Bootstrapping EEG-based auditory attention detection systems: boundary conditions for background noise and speaker positions

Neetha Das[†⋆], Alexander Bertrand[†], Tom Francart[⋆]

Abstract

Objective. A listener's neural responses can be decoded to identify the speaker the person is attending to in a cocktail party environment. Such auditory attention detection methods have the potential to provide noise suppression algorithms in hearing devices with information about the listener's attention. A challenge is the effect of noise and other acoustic conditions that can reduce the attention detection accuracy. Specifically, noise can impact the ability of the person to segregate the sound sources and perform selective attention, as well as the external signal processing necessary to decode the attention effectively. The aim of this work is to systematically analyze the effect of noise level and speaker position on attention decoding accuracy. *Approach.* 28 subjects participated in the experiment. Auditory stimuli consisted of stories narrated by different speakers from 2 different locations, along with surrounding multi-talker background babble. EEG signals of the subjects were recorded while they focused on one story and ignored the other. The strength of the babble noise as well as the spatial separation between the two speakers were varied between presentations. Spatio-temporal decoders were trained for each subject, and applied to decode attention of the subjects from every 30s segment of data. Behavioral speech recognition thresholds were obtained for the different speaker separations. *Main results.* Both the background noise level and the angular separation between speakers affected attention decoding accuracy. Remarkably, performance was seen to increase with the inclusion of moderate background noise (versus no noise), while across the different noise conditions performance dropped significantly with increasing noise level. We also observed that decoding accuracy improved with increasing speaker separation, exhibiting the advantage of spatial release from masking. Furthermore, the effect of speaker separation on the decoding accuracy became stronger when the background noise level increased. A significant correlation between speech intelligibility and attention decoding accuracy was found across conditions. *Significance.* This work shows how the background noise level and relative positions of competing talkers impact attention decoding accuracy. It indicates in which circumstances a neuro-steered noise suppression system may need to operate, in function of acoustic conditions. It also indicates the boundary conditions for bootstrapping EEG-based attention detection systems in neuro-steered hearing prostheses.

Index Terms

Auditory attention detection, EEG processing, neuro-steered auditory prostheses, brain-computer interface, cocktail party, acoustic conditions.

The work is funded by KU Leuven Special Research Fund C14/16/057 and OT/14/119, FWO project nrs. 1.5.123.16N and G0A4918N, the ERC (637424) under the European Union's Horizon 2020 research and innovation programme, and a research gift of Starkey Hearing Technologies. The scientific responsibility is assumed by its authors.

[†] KU Leuven, Dept. Electrical Engineering (ESAT), Stadius Center for Dynamical Systems, Signal Processing and Data Analytics. Kasteelpark Arenberg 10, B-3001 Leuven, Belgium.

[⋆] KU Leuven, Dept. Neurosciences, ExpORL. Herestraat 49 bus 721, B-3000 Leuven, Belgium.

I. Introduction

Hearing aids have noise reduction algorithms that help improve speech perception by reducing undesirable noise. They however lack an important piece of information - the intent of the user. A multi-speaker environment with a lot of background noise is a challenging situation for a hearing impaired person who wants to listen to one speaker, and ignore the rest of the auditory stimuli present. Although a hearing aid can help reduce the surrounding noise, most hearing aids are not 'smart' enough to work in synchrony with the user's intent. Hearing devices equipped with multiple microphones are capable of using beamforming techniques to improve the signal-to-noise ratio (SNR) in a given direction, but this information about the desired direction is not available.

It is well established that a listener's cortical responses entrain to the envelopes of the speech streams present, and that the entrainment is stronger for the attended speech stream (Ding & Simon 2012*a*, Golumbic et al. 2013). Auditory attention detection (AAD) algorithms allow to decode neural responses to auditory stimuli to detect the speaker a person is attending to, based on the auditory sources present, and the electroencephalogram (EEG) or magnetoencephalogram (MEG) of the person. Recent years saw progress in this area such as attention detection based on forward and backward modeling using various speech representations such as envelopes (O'Sullivan et al. 2014), spectrograms, phonetic features etc (Di Liberto et al. 2015), addressing issues such as the number of electrodes (Mirkovic et al. 2015), electrode placements (Fiedler et al. 2016, Mirkovic et al. 2016), the length of the data window to decode attention (Mirkovic et al. 2015, Zink et al. 2017), the use of concealable miniature devices (Abhijith & Bertrand 2018, Fiedler et al. 2016, Mirkovic et al. 2016), auditory-inspired envelopes (Biesmans et al. 2016) etc. We also saw the application of neural network based approaches to decode attention (Taillez et al. 2017, Wong et al. 2018). Attention decoding has also been explored using canonical correlation analysis based decoders (de Cheveigné et al. 2018) as well as state-space estimators of attentional state rather than the more often used correlation measures (Akram et al. 2016, Miran et al. 2017). These developments bring us closer towards achieving the goal of faster attention detection which is a necessity if neuro-steered hearing aids are to become a reality. In addition, there has been some work towards building an AAD integrated noise suppression system (Das et al. 2017, O'Sullivan et al. 2017, Van Eyndhoven et al. 2017).

While work on algorithms for attention decoding and its integration into noise-suppression systems advance, there is a need for insights into their performance in realistic auditory scenarios. From the work by Das et al. (2017), we have seen that from an audio signal processing point of view, good attention decoding accuracies, as well as output signal-to-noise ratios are possible for a range of background noise levels and speaker positions. However, Das et al. (2017) used EEG data from recordings where the stimuli did not include background noise and with a maximal spacing between the two competing speakers, and hence the outcome could be considered rather optimistic. Since we rely on the SNR of the neural responses to perform attention decoding, it is important to have a systematic understanding of the impact of changing noise level and speaker positions on the quality of these neural responses. The boundary conditions for factors like noise level and speaker separation necessary for bootstrapping attention detection systems are still unknown. Fuglsang et al. (2017) analyzed the effect of reverberation and number of competing talkers, showing that cortical activity tracked the attended speech stream and the tracking remained robust

irrespective of the realistic acoustic environments used in the study. It was also found that more realistic stimuli (filtered by head-related transfer functions) result in better attention decoding performance compared to dichotically presented stimuli (Das et al. 2016). There has also been research into how attention decoding is impacted when unprocessed binaural signals are used as reference signals in different acoustic environments (reverberant and/or noisy) (Aroudi & Doclo 2017).

In this study, we focus on the impact of background babble noise as well as relative separation between the two competing talkers, on AAD. We used negative signal to noise ratios that can be expected to result in significantly different speech intelligibilities, as well as relative speaker positions which can be perceived to be easy to hard in terms of speech intelligibility and the effort necessary to focus on one speaker and ignore the other. Finally, drawing inspiration from the work of Vanthornhout et al. (2018), we looked into the relationship between speech intelligibility and attention decoding accuracies.

II. Methods

A. Data Acquisition

Twenty-two female and six male subjects between the ages of 20 and 25 years participated in the experiment. All subjects were native Dutch speakers. Pure-tone audiometry was performed by which all subjects were found to be normal-hearing (thresholds < 20 dB HL from 125 Hz to 8000 Hz). Every subject signed an informed consent form approved by the local ethical committee.

During the experiment, 64-channel EEG data at a sampling rate of 8192 Hz was recorded using the BioSemi ActiveTwo system. The speech stimuli (competing talkers) were administered to the subject at 65 dBA through Etymotic ER1 insterphones. The experiments were conducted in a soundproof, electromagnetically shielded room. We used the APEX 3 program developed at ExpORL (Francart et al. 2008) to present sounds and an interactive screen to the subjects where they could use a mouse to navigate through the experiment.

B. Stimuli and experiment design

We chose Dutch short stories narrated by male speakers as stimuli. In the audio files, silences longer than 300 ms were shortened to 300 ms. For each story, the root mean square (RMS) value of the speech material was computed for every 0.5 second frame, which formed the vector of RMS values. The RMS intensity over time of the stories was smoothed so that there were no large variations in the sound pressure level of the stimuli over time. This was done to avoid differences in speech intelligibility and also to prevent a shift in attention of the subject if and when significant sound pressure level differences occur between the attended and the unattended speakers. All stories were then normalized to the same overall RMS intensity.

We made 9 babble noise files by combining speech streams from 2 unique male and 2 unique female speakers at a time. Head-related transfer functions (HRTFs) (Kayser et al. 2009) were used to filter each of these babble noise streams such that babble noise was perceived to come from 9 equidistant (from -180 degrees to 140 degrees in steps of 40 degrees) locations around the subject. The long-term spectrum of the babble noises was matched to the average spectrum of all the stories used in the experiments. The overall effect to the subject was that of a 36-talker

Presentation	Speaker positions (degrees)	Attended direction (degrees)	Attended Stimulus	Ignored Stimulus	SNR (dB)
1	-5,5	-5	StoryA, part1	StoryB, part1	20
2	-5,5	-5	StoryA, part2	StoryB, part2	-4.1
3	-5,5	-5	StoryA, part3	StoryB, part3	-1.1
4	-5,5	-5	StoryA, part4	StoryB, part4	-7.1
5	-5,5	5	StoryB, part1	StoryA, part1	20
6	-5,5	5	StoryB, part2	StoryA, part2	-4.1
7	-5,5	5	StoryB, part3	StoryA, part3	-1.1
8	-5,5	5	StoryB, part4	StoryA, part4	-7.1

TABLE I: An example of the design of an experiment where speech streams were presented from -5 degrees (left) and 5 degrees (right) with respect to the frontal direction of the subject.

Presentation	Speaker separation (degrees)	Duration (minutes)
1 & 5	180	4
2 to 4 & 6 to 8	180	5
1 & 5	60	2
2 to 4 & 6 to 8	60	2.5
1 & 5	10	4
2 to 4 & 6 to 8	10	5

TABLE II: Duration of each presentation for different angular separations between speakers.

(9 directions X 4 speakers) surrounding babble. We define an experiment here as a sequence of 8 presentations. An example for the design of an experiment is shown in Table I. Within an experiment, each presentation consisted of the subject being presented with two story parts at 65 dBA - each from a particular direction with respect to the subject, as well as surrounding babble noise at a particular signal to noise ratio (SNR) with respect to the power of the speakers. The screen in front of the subjects showed a figure with indication towards the direction of the speech stream to which the subjects were expected to focus their attention. Each subject participated in 4 experiments where the speaker positions were at -90 (extreme left) and 90 (extreme right) degrees, 30 and 90 degrees, -30 and -90 degrees, and -5 and 5 degrees. Thus, the design allowed to analyze situations where the angular separation between the speakers were 180 degrees, 60 degrees and 10 degrees respectively.

Before the first presentation, the subject was presented with a test screen and a 5 s audio of only the attended speaker so that the subject could get an idea of the attended speaker's voice and direction before the presentations began. In the first presentation, the two speakers were presented to the subject with no background babble. In the following sections, we will be referring to this as the 'no noise' condition. For convenience, the 'no noise' condition

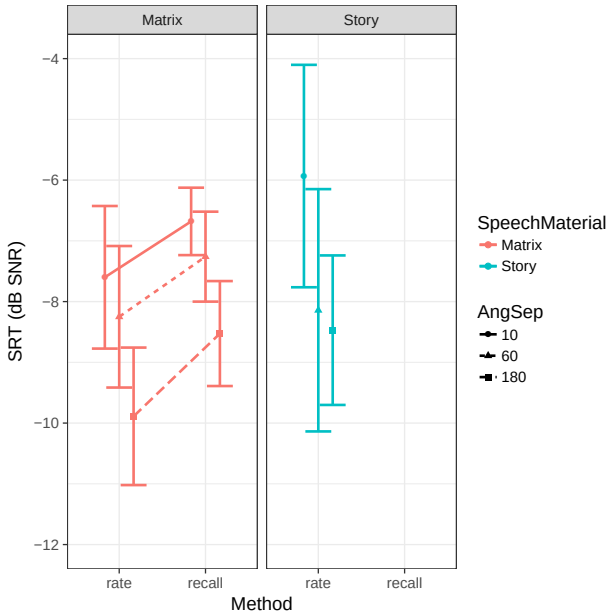

Fig. 1: Speech recognition thresholds for different speech material. Offset between the mean SRTs of the 'recall' and 'rate' methods using Matrix speech material is used as a correction for the mean 'rate' SRT when story speech material is used.

is shown as 20 dB SNR in all the tables and figures. In the following 3 presentations, the speakers were presented along with surrounding babble at 3 different SNRs: -1.1 dB, -4.1 dB and -7.1 dB. After the first 4 presentations, the subject was again presented with a test screen, this time accompanied by indication and a 5 s audio of the other speaker. For presentations 5 to 8, the subject was instructed to focus his attention on the other speaker. The order of SNRs, the presented stories and the order of attended direction were randomized between experiments as well as subjects to avoid any possible bias. After each presentation, the subjects had to answer a multiple choice question about the content of the attended story. The subjects also had to mark on a scale on the screen, what percentage of words of the attended speaker they could understand. After each experiment, we also repeated the presentations 1 and 5 so that we could collect enough data in the 'no noise' condition to independently train decoders on data with no exposure to babble noise, if the need arose. Table II shows the duration of presentations for different speaker separations.

Choice of SNRs: We chose signal to noise ratios that would ensure significantly different speech intelligibility levels. For this, we conducted an adaptive procedure to find the speech recognition threshold (SRT), i.e. the SNR at which speech intelligibility was 50%. The subject had to attend to a list of 'Matrix' sentences (Luts et al. 2014), in the presence of a competing talker, and surrounding babble noise for the 3 speaker separations. The SNR was changed adaptively (Brand & Kollmeier 2002) until the point where the subject could recognize 50% of the words in the sentence correctly. This behavioral test was done before the start of the main experiment. Thus, standard SRTs for the 3 speaker separations were found for 13 out of the 28 subjects who participated. An experiment was also designed to find subjective SRTs in the case of non-standardized speech material such as the short stories used in our main experiment (Decruy et al. 2018). During this adaptive procedure, subjects were asked to indicate

on the screen whether they understood more or less than 50% of the attended story or sentence. The SNRs were adapted accordingly until the subjects reached the 50% speech intelligibility point. Subjective SRTs found using this method ('rate' method) as well as standard SRTs found using the standard adaptive procedure ('recall' method) were compared for the Matrix speech material. On a set of 8 subjects who participated in a pilot study, the results for the 'rate' and the 'recall' methods using the Matrix speech material differed, for the 3 speaker separations (figure 1). We found a difference of 1.3 dB between the mean 'rate' (subjective) and 'recall' (standard) SRT, for the 180 degree speaker separation for the Matrix speech material. Since standard tests can not be used for finding the SRT for story speech material, we applied the 1.3 dB as a correction on the mean 'rate' SRT for the story speech material to find an estimated standard SRT for story speech material. At this estimated SRT (-7.1 dB), we could expect 50% intelligibility for the story speech material for the speaker separation of 180 degrees. This was chosen to be the lowest SNR for the main experiment. We then chose 2 more SNRs -4.1 dB and -1.1 dB, increasing the first chosen SNR in steps of 3 dB, similar to the method of choosing SNRs by Petersen et al. (2016). We, however, refrained from using individualized SNR levels as done by Petersen et al. (2016), since we tested only normal hearing subjects. In addition to these 3 SNRs, we also recorded data with no babble noise in the background.

Choice of speaker positions: Speaker positions were chosen to be representative of different real-life situations. We chose 180 degree speaker separation to represent an intuitively easy scenario, where spatial release from masking would make auditory object formation and selection an easy task. We also included the speaker positions -5 and 5 degrees which is an extreme case in which one can imagine two speakers standing shoulder to shoulder at a distance of 3 meters in front of the subject. Compared to the 180 degree condition, this scenario can be expected to be difficult since the speakers are almost co-located, and a certain amount of effort would be necessary in order to attend to one speaker and ignore the other. This is confirmed by the SRTs in 180 and 10 degree speaker separations shown in figure 1 for the 'recall' method using Matrix speech material. The third choice for speaker separation was 60 degrees, in which case, both the speakers were on the same side of the head. Das et al. (2017) found that in a system performing neuro-steered noise suppression at low input SNRs, speaker positions on the same side of the head resulted in a relatively lower improvement in output SNRs compared to symmetric speaker set ups. To balance between the two sides of the head, to avoid introducing a lateralization bias during the training of decoders (Das et al. 2016), we split this condition between two experiments where the speakers were at -30 and -90 degrees, and 30 and 90 degrees. During our analysis, we combined the results of these two experiments, and only analyzed the differences between the three speaker separations: 10 degrees, 60 degrees and 180 degrees.

C. Attention decoding

1) Data Preprocessing: The recorded EEG data was bandpass filtered between 0.5 and 10 Hz, using an equiripple bandpass filter with a maximal pass band attenuation of 0.5 dB, transition width of 0.9 Hz, and stop band attenuations of 20 dB (lower stop band) and 15 dB (upper stop band). This choice of the frequency range was based on earlier studies (Das et al. 2016, Ding & Simon 2012*b*, Golumbic et al. 2013, Pasley et al. 2012) that showed the best attention decoding performance based on cortical envelope tracking in this range. The filtered signal was then downsampled to 20 Hz. For the story stimuli presented to the subjects, a gammatone filterbank was applied to split

the speech signal into 15 filter bands, and a powerlaw compression with exponent 0.6 was then applied to each of these subbands (Patterson et al. 1995). The subbands were bandpass filtered using the same equiripple filter used for the EEG signals, and finally they were combined with equal weights, and downsampled to 20 Hz, to form a single 'powerlaw subband' envelope per stimulus. Biesmans et al. (2016) found that this method of auditory inspired envelope extraction resulted in significantly better attention decoding accuracies than other envelope extraction methods. The data set was divided into 30 second trials which resulted in 276 trials per subject, equally divided between the three speaker separations.

2) Stimulus Reconstruction and Decoder Design: We used the algorithm described by Biesmans et al. (2016) for our decoder design. 'Powerlaw subband' envelopes $s_a(t)$ and $s_u(t)$ were extracted from the attended and unattended speech signals respectively. A linear spatio-temporal decoder was used to reconstruct the envelope of the attended speech signal from the EEG data. A decoder consists of weights $D_c(n)$ for each EEG channel c and each time lag n. The time lags represent the difference between the time of stimulus presentation and the time when this stimulus is reflected in the cortical response. We used time lags of up to 250ms (O'Sullivan et al. 2014). If EEG data of channel c is represented by $M_c(t)$ where t is the time index, the reconstructed envelope on the attended speech signal is given by:

$$\tilde{s}_a(t) = \sum_{n=0}^{N_l-1} \sum_{c=1}^{C} D_c(n) M_c(t+n). \tag{1}$$

where C denotes the number of channels and N_l denoted the number of time-lags used in the decoding. For the sake of notation, we stack all the $D_c(n)$ in a vector $\mathbf{D} = [D_1(0), D_1(1), ..., D_1(n), D_2(0), ..., D_2(n), ..., D_C(0), ..., D_C(n)]^T$, and we stack all $M_c(t)$ and lags in a vector $\mathbf{M}(t) = [M_1(t), M_1(t+1), ..., M_1((t+N_l-1)), M_2(t), M_2(t+1), ..., M_2((t+N_l-1)), ..., M_C(t), ..., M_C((t+N_l-1))]^T$. The design of the decoder is in such a way as to minimize the mean squared error between the attended speech envelope $s_a(t)$ and its reconstruction $\tilde{s}_a(t)$. The decoder for each trial is found based on all other trials by solving the standard minimum mean squared error problem:

$$\mathbf{D} = R_{\mathbf{MM}}^{-1} \mathbf{r}_{\mathbf{M}S}, \tag{2}$$

where $R_{\mathbf{MM}} = E\{\mathbf{M}(t)\mathbf{M}(t)^T\}$ is the correlation matrix of the EEG data, and $\mathbf{r}_{\mathbf{M}S} = E\{\mathbf{M}(t)s_a(t)\}$ is the cross-correlation vector between the attended speech envelope and the EEG data, over all channels and time lags (where $E\{.\}$ denotes the expectation operator). For each subject, we used 30 s trials for decoding attention. Adopting a leave-one-trial-out approach for training decoders, the correlation matrices and the cross-correlation vectors are estimated over the samples from all trials except the test trial to compute the decoder from 2, after which it is applied to the the test trial's EEG data. Thus, a reconstructed attended envelope $\tilde{s}_a(t)$ was found per trial, for which Pearson's correlation coefficient was then computed with that trial's $s_a(t)$ and $s_u(t)$. Attention was considered to be decoded correctly if the correlation of $\tilde{s}_a(t)$ with $s_a(t)$ was found to be higher than with $s_u(t)$. These results were then grouped based on stimuli conditions and averaged across trials to obtain percentage decoding accuracies per subject for these conditions.

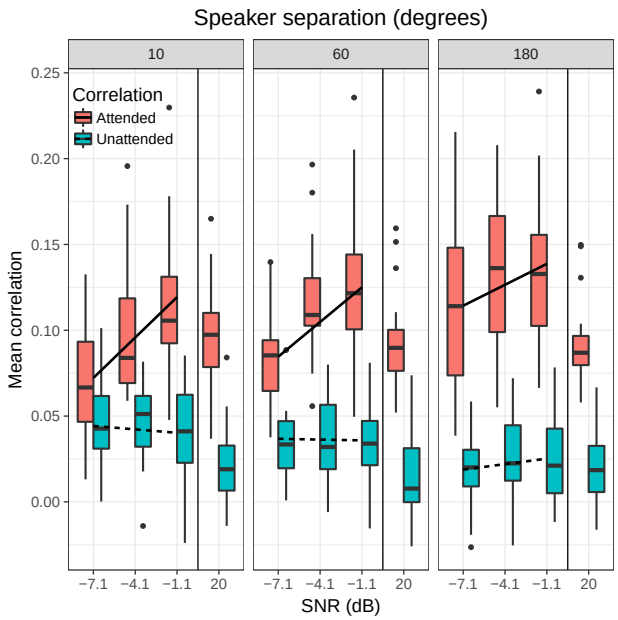

Fig. 2: Boxplots showing mean attended and unattended correlations across subjects for different SNRs and speaker separations. The horizontal lines show predictions of the linear mixed effect models for each speaker separation.

III. Results

A. Mean attended and unattended correlations

Attended and unattended correlations were calculated for each trial, and we analyzed the influence of SNR and speaker separation on the mean correlations. Remarkably, as can be seen in the figure 2, we observe that both mean attended and mean unattended correlations were lower in the 'no noise' condition (to the right of the vertical lines) compared to when some noise is added. However, within the noisy conditions, we observed that mean attended correlations had a tendency to drop linearly with increasing noise power. Therefore we excluded the 'no noise' condition data while fitting linear mixed effect models for the correlations. SNR and speaker separation were taken to be fixed factors, and subject as a random factor. We found that while mean attended correlations dropped significantly ($p<0.001$) with lower SNRs, SNR did not have a significant influence on mean unattended correlations. We also found that speaker separation significantly influenced both mean attended ($p<0.001$) and mean unattended correlations ($p<0.001$).

Post-hoc Tukey tests were performed to compare between groups of SNR and speaker separation. We found that mean attended correlations at -7.1 dB SNR were significantly lower than those at -1.1 dB and -4.1 dB ($p<0.001$). The mean attended correlations at 180 degrees speaker separation were found to be significantly higher than at 60 and 10 degrees ($p<0.001$). The post-hoc tests on mean unattended correlations revealed that the unattended correlations dropped as the speaker separation increased (unattended correlations at: 10 degrees > 60 degrees with $p = 0.039$, 60 degrees > 180 degrees with $p = 0.002$, and 10 degrees > 180 degrees with $p<0.001$).

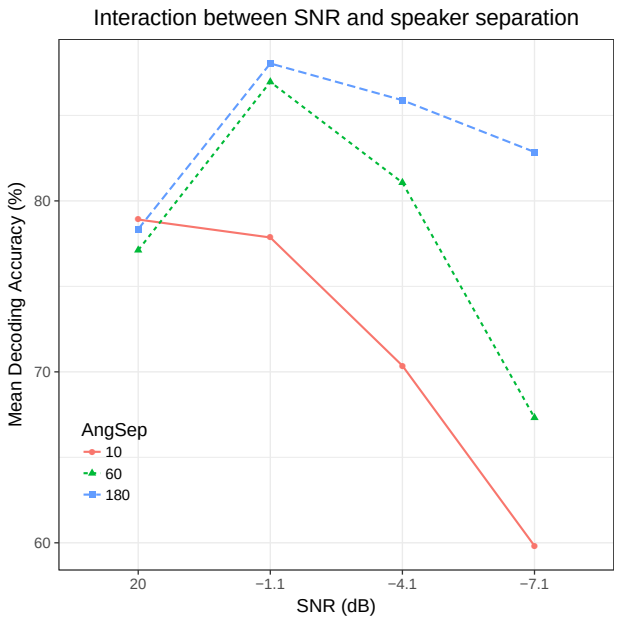

Fig. 3: Mean decoding accuracies for the 4 SNRs and 3 speaker separations showing interaction between noise power and the angular separation between the competing talkers.

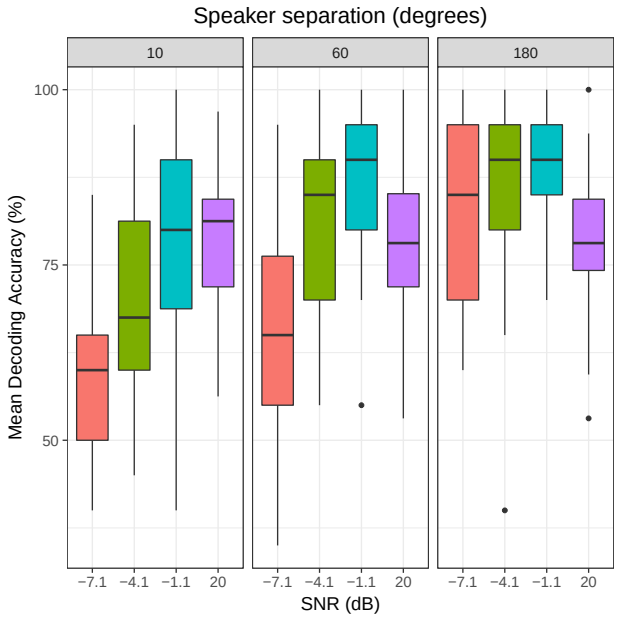

Fig. 4: Decoding accuracies based on trials grouped by SNRs.

B. Attention decoding accuracy

Attention decoding accuracy was calculated based on trials grouped with respect to the different auditory conditions in the experiments. Figure 3 shows mean decoding accuracy (over all 28 subjects) for the different SNRs and speaker separations. Thus, we analyzed the performance differences across the 4 SNR conditions: -7.1 dB,

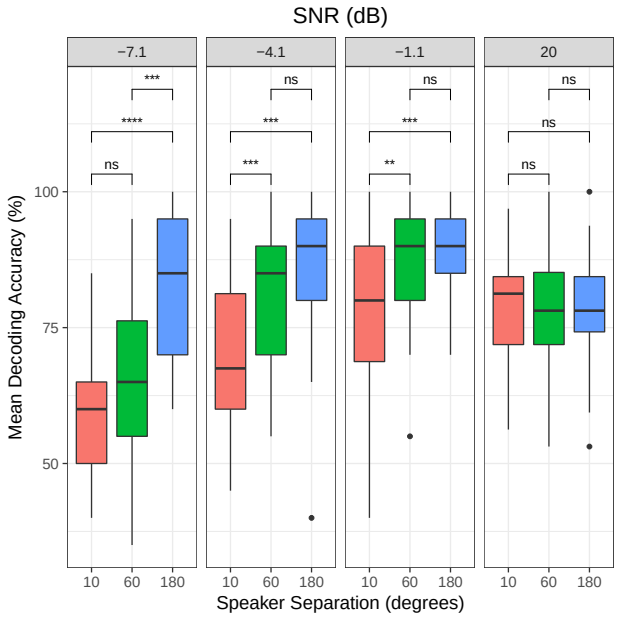

Fig. 5: Decoding accuracies based on trials grouped by speaker separation.

-4.1 dB, -1.1 dB and 20 dB, as well as across the 3 speaker separations: 10, 60 and 180 degrees.

To investigate the overall effect of SNR and speaker separation on attention decoding accuracy, a linear mixed effect model was used to model the relationship between mean decoding accuracy and SNR as well as speaker separation, with subject as random factor. The linear mixed effect model was fitted using all the data excluding the 'no noise' condition. The results showed that the attention decoding accuracies dropped significantly ($p<0.001$) with SNR as well as speaker separation (figure 3). Post-hoc comparisons (Tukey test) between decoding accuracies grouped by SNR revealed that decoding accuracies dropped significantly with SNRs. The relationship between decoding accuracy and SNR can be observed in figure 4.

We also analyzed the impact of SNR on attention decoding accuracy for each of the speaker separations. We performed paired t-tests between decoding accuracies grouped by SNR for each speaker separation (figure 4). The significance level for differences between different groups are indicated by: '*' for $p<0.050$, '**' for $p<0.010$, '***' and '****' for $p<0.001$. We found an interesting trend here: with respect to the 'no noise' condition, attention decoding accuracy increased when some noise was present for both 180 and 60 degree speaker separations. In the noisy conditions, as can be seen in figure 4, decoding accuracies dropped with SNR for the 10 and 60 degree speaker separations.

Post-hoc comparisons were also done to look for differences in decoding accuracies grouped by speaker separation (figure 5). A comparison between groups revealed that decoding performance decreased with speaker separation (performance for 180 > 60 > 10 degrees with $p<0.001$). The relationship between decoding performance and speaker separation can be observed in figure 5.

We also investigated the effect of speaker separation on the decoding accuracies, for each of the SNR conditions.

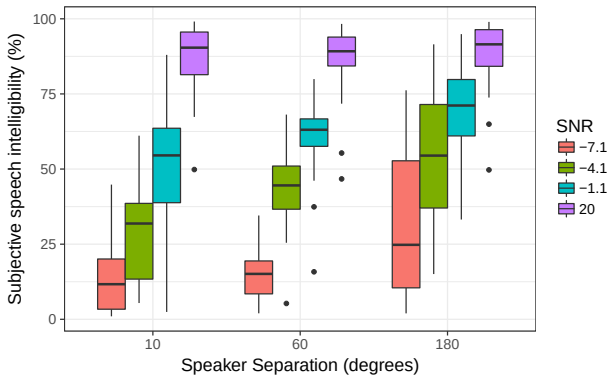

Fig. 6: Reported speech intelligibility at different SNRs for each speaker separation.

As can be seen in figure 5, for the 'no noise' condition, there was no significant effect of speaker separation on decoding accuracy. However, for the noisy conditions, it was seen that the decoding accuracy decreased with speaker separation.

C. Comparison with speech intelligibility (SI)

Out of the 28 subjects who participated in the study, 13 subjects also participated in adaptive behavioral speech tests ('recall' method) to find their SRTs for the 3 speaker separations, using the Matrix speech material as the attended sentence and a story as the competing talker, with the power of the background babble adapted based on the subjects' responses. From this data, we fitted psychometric curves for each subject and speaker separation to get speech intelligibility scores (for Matrix speech material) at the SNRs presented in the main experiments. For convenience we will refer to these scores as 'standard' speech intelligibility scores. We used a linear mixed effect model to investigate standard speech intelligibility with fixed factors as SNR and speaker separation and subject as the random factor. We found that standard speech intelligibility (for Matrix speech material) significantly dropped with SNR as well as speaker separation ($p<0.001$). We calculated the Spearman correlation between attention decoding accuracy and standard speech intelligibility under the same conditions, and found a significant effect (ρ = 0.445, $p<0.001$). We excluded the 'no noise' condition from the data for fitting the linear model because we found attention decoding accuracies to be significantly higher in the presence of moderate noise as compared to the 'no noise' condition.

During the experiments, all subjects reported their speech intelligibility (subjective) after each presentation. Figure 6 shows the mean of the speech intelligibility scores indicated by the 28 subjects grouped by the different SNR conditions and speaker separations. For convenience, we refer to this as 'subjective' speech intelligibility. We also investigated the relationship between subjective speech intelligibility and attention decoding accuracy. We fitted a linear mixed effect model for the subjective speech intelligibility with SNR and speaker separation as fixed factors, and subject as random factor. We found that the subjective speech intelligibility dropped significantly with dropping SNR as well as speaker separation ($p<0.001$).

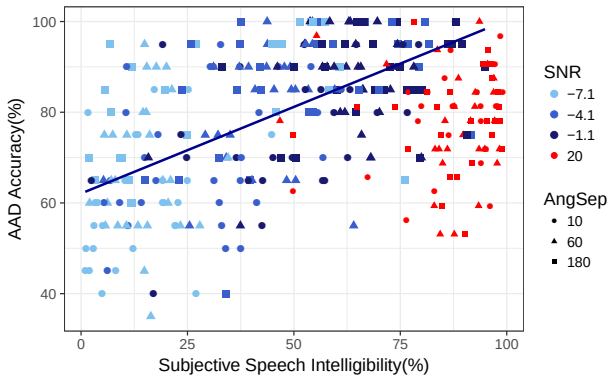

Fig. 7: Attention decoding accuracies plotted against subjective speech intelligibilities of 28 subjects for the 4 SNRs and 3 angular separations between speakers. The points in red correspond to the 'no noise' condition. These points were excluded in the linear mixed model fitting (the dark blue line).

Figure 7 shows how attention decoding accuracy varied with subjective speech intelligibility. We fitted a linear mixed effects model for attention decoding accuracy with the subjective speech intelligibility as fixed factor, and subject as random factor. The decoding accuracies from the 'no noise' condition were excluded from the model fitting. We found that attention decoding accuracy dropped significantly with subjective speech intelligibility ($p<0.001$). The predicted values of the linear mixed effects model for attention decoding accuracy are plotted as a line in the figure 7. We also found a significant Spearman correlation ($\rho = 0.585$, $p<0.001$) between attention decoding accuracy and subjective speech intelligibility.

IV. Discussion

This study was designed to investigate the impact of challenging acoustic conditions on attention decoding performance. The chosen SNR values ensured significantly different speech intelligibility between conditions (figure 6). The mean attention decoding performance on 30s trials in the 'no noise' condition over all speaker separation conditions spanned a range between 58% and 96% which is similar to the attention decoding performance reported by Biesmans et al. (2016) and Das et al. (2016).

A. *Influence of noise power on attention decoding*

We found that SNR adversely affects the attended speech representation in the neural responses, whereas the representation of the unattended speech envelope remains more or less unaffected. This is also in line with the study of Petersen et al. (2016) where the tracking of the attended speech was found to increase in magnitude with lower noise levels, while no such effect could be observed for the tracking of the ignored speech. Since we found that attended correlations dropped significantly with SNR, the separation between the attended and unattended correlations decreased with lower SNRs. This can be expected to have a direct influence on the attention decoding accuracy. Indeed, we found that attention decoding accuracy significantly drops with SNR.

Our analysis also shows how the interaction between background noise power and separation of speakers affect attention decoding accuracy (figure 3). We found that mean decoding performance was significantly lower in the

‘no noise’ condition compared to -1.1 dB SNR. With the inclusion of background noise, attention to one speaker becomes harder, and consequently demands more effort from the subject. The top-down processes involved in suppressing the additional noise could also result in stronger neural entrainment of the attended speech stream, and hence better signal-to-noise ratio in the EEG data. This could explain the relatively better decoding accuracies compared to the less challenging ‘no noise’ condition. However, as the background noise power increases, this benefit fails to outweigh the disadvantage of the noise. With rising noise power, sound segregation cues are degraded, and as a consequence, features extracted from the auditory scene can no longer be effectively grouped to represent a specific attended stream (Ding et al. 2014). This results in reduced cortical entrainment, and hence worse decoding performance as can be seen for -4.1 dB and -7.1 dB SNRs.

B. Influence of speaker positions on attention decoding

From figure 5, it can be inferred that in challenging SNR conditions, decoding performance for smaller speaker separations is more severely affected than for speaker positions that are far apart (for eg. 180 degrees). In a cocktail party scenario, the listener performs sound segregation based on many features such as spectro-temporal cues, perceptual differences, and speaker locations. Binaural cues such as inter-aural time differences (ITD) and inter-aural level differences (ILD) occurring due to difference in spatial locations of the competing talkers are exploited to perform segregation of speech streams (Carlyon 2004). Angular separation between speakers is expected to affect spatial release from masking (SRM). Kidd Jr et al. (1998) and Jones & Litovsky (2011) found the advantage of speaker separation in the identification of audio signal patterns, thereby improving speech intelligibility in a cocktail party scenario. In our study, therefore, the 10 degree speaker separation can be expected to be a more difficult auditory scene in terms of segregation of speech streams, in comparison to the wider speaker separations 60 and 180 degrees. This is in line with our findings that the 10 degrees speaker separation resulted in the worst SRT values from the behavioral tests. From figure 3, it can also be seen that the 10 degree speaker separation is indeed the least robust to background noise power, with decoding accuracies dropping more sharply than the other speaker separations as the noise power increases.

As opposed to SNR, speaker separation did have a significant effect on the correlation with the unattended speaker. Horton et al. (2013) describe an entrainment-based suppression mechanism that reduces correlations with the unattended speech stream during selective attention. In our study, the finding that unattended correlations increase with reducing speaker separation suggests that the suppression mechanism is adversely affected by noise for closer speaker separations. The difference between mean attended and unattended correlations was also found to decrease with decreasing speaker separation, which can be expected to result in reduction in attention decoding accuracy (figure 2).

C. Relationship between speech intelligibility and attention decoding accuracy

We hypothesized that the top-down processes that support auditory object formation and selection, and suppression of noise to improving speech perception, are also responsible for the improvement in signal-to-noise ratio of the neural responses that support attention decoding under challenging auditory conditions. Zekveld et al. (2006)

found brain regions involved in top-down processing supplementing speech comprehension to be more active when the speech was less intelligible or presented at low SNRs. Riecke et al. (2017) demonstrated that speech-brain entrainment contributes functionally to speech intelligibility, and that speech-entrainment and intelligibility interact reciprocally. Ding & Simon (2013) also found subjective speech intelligibility to be correlated with measures of neural synchronization with the stimulus. We therefore expected to see a significant correlation between the speech intelligibility reported by the subjects and their attention decoding accuracy. In figure 7, although we observed large variability in attention decoding accuracy with respect to subjective speech intelligibility, we indeed found a significant correlation between speech intelligibility and attention decoding accuracy. This is also supported by the findings of Vanthornhout et al. (2018) where the correlation between the reconstructed envelope from neural responses, and the acoustic speech envelope was proposed to be used as an objective measure of speech intelligibility. Rimmele et al. (2015) found more precise speech-tracking of attended speech to be related to processing its fine structure, possibly reflecting the application of higher-order linguistic processes. They did not find a similar dependence on the fine structure for the unattended speech. This, along with the dependence of attended correlations on the signal to noise ratio that we found in our study, suggests that the top-down processes that are involved in sound segregation and speech perception, thereby influencing speech intelligibility, could also be responsible for the cortical entrainment to the attended speech stream.

In figure 7, mean attention decoding accuracy can be seen to exceed 80% when subjective speech intelligibility is close to 50%. We observe that even when the noise power is higher than the attended speech power, and speech intelligibility has degraded, auditory attention decoding is still possible. Both background noise level and angular separation between competing talkers have a significant effect on speech intelligibility as well as on the neural representation of the competing speech streams. For attention decoding algorithms to be incorporated in neuro-steered hearing prostheses, this study is a step towards better understanding the limitations posed by acoustic conditions on decoding performance. There is a need to use EEG-data from realistic challenging acoustic conditions to evaluate nuero-steered noise suppression systems such as proposed by Van Eyndhoven et al. (2017) and Das et al. (2017) in order to study the limitations of the system, including its tolerance towards realistic attention decoding errors. In this study we used simple spatio-temporal linear decoders in order to reconstruct the attended speech envelope. Newer approaches to decoding such as neural-network (Taillez et al. 2017, Wong et al. 2018), CCA (de Cheveigné et al. 2018) or state-space modelling based (Akram et al. 2016, Miran et al. 2017) decoders have the possibility of more robust attention decoding by which the boundary conditions of a neuro-steered system can be further improved.

V. Conclusion

Auditory attention decoding has the potential to improve state-of-the-art noise suppression algorithms in hearing prostheses by incorporating information about the attention of the listener. The main goal of this paper was to analyze the impact of background noise and speaker separation, and their interaction, on the neural entrainment of the attended speech stream, and thereby the attention decoding accuracy. We compared the attention decoding

accuracy for a range of SNRs and relative speaker positions. We also conducted behavioral tests to estimate speech intelligibility under these conditions.

We found an improvement in decoding performance with the inclusion of moderate background noise. It could possibly be due to improvement in the SNR of the EEG data because of the additional effort by the subjects. Decoding performance decreased with increasing noise power, and the degrading effect of the noise was stronger when the competing talkers were located closer to each other. We also found a significant correlation between attention decoding performance and speech intelligibility across conditions indicating that acoustic conditions with 100% speech intelligibility are not necessary to bootstrap an attention decoding system.

ACKNOWLEDGMENTS

The authors would like to thank Jelien Geivers, Tine Arras and Astrid Vandevelde for their help in conducting the experiments. We would also like to thank all the subjects for their participation in the study.

REFERENCES

Abhijith, M. & Bertrand, A. (2018), The effect of miniaturization and galvanic separation of eeg sensor nodes in an auditory attention detection task, *in* 'Proc. 40th International Conference of the IEEE EMBS', number accepted.

Akram, S., Presacco, A., Simon, J. Z., Shamma, S. A. & Babadi, B. (2016), 'Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling', *NeuroImage* **124**, 906–917.

Aroudi, A. & Doclo, S. (2017), EEG-based auditory attention decoding using unprocessed binaural signals in reverberant and noisy conditions, *in* 'Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE', IEEE, pp. 484–488.

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', *Neural Systems and Rehabilitation Engineering, IEEE Transactions on* . accepted for publication.

Brand, T. & Kollmeier, B. (2002), 'Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests', *The Journal of the Acoustical Society of America* **111**(6), 2801–2810.

Carlyon, R. P. (2004), 'How the brain separates sounds', *Trends in cognitive sciences* **8**(10), 465–471.

Das, N., Biesmans, W., Bertrand, A. & Francart, T. (2016), 'The effect of head-related filtering and ear-specific decoding bias on auditory attention detection', *Journal of neural engineering* **13**(5), 056014.

Das, N., Van Eyndhoven, S., Francart, T. & Bertrand, A. (2017), EEG-based attention-driven speech enhancement for noisy speech mixtures using N-fold multi-channel Wiener filters, *in* 'Signal Processing Conference (EUSIPCO), 2017 25th European', IEEE, pp. 1660–1664.

de Cheveigné, A., Wong, D. D., Liberto, G. M. D., Hjortkjaer, J., Slaney, M. & Lalor, E. (2018), 'Decoding the auditory brain with canonical component analysis', *NeuroImage* **172**, 206 – 216.
URL: *http://www.sciencedirect.com/science/article/pii/S1053811918300338*

Decruy, L., Das, N., Verschueren, E. & Francart, T. (2018), 'Measuring intelligibility of connected discourse', *Trends in Hearing* . submitted.

Di Liberto, G. M., O'Sullivan, J. A. & Lalor, E. C. (2015), 'Low-frequency cortical entrainment to speech reflects phoneme-level processing', *Current Biology* **25**(19), 2457–2465.

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

Ding, N. & Simon, J. Z. (2012*a*), 'Emergence of neural encoding of auditory objects while listening to competing speakers', *Proc. National Academy of Sciences* **109**(29), 11854–11859.

Ding, N. & Simon, J. Z. (2012*b*), 'Neural coding of continuous speech in auditory cortex during monaural and dichotic listening', *Journal of neurophysiology* **107**(1), 78–89.

Ding, N. & Simon, J. Z. (2013), 'Adaptive temporal encoding leads to a background-insensitive cortical representation of speech', *Journal of Neuroscience* **33**(13), 5728–5735.

Fiedler, L., Obleser, J., Lunner, T. & Graversen, C. (2016), Ear-EEG allows extraction of neural responses in challenging listening scenarios-a future technology for hearing aids?, *in* 'Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the', IEEE, pp. 5697–5700.

Francart, T., Van Wieringen, A. & Wouters, J. (2008), 'Apex 3: a multi-purpose test platform for auditory psychophysical experiments', *Journal of Neuroscience Methods* **172**(2), 283–293.

Fuglsang, S. A., Dau, T. & Hjortkjær, J. (2017), 'Noise-robust cortical tracking of attended speech in real-world acoustic scenes', *NeuroImage* **156**, 435 – 444.
URL: *http://www.sciencedirect.com/science/article/pii/S105381191730318X*

Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z. et al. (2013), 'Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party"', *Neuron* **77**(5), 980–991.

Horton, C., D'Zmura, M. & Srinivasan, R. (2013), 'Suppression of competing speech through entrainment of cortical oscillations', *Journal of neurophysiology* **109**(12), 3082–3093.

Jones, G. L. & Litovsky, R. Y. (2011), 'A cocktail party model of spatial release from masking by both noise and speech interferers', *The Journal of the Acoustical Society of America* **130**(3), 1463–1474.

Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V. & Kollmeier, B. (2009), 'Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses', *EURASIP Journal on Advances in Signal Processing* **2009**, 6.

Kidd Jr, G., Mason, C. R., Rohtla, T. L. & Deliwala, P. S. (1998), 'Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns', *The Journal of the Acoustical Society of America* **104**(1), 422–431.

Luts, H., Jansen, S., Dreschler, W. & Wouters, J. (2014), 'Development and normative data for the Flemish/Dutch Matrix test'.

Miran, S., Akram, S., Sheikhattar, A., Simon, J. Z., Zhang, T. & Babadi, B. (2017), 'Real-time tracking of selective auditory attention from M/EEG: A bayesian filtering approach', *bioRxiv* p. 222661.

Mirkovic, B., Bleichner, M. G., De Vos, M. & Debener, S. (2016), 'Target speaker detection with concealed EEG around the ear', *Frontiers in neuroscience* **10**, 349.

Mirkovic, B., Debener, S., Jaeger, M. & De Vos, M. (2015), 'Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications', *Journal of neural engineering* **12**(4), 046007.

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. (2014), 'Attentional selection in a cocktail party environment can be decoded from single-trial EEG', *Cerebral Cortex* p. bht355.

O'Sullivan, J., Chen, Z., Herrero, J., McKhann, G. M., Sheth, S. A., Mehta, A. D. & Mesgarani, N. (2017), 'Neural decoding of attentional selection in multi-speaker environments without access to clean sources', *Journal of neural engineering* **14**(5), 056001.

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T. & Chang, E. F. (2012), 'Reconstructing speech from human auditory cortex', *PLoS-Biology* **10**(1), 175.

Patterson, R. D., Allerhand, M. H. & Giguere, C. (1995), 'Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform', *The Journal of the Acoustical Society of America* **98**(4), 1890–1894.

Petersen, E. B., Wöstmann, M., Obleser, J. & Lunner, T. (2016), 'Neural tracking of attended versus ignored speech is differentially affected by hearing loss', *Journal of neurophysiology* **117**(1), 18–27.

Riecke, L., Formisano, E., Sorger, B., Başkent, D. & Gaudrain, E. (2017), 'Neural entrainment to speech modulates speech intelligibility', *Current Biology* .

Rimmele, J. M., Golumbic, E. Z., Schröger, E. & Poeppel, D. (2015), 'The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene', *Cortex* **68**, 144–154.

Taillez, T., Kollmeier, B. & Meyer, B. T. (2017), 'Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech', *European Journal of Neuroscience* .

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 Transactions on Biomedical Engineering* **64**(5), 1045–1056.

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z. & Francart, T. (2018), 'Speech intelligibility predicted from neural entrainment of the speech envelope', *Journal of the Association for Research in Otolaryngology* **19**(2), 181–191.
URL: *https://doi.org/10.1007/s10162-018-0654-z*

Wong, D., S., F., Hjortjaer, J., Di Liberto, G. & de Cheveigné, A. (2018), 'Classifying attended talker from EEG using artificial neural networks', *ARO Midwinter meeting (abstract)* .

Zekveld, A. A., Heslenfeld, D. J., Festen, J. M. & Schoonhoven, R. (2006), 'Top–down and bottom–up processes in speech comprehension', *NeuroImage* **32**(4), 1826–1836.

Zink, R., Proesmans, S., Bertrand, A., Van Huffel, S. & De Vos, M. (2017), 'Online detection of auditory attention with mobile EEG: closing the loop with neurofeedback', *bioRxiv* .
URL: *https://www.biorxiv.org/content/early/2017/11/13/218727*