Real-time decoding of selective attention from the human auditory brainstem response to continuous speech

Octave Etard[1,*], Mikolaj Kegler[1*], Chananel Braiman[2], Antonio Elia Forte[1] and Tobias Reichenbach[1,†]

[1]Department of Bioengineering and Centre for Neurotechnology, Imperial College London, South Kensington Campus, SW7 2AZ, London, U.K.

[2]Tri-Institutional Training Program in Computational Biology and Medicine, Weill Cornell Medical College, New York, NY 10065, U.S.A.

* These authors contributed equally to this work.

[†]To whom correspondence should be addressed (email: reichenbach@imperial.ac.uk)

Abstract

Humans are highly skilled at analysing complex auditory scenes. Previously we showed that the auditory brainstem response to speech is modulated by selective attention, a result that we achieved through developing a novel method for measuring the brainstem's response to running speech (Forte et al. 2017). Here we demonstrate that the attentional modulation of the brainstem response to speech can be employed to decode the attentional focus of a listener in real time, from short measurements of ten seconds or less in duration. The decoding is based on complex statistical models for extracting the brainstem response from multi-channel scalp recordings and subsequent classification of the model performances according to the focus of attention. We further show how a few recording channels as well as out-of-the-box decoding that employs population average models achieve a high accuracy from short recordings as well.

Introduction

Humans have an extraordinary capability to analyze crowded auditory scenes. We can, for instance, focus our attention on one of two competing speakers and understand her or him despite the distractor voice (Bronkhorst 2000). People with hearing impairment such as sensorineural hearing loss, however, face major difficulty with understanding speech in noisy environments, and this difficulty persists even when they wear auditory prosthesis such as hearing aids or cochlear implants (Kenyon et al. 1998). Auditory prosthesis could potentially aid with understanding speech in noise through selectively enhancing a target speech, for instance based on its direction, using algorithms such as beam forming (Kidd et al. 2015). However, such selective enhancement requires knowledge of which sound the user aims to attend to. Current research therefore attempts to decode an individual's focus of selective attention to sound from non-invasive brain recordings. If such decoding worked in real time, it could inform the sound processing in an auditory prosthesis. It could also form the basis of a non-invasive brain-computer interface for motor-impaired patients with brain injury, for instance, who may not be able to respond behaviourally. Moreover, such real-time decoding of selective attention could be employed clinically for a better understanding and characterization of hearing loss.

Neural activity in the cerebral cortex, especially in the delta (1 – 4 Hz) and theta (4 – 8 Hz) frequency bands, tracks the amplitude envelope of a complex auditory stimulus such as speech (Giraud & Poeppel 2012; Power et al. 2012; Ding & Simon 2014). The tracking is shaped by selective attention to one of several sound sources and can be measured noninvasively from magnetoencephalograpy (MEG) (Ding & Simon 2012), from electrocorticography (ECoG) (Mesgarani and Chang, 2012) as well as from the clinically more applicable electroencephalography (EEG) (Kerlin et al.,2010; Horton et al. 2013). Attention to one of two competing voices has been successfully decoded from single trials of one minute in duration using MEG (Ding & Simon 2012) as well as EEG (O'Sullivan et al. 2015; Mirkovic et al. 2015). Further optimization of the involved statistical modelling led to an accurate decoding of the focus of selective attention from still shorter recordings lasting less than 30 s (Biesmans et al. 2017; Van Eyndhoven et al. 2017).

Recently we showed that subcortical neural activity is consistently modulated by selective attention as well (Forte et al. 2017). To this end we developed a method to measure the response of the auditory brainstem to running non-repetitive speech. We employed empirical mode decomposition (EMD) to extract a nonlinear waveform from the speech signal that, at each time instance, oscillates at the fundamental frequency of the voice. We then correlated this fundamental waveform to the neural recording obtained from a few scalp electrodes. We observed a peak in the cross-correlation at a latency of 9 ms, evidencing a neural response at the fundamental frequency with a subcortical origin. When volunteers listened to two competing speakers, we then observed that the brainstem response to

the fundamental frequency of each speaker was larger when the speaker was attended than when she or he was ignored.

Because the brainstem response to speech that we measured occurs at the fundamental frequency of speech, typically between 100 – 300 Hz, it is ten- to hundredfold faster than the cortical tracking of the speech envelope. We therefore wondered if the brainstem response, despite its smaller magnitude than cortical responses, may allow to decode attention to one of several sounds in real time.

Results

We first measured neural responses to a single non-repetitive speech signal from 64-channel EEG. We employed empirical mode decomposition to obtain a fundamental waveform from the speech signal (Forte et al. 2017). We then employed linear regression with regularization to reconstruct the fundamental waveform from the multi-channel EEG data for each individual subject (linear backward model, Methods). The performance of the reconstruction was assessed through the correlation of the reconstructed fundamental waveform to the actual one.

We first verified that the linear backward model did extract a significant brainstem response to speech. To this end we also constructed models of the fundamental waveform of unrelated speech signals from the neural data. For almost all subjects that we assessed, the model that reconstructed the actual fundamental waveform significantly outperformed the one that attempted to reconstruct an unrelated fundamental waveform, showing that the former model was able to extract a meaningful brainstem response (Fig. 1A)

To investigate the spatio-temporal characteristics of the brainstem response we then computed a generic linear forward model that estimated the EEG recording from the fundamental waveform using the data from all the subjects (Haufe et al. 2014; Methods). The magnitude of the obtained complex coefficients peaked at 9 ms, demonstrating the subcortical origin of the neural activity and in agreement with previous recordings of speech-evoked brainstem responses (Fig. 1B; Skoe & Kraus 2010; Reichenbach et al. 2016; Forte et al. 2017). The magnitude of the coefficients at that latency showed major contributions from the mastoids as well as from the central scalp areas (Fig. 1C). The coefficients at the central area were approximately in antiphase to those near the mastoids (Fig. 1D). This topography accords with the dipolar nature of scalp-recorded auditory brainstem activity (Bidelman 2015; Ono et al. 1984; Grandori 1986; Norrix & Glattke 1996).

We then investigated how attention could be decoded from the brainstem response. Following a classic auditory attention paradigm we presented subjects with a male and a female voice diotically and simultaneously, instructing them to attend to either the male or the female speaker, while recording their neural activity from 64-channel EEG (Ding & Simon 2012; Forte et al. 2017). For

each subject, we computed four linear backward models. The first model, MA, reconstructed the fundamental waveform of the male voice when the subject attended to it. The second model, MI, reconstructed the fundamental waveform of the male speaker when the subject ignored it. Analogously, a third and fourth model, FA and FI, reconstructed the fundamental waveform of the female voice when it was attended or ignored, respectively. We observed that the performance of the two models that reconstructed the fundamental waveform of a speaker when they were attended was, in most subjects, significantly larger than that of the corresponding model for the ignored voice (Fig. 2). The average ratio between the reconstruction performance of the model for the attended male voice to that for the ignored male voice was 1.22, significantly larger than one ($p < 0.001$). The ratio was lower in the case of the female voice, 1.15, which was significantly above one as well ($p < 0.05$). The two ratios did not differ significantly ($p > 0.05$). The better reconstruction performance of the fundamental waveform of an attended speech signal demonstrates the attentional modulation of the brainstem response to speech that we described previously (Forte et al. 2017).

Having verified the attentional modulation of the brainstem response to speech using high-density EEG recordings and linear backward models, we sought to investigate whether this approach could be used to decode auditory attention in real time. We therefore applied the four linear backward models MA, MI, FA and FI to unseen data that had not been used to train the linear models. We expected the focus of attention to emerge, for instance, from the difference in the performances of the models MA and FA. This difference should typically be positive when the subject attended to the male voice, and be negative otherwise. Similarly, attention could potentially be decoded from the difference of the reconstruction performance of the models FI and MI. A subject's attention to the male voice should mostly lead to a positive difference, and a focus on the female voice to a negative difference. A more complex approach can take the performance of these four models into account, and we accordingly trained a Support Vector Machine (SVM) algorithm to classify the focus of attention (Fig. 3A).

We tested the accuracy of the decoding on samples of a duration that varied from half a second to 60 seconds (Fig. 3B). For durations of less than 30 seconds, the SVM achieved a higher accuracy than the classification based on the models MA and FA for the attended voices only. In particular, the accuracy remained significantly above chance even for very short samples that lasted only half a second. It was, for instance, 62% and 77% and for one-second respectively ten-second samples. In contrast, the models MI and FI by themselves still allowed for a decoding of the attentional focus with an accuracy that was better than chance, but below that of the other two approaches. Because of a smaller set of samples with a duration above 30 seconds, the accuracy of the SVM approached that of the simpler decoding based on the models MA and FA for such samples, and reached 91% for a 60-second sample.

Practical applications of the decoding of auditory attention benefit from a small number of required recording channels. We therefore investigated how well the developed decoding works if the linear backward models use only three EEG channels, the left and right mastoid as well as the central channel Cz. Strikingly, the decoding accuracy was barely smaller than that of the 64-channel model; for instance, it remained at an impressive 77% for a short ten-second sample and when the SVM was used as classifier (Fig. 3C).

The decoding described above utilized linear backward models that were subject specific and hence required prior training from EEG recordings for each individual. Such subject-specific training may, however, not always be available. We thus assessed how well a linear backward model that was trained on the whole population of subjects, and thus represented an average model that could be used out-of-the-box, allowed to decode attention. As expected, the decoding accuracies were then lower than those for the subject-specific models, but nonetheless remained higher than chance level for short durations of one second or less (Fig. 3D, E). In particular, the accuracy was 73% for a ten-second sample when the SVM was used as the classifier and when 64 EEG channels were used. It dropped only very slightly to 72% for three EEG channels. The SVM outperformed the other two classifiers for all sample durations, presumably due the larger amount of training data that was available compared to the subject-specific models. The superior performance of the SVM classifier showed that the models MI and FI for the ignored voices contributed relevant information to the attentional decoding, although decoding based on them alone did not achieve high accuracy.

Discussion

We have shown that attention to one of two competing speakers can be decoded accurately from the brainstem response to natural speech, from data of a short duration of ten seconds or less. The decoding is best when high-density EEG data is available, as well as when models that relate the neural recording to the speech signal are computed for each subject individually. However, for short samples, attention could be decoded from only three EEG channels as accurately as from 64 channels, evidencing the potential of the obtained results for clinical and technological applications. Decoding auditory attention to steer an auditory prostheses or a brain-computer interface will indeed benefit from a small number of required recoding channels. Conversely, subject-specific models may cause difficulty in practice as sufficient training data per subject may not always be obtainable. We have shown, however, that even three-channel generic out-of-the-box models yield a high decoding accuracy. The out-of-the box models reflect the average over many subjects and that can be readily applied to other subjects for which no training data is available.

The decoding that we have described here is based on linear backward models that reconstruct the fundamental waveform of the speech signal from the EEG recordings. Improved performance may be obtained through canonical correlation analysis that relates the neural recording

to more speech features in an optimized space (de Cheveigne et al. 2017) or through an artificial neural network that is able to extract highly nonlinear relations between the two datasets (Yang et al 2015).

The correlation between a reconstructed fundamental waveform and the actual one is small, typically between 0.05 and 0.1 (Fig. 1A, Fig. 2). Cortical responses allow to reconstruct the speech envelope from EEG recordings and yield somewhat higher correlation coefficients. However, decoding of attention from short amounts of data based on cortical responses shows similar performance to the decoding based on brainstem responses that we have described here. We attribute this to the rapidness of the brainstem response to the fundamental frequency of speech that is ten- to hundredfold faster than the cortical response to the speech envelope. Although smaller in magnitude, the brainstem response has therefore a higher information rate that enables real-time decoding of attention. Moreover, the framework for attentional decoding based on the brainstem response to running speech that we have developed here allows integration with cortical responses to the speech envelope which may boost the decoding accuracy even further.

Methods

Participants. 18 healthy adult English native speakers (aged 22.8 ± 1.9 year, four females), with no history of auditory or neurological impairments participated in the study. All participants provided written informed consent. The experimental procedures were approved by the Imperial College Research Ethics Committee.

Experimental design and stimuli. We employed the same experimental design that we used previously to measure the brainstem response to non–repetitive speech and its modulation through selective attention (Forte et al. 2017). Briefly, ten-min long continuous speech samples from a male and female speaker were obtained from publicly available audiobooks (librivox.org). One sample from the female speaker was used when presenting speech in quiet. Two distinct samples from each speaker were mixed at equal root-mean-square amplitude to produce stimuli with two competing speakers.

Participants first listened to a single speaker without background noise. They then listened to two stimuli with two competing speakers each. They were instructed to attend either the male or female voice in the first stimulus, and to attend to the speaker they previously ignored in the second one. The presentation order was decided at random for each subject. Each stimulus was presented in four parts of approximately equal duration, and comprehension questions were asked after each part. All stimuli were delivered diotically at 76 dB(A) SPL (A-weighted frequency response).

Neural data acquisition and processing. Neural activity was recorded at 1 kHz through a 64-channel scalp EEG system using active electrodes (actiCAP, BrainProducts, Germany) and a multi-channel

EEG amplifier (actiCHamp, BrainProducts, Germany). The electrodes were positioned according to the standard 10-20 system and referenced to the right earlobe. The EEG recordings were band-passed filtered between 100 and 300 Hz (low pass: linear phase FIR filter, cutoff (-6 dB) 325 Hz, transition bandwidth 50 Hz, order 66 ; high pass: linear phase FIR filter, cutoff (-6 dB) 95 Hz, transition bandwidth 10 Hz, order 364 ; both: one-pass forward and compensated for delay) and then referenced to the average. When using three channels for the decoding only, all channels except the two mastoids TP9 and TP10 and the vertex Cz were discarded before filtering the recordings as described above. The audio signals were simultaneously recorded by the amplifier through an acoustic adapter (Acoustical Stimulator Adapter and StimTrak, BrainProducts, Germany), and were used to align the neural responses and stimuli. A 1 ms delay of the acoustic signal introduced by the earphones was taken into account.

Computation of the fundamental waveform of speech. The fundamental waveform of each speech sample was computed using empirical mode decomposition as described previously (Forte et al. 2017). It was then downsampled to 1 kHz, the sampling rate of the neural recordings, and filtered as described above. Silent or unvoiced parts of the speech produced some segments where the fundamental waveform was equal to zero. For the stimuli with a single speaker, we excluded such segments from the further analysis. For the stimuli with two competing speakers we excluded the few segments where the fundamental waveform of one of the two voices was entirely zero as attention could not be decoded in this case.

Backward model. We first used a linear spatio-temporal backward model to reconstruct the fundamental waveform of speech from the neural recordings. Specifically, at each time instance t_n, the fundamental waveform $y(t_n)$ was expressed as a linear combination of the neural recordings $x_j(t_n + \tau_k)$ as well as their Hilbert transform $x_j^h(t_n + \tau_k)$ at a delay τ_k:

$$y(t_n) = \sum_{j=1}^{N} \sum_{k=1}^{T} \left[\beta_{j,k}^{(r)} x_j(t_n + \tau_k) + \beta_{j,k}^{(i)} x_j^h(t_n + \tau_k) \right]. \quad (1)$$

The index *j* refers hereby to the recording channel, and $\beta_{j,k}^{(r)}, \beta_{j,k}^{(i)}$ are a set of real coefficient to determine. The Hilbert transform of each recording channel was included to allow the reconstruction of the fundamental waveform from these signals as well. The model's coefficients can be assembled into complex coefficients $\beta_{j,k} = \beta_{j,k}^{(r)} + i\beta_{j,k}^{(i)}$ that encode accordingly the amplitude of the brainstem response, the temporal delay as well as the phase difference between stimulus and response. Following typical latencies of auditory brainstem responses, we used lags from -5 ms to 19 ms with an increment of 1 ms. We thus obtained T=25 temporal delays that, together with the N=64 recording channels, led to 1,600 complex model coefficients. The model coefficients were then estimated using a regularised ridge regression as $\beta = (X^t X + \lambda I)^{-1} X^t Y$, in which X is the design matrix and λ is the regularisation parameter (Hastie et al. 2009). A five-fold cross-validation procedure was

implemented, and the regularisation coefficient that produced the highest Pearson's correlation coefficient between the reconstructed fundamental waveform and the target one was selected. The model's performance was quantified through the obtained correlation coefficient.

Significance of the auditory brainstem response. To determine if the linear backward models showed a significant brainstem response to the fundamental frequency, we also computed noise models as linear backward models that attempted to reconstruct the fundamental waveform of an unrelated speech segment that was not heard by the participant. We then compared the performances of the correct linear backward model and the noise model through a two-sample and two-tailed Student's t-test. The results of the statistical tests are indicated in Fig. 1 through asterisks: no asterisk is given when results are not significant ($p > 0.05$), one asterisk when results are significant (*, $0.01 < p < 0.05$), two asterisks when significance is high (**, $0.001 < p < 0.01$), and three asterisks when significance is very high (***, $p < 0.001$).

Estimation of the neural response (forward model). We also computed a linear forward model that relates the EEG recording $x_j(t_n)$ at time t_n to the fundamental waveform $y(t_n - \tau_k)$ as well as its Hilbert transform $y^h(t_n - \tau_k)$ at a delay τ_k:

$$x_j(t_n) = \sum_{k=1}^{T} \left[\alpha_k^{(r)} y(t_n - \tau_k) + \alpha_k^{(i)} y^h(t_n - \tau_k) \right], \quad (1)$$

in which $\alpha_k^{(r)}$ and $\alpha_k^{(i)}$ are the model's real coefficients. They can be interpreted as real and imaginary parts of the complex coefficients $\alpha_k = \alpha_k^{(r)} + i\,\alpha_k^{(i)}$. The model coefficients were estimated by pooling the data from all subjects together and using a regularised ridge regression. The coefficients of this forward model, but not those of the backward model, allow for a neurobiological interpretation of their spatio-temporal characteristics (Haufe et al. 2014).

Attentional modulation of the auditory brainstem response. To analyze the attentional modulation of the brainstem response to one of two competing speakers, we computed two pairs of backward models for each subject using five-fold cross-validation. The first pair of models reconstructed the fundamental frequency of the male voice while it was either attended or ignored. The second pair of models reconstructed the fundamental waveform of the female voice (attended or ignored). For each pair of models the cross-validated performances were then compared using a two-sample and two-tailed Student's t-test. The results are indicated in Fig. 3 through asterisks as described above. We further employed a one-sample one-tailed Student's t test to verify that the ratio of the performances was significantly larger than one, and a two-sample two-tailed Student's t test to check if the ratios obtained from the responses to the male voice and to the female voice were significantly different.

Decoding of auditory attention. We investigated how attention could be decoded from short segments of neural data that were obtained in response to competing speakers. We first assessed the

performances of the two pairs of subject-specific linear backward models on the short EEG data, using five-fold cross-validation and yielding four different performances. These were then employed to decode attention either by comparing the performances of the models for the attended male and the attended female voice, or by comparing the performances of the models for the ignored male and the ignored female voice. We also took all four performances into account through training a Support Vector Machine algorithm (linear kernel). The decoding accuracy was computed from fivefold cross-validation. Since we had a finite number of samples, we computed the 95% chance level, that is, the accuracy that a random classifier would at most achieve in 95% of cases, using a binomial distribution (Combrisson & Jerbi 2015).

For the out-of-the-box models, we trained linear backward models on the pooled data from all subjects and quantified their performances using a leave-one-subject-out cross-validation coupled with a five-fold cross-validation regarding the auditory stimuli. In particular, the models were trained on all subjects except for one, and on four out of five folds of the stimulus data. The model performances were then evaluated on the remaining subject and on the remaining stimulus fold. The classification accuracies of the three decoding methods were then evaluated using a five-fold cross-validation as described above.

Both the decoding using subject-specific linear backward models as the one using out-of-the-box models were implement using 64 as well as three EEG channels.

Acknowledgement

We thank Steve Bell, Karolina Kluk-de Kort, Patrick Naylor, David Simpson, Alain de Cheveigne and Malcolm Slaney for discussion. This research was supported by EPSRC grant EP/M026728/1 to T.R. by the Royal British Legion Centre for Blast Injury Studies, as well as in part by the National Science Foundation under Grant No. NSF PHY-1125915.

Competing financial interests

The authors declare no competing financial interests.

