Published in final edited form as:
J Neural Eng. 2017 October ; 14(5): 056001. doi:10.1088/1741-2552/aa7ab4.

Neural decoding of attentional selection in multi-speaker environments without access to clean sources

James O'Sullivan[1,2], **Zhuo Chen**[1], **Jose Herrero**[4], **Guy M McKhann**[3], **Sameer A Sheth**[3], **Ashesh D Mehta**[4], and **Nima Mesgarani**[1,2]

[1]Department of Electrical Engineering, Columbia University, New York, NY, United States of America

[2]Mortimer B Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, United States of America

[3]Department of Neurological Surgery, The Neurological Institute, 710 West 168 Street, New York, NY 10032, United States of America

[4]Department of Neurosurgery, Hofstra-Northwell School of Medicine and Feinstein Institute for Medical Research, Manhasset, NY 11030, United States of America

Abstract

Objective—People who suffer from hearing impairments can find it difficult to follow a conversation in a multi-speaker environment. Current hearing aids can suppress background noise; however, there is little that can be done to help a user attend to a single conversation amongst many without knowing which speaker the user is attending to. Cognitively controlled hearing aids that use auditory attention decoding (AAD) methods are the next step in offering help. Translating the successes in AAD research to real-world applications poses a number of challenges, including the lack of access to the clean sound sources in the environment with which to compare with the neural signals. We propose a novel framework that combines single-channel speech separation algorithms with AAD.

Approach—We present an end-to-end system that (1) receives a single audio channel containing a mixture of speakers that is heard by a listener along with the listener's neural signals, (2) automatically separates the individual speakers in the mixture, (3) determines the attended speaker, and (4) amplifies the attended speaker's voice to assist the listener.

Main results—Using invasive electrophysiology recordings, we identified the regions of the auditory cortex that contribute to AAD. Given appropriate electrode locations, our system is able to decode the attention of subjects and amplify the attended speaker using only the mixed audio. Our quality assessment of the modified audio demonstrates a significant improvement in both subjective and objective speech quality measures.

Correspondence to: Nima Mesgarani.

ORCID iDs
James O'Sullivan https://orcid.org/0000-0002-3501-9647

Significance—Our novel framework for AAD bridges the gap between the most recent advancements in speech processing technologies and speech prosthesis research and moves us closer to the development of cognitively controlled hearable devices for the hearing impaired.

Keywords

ECoