T, Kalaska JF, eds), pp 493–507 Progress in Brain Research. Elsevier.

Peelle JE (2017) Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior. Ear Hear.

Petersen EB, Wöstmann M, Obleser J, Lunner T (2016) Neural tracking of attended versus ignored speech is differentially affected by hearing loss. J Neurophysiol 117:18–27.

Pichora-Fuller MK, Schneider BA, Daneman M (1995) How young and old listen to and remember speech in noise. J Acoust Soc Am 91:593–608.

Picton T (2013) Hearing in Time: Evoked Potential Studies of Temporal Processing. Ear Hear 34:385–401.

Picton TW, Hillyard SA (1974) Human auditory evoked potentials. ii: effects of attention. Electroencephalogr Clin Neurophysiol 36:191–199.

Pomper U, Chait M (2017) The impact of visual gaze direction on auditory object tracking. Sci Rep 7:1–16.

Power AJ, Foxe JJ, Forde EJ, Reilly RB, Lalor EC (2012) At what time is the cocktail party? A late locus of selective attention to natural speech. Eur J Neurosci 35:1497–1503.

Scherg M, Picton TW (1989) A Source Analysis of the Late Human Auditory Evoked Potentials. J Cogn Neurosci 1:336–355.

Smeds K, Wolters F, Rung M (2015) Estimation of Signal-to-Noise Ratios in Realistic Sound Scenarios. J Am Acad Audiol 196:183–196.

Spaak E, Lange FP De, Jensen O (2014) Local Entrainment of Alpha Oscillations by Visual Stimuli Causes Cyclic Modulation of Perception. J Neurosci 34:3536–3544.

Tavabi K, Obleser J, Dobel C, Pantev C (2007) Auditory evoked fields differentially encode speech features: an MEG investigation of the P50m and N100m time courses during syllable processing. Eur J Neurosci 25:3155–3162.

Veen BD Van, Drongelen W Van, Yuchtman M, Suzuki A (1997) Localization of Brain Electrical Activity via Linearly Constrained Minimum Variance Spatial Filtering. IEEE Trans Biomed Eng 44:867–880.

Willmore BDB, Cooke JE, King AJ (2014) Hearing in noisy environments: noise invariance and contrast gain control. J Physiol 16:3371–3381.

Woolgar A, Jackson J, Duncan J (2016) Coding of Visual, Auditory, Rule, and Response Information in the Brain: 10 Years of Multivoxel Pattern Analysis. J Cogn Neurosci 28:1433–1454.

Zar JH (1999) Biostatistical Analysis, 4th ed. Michigan: Prentice Hall.

Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE (2013) Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party.” Neuron 77:980–991.