References

Baken, R.J. & Orlikoff, R.F., 2000. Clinical measurement of speech and voice. *Singular Thomson Learning*

Bidelman, G.M., 2015. Multichannel recordings of the human brainstem frequency-following response: scalp topography, source generators, and distinctions from the transient ABR. *Hear. Res.*, 323: 68-80

Biesmans, W., Das, N., Francart, T. & Bertrand, A., 2017. Auditory-inspired speech envelope extraction methods for improved eeg-based auditory attention detection in a cocktail party scenario. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 25: 402-412

Bronkhorst, A. W., 2000. The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions. *Acustica*, 86: 117-128

Cheveigne, A., Wong, D., Liberto, G. Di, Hjortkjaer, J., Slaney, M., & Lalor, E. (2017). Decoding the auditory brain with canonical component analysis. bioRxiv, 217281.

Ding, N. & Simon, J.Z., 2012. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci. U. S. A.*, 109: 11854–11859

Ding, N. and Simon, J. Z.,2014. Cortical entrainment to continuous speech: functional roles and interpretations. *Front. Hum. Neurosci.,* 8: 311

Combrisson, E. & Jerbi, K., 2015. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *J. Neurosci. Methods,* 250: 126-136

Forte, A.E., Etard, O. & Reichenbach, T., 2017. The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *Elife*, 6: e27203

Giraud, A-L. & Poeppel, D., 2012. Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.,* 15: 511-517

Grandori, F., 1986. Field analysis of auditory evoked brainstem potentials. *Hear. Res.*, 21: 51-58

Hashimoto, I., Ishiyama, Y., Yoshimoto, T. & Nemoto, S., 1981. Brain-stem auditory-evoked potentials recorded directly from human brain-stem and thalamus. *Brain*, 104: 841–859

Hastie, T., Tibshirani, R., & Friedman, J., 2009. The Elements of Statistical Learning. *Springer Series in Statistics*

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., & Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87: 96-110

Horton, C., D'Zmura, M., Srinivasan, R. 2013. Suppression of competing speech through entrainment of cortical oscillations. *J. Neurophysiol.* 109: 3082-3093.

Horton, C., Srinivasan, R. & D'Zmura, M., 2014. Envelope responses in single-trial EEG indicate attended speaker in a "cocktail party." *J. Neural Eng.*, 11: 046015

Johnson, K.L., Nicol, T.G. & Kraus, N., 2005. Brain stem response to speech: a biological marker of auditory processing. *Ear Hear.*, 26: 424-434

Kenyon, E.L., Leidenheim, S.E., Zwillenberg, S.,1998. Speech discrimination in the sensorineural hearing loss patient: how is it affected by background noise? *Mil. Med.*, 163: 647-650

Kerlin, J.R., Shahin, A.J., Miller, L.M. 2010. Attentional gain control of ongoing cortical speech representations in a "cocktail party". *J. Neurosci.* 30: 620-628

Kidd, G., Mason, C. R., Best, V., & Swaminathan, J., 2015. Benefits of Acoustic Beamforming for Solving the Cocktail Party Problem. *Trends in Hearing* 19: 1-15

Mesgarani, N. & Chang, E.F., 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: 046007

Norrix, L.W. & Glattke, T.J., 1996. Multichannel waveforms and topographic mapping of the auditory brainstem response under common stimulus and recording conditions. *J. Commun. Disord.*, 29: 157-182

O'Sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Slaney, M., Shamma, S.A. & Lalor, E.C., 2015. Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cereb. Cortex*, 25: 1697-1706

Ono, I., Ichikawa, G., Yosikawa, H., Kato, E. & Fukuda, M., 1984. The scalp topography of abr. *Audiol. Japan*, 27: 292-299

Picton, T.W., Stapells, D.R. & Campbell, K.B., 1981. Auditory evoked potentials from the human cochlea and brainstem. *J. Otolaryngol. Suppl.*, 9: 1-41

Power ,A.J., Foxe, J.J., Forde, E.J., Reilly, R.B., Lalor, E.C. 2012 At what time is the cocktail party? A late locus of selective attention to natural speech. *Eur. J. Neurosci.* 35: 1497-1503

Reichenbach, C.S., Braiman, C., Schiff, N.D., Hudspeth, A.J. & Reichenbach, T., 2016. The auditory-brainstem response to continuous, non-repetitive speech is modulated by the speech envelope and reflects speech processing. *Front. Comp. Neurosci*. 10: 47

Skoe, E. & Kraus, N., 2010. Auditory brain stem response to complex sounds: a tutorial. *Ear Hear.*, 31: 302-324

Sohmer, H., Pratt, H. & Kinarti, R., 1977. Sources of frequency following responses (ffr) in man. *Electroencephalogr. Clin. Neurophysiol.*, 42: 656-664

Titze, I.R. & Martin, D.W., 1998. Principles of voice production. *National Center for Voice and Speech*

Van Eyndhoven, S., Francart, T., & Bertrand, A., 2017. EEG-Informed Attended Speaker Extraction from Recorded Speech Mixtures with Application in Neuro-Steered Hearing Prostheses. *IEEE Trans. Biomed. Eng.*, 64: 1045-1056

Yang, M.; Sheth, S. A.; Schevon, C. A.; II, G. M. M., Mesgarani, N. (2015), Speech reconstruction from human auditory cortex with deep neural networks., *INTERSPEECH* , ISCA

Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., Poeppel, D. & Schroeder, C.E., 2013. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." *Neuron*, 77: 980-99

Figures

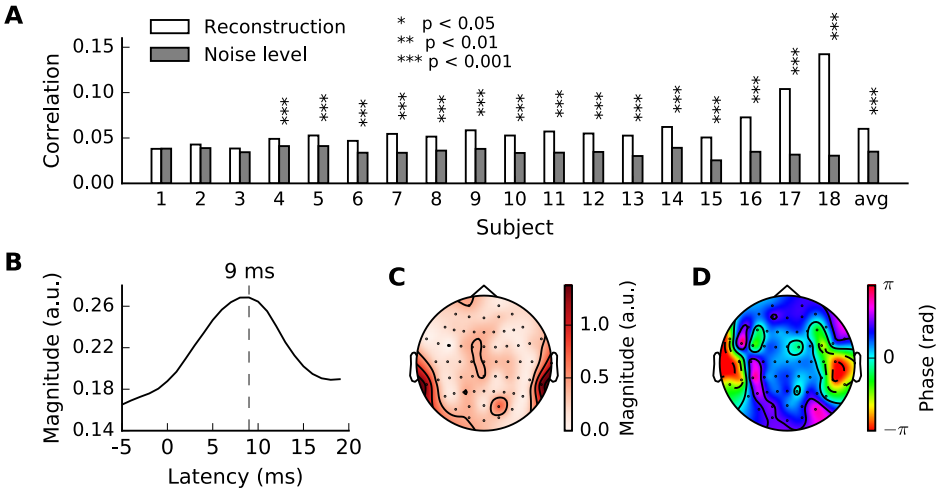

Figure 1. The brainstem response to natural speech detected from high-density EEG recordings using a linear model. (**A**) The performance of the linear backward model is assessed through the correlation of the reconstructed fundamental waveform and the actual one. The performance is significantly larger in almost all subjects than that of a noise model. Subjects have been ordered by increasing significance. (**B**) The magnitude of the coefficients of the forward model obtained from all subjects and averaged over all EEG channels peaks at a latency of 9 ms. (**C**) At the delay of 9 ms, the channels from the left and right temporal scalp areas as well as from the central area have model coefficients with the largest magnitudes. (**D**) The phase of the coefficients at the delay of 9 ms shows a phase difference of around π between the temporal areas and the central one.

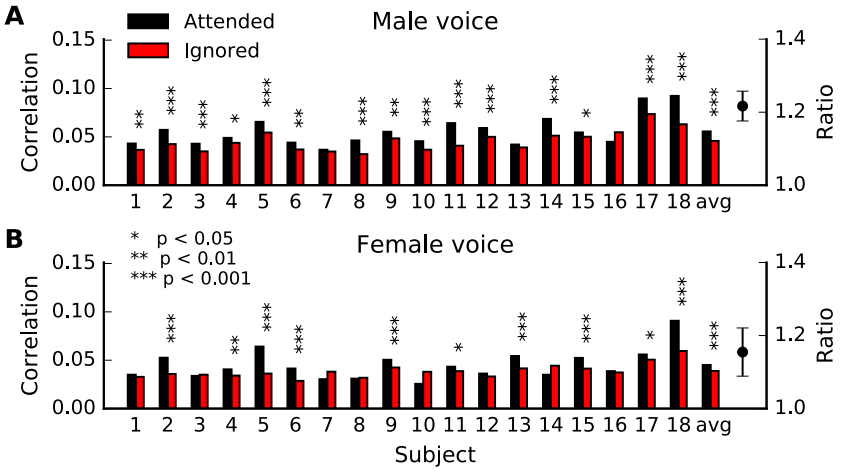

Figure 2. Attentional modulation of the auditory brainstem response to natural speech. The order of the subjects is as in Fig. 1A. (**A**) The performance of the linear backward model for the male voice is larger when the male speaker is attended (black) then when he is ignored (red). The two performances differ significantly in most subjects, and so do the two average performances (avg). The average ratio between the two performances is 1.22 and is significantly larger than one. (**B**) The performance of the linear backward model that reconstructs the fundamental waveform of the attended female voice is likewise significantly larger than that for the ignored female voice in most subjects, as well as on average (avg). The average ratio of the two performances is 1.15, significantly larger than one.

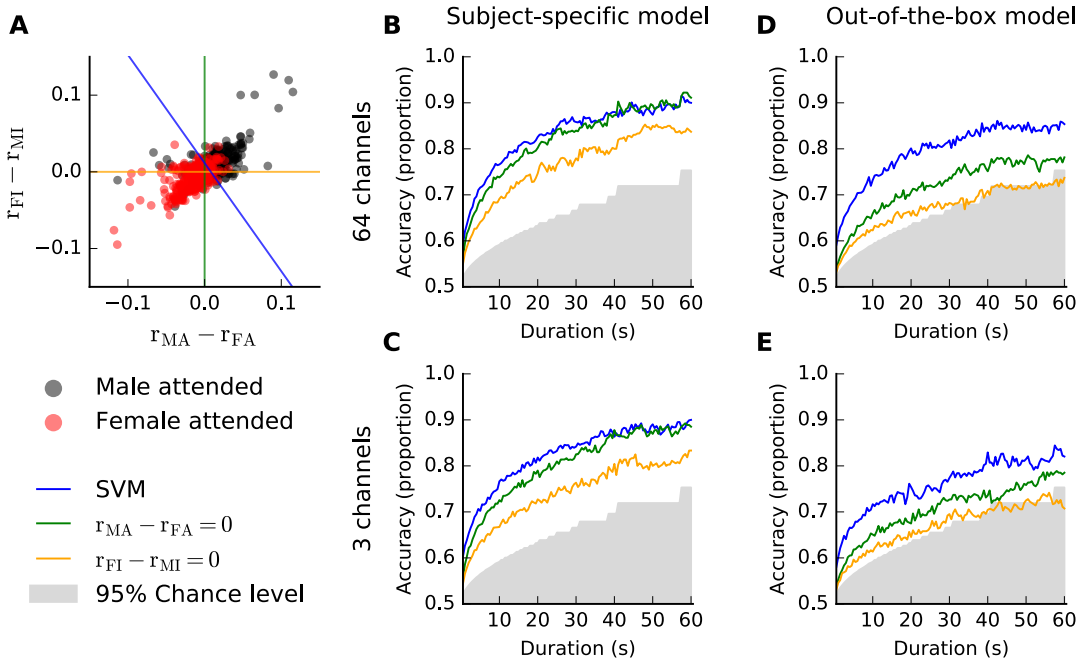

Figure 3. Real-time decoding of auditory attention. (**A**) Short data of a duration of 30 s that were obtained from a subject listening to the male speaker (black) can be discriminated from those obtained when a subject listened to the female voice (red) through the performances *r* from four linear backward models (MA, MI, FA, FI; main text). The classification can employ the difference in the performances between the models MA and FA (green), the difference between the models FI and MI (orange), or both (Support Vector Machine or SVM, blue). (**B**) The average decoding accuracy obtained from the models MA and FA reaches 91% at a duration of 60 seconds, and remains above chance level (grey) for very short durations of 500 ms. (**C**) Employing only three recording channels to decode attention reduces the performance of the three classifiers only slightly, if at all. (**D, E**) Employing 64 recording channels (**D**) or 3 channels (**E**), but out-of-the-box models for reconstructing the fundamental waveforms of the different speech signals, leads to only slightly reduced decoding accuracies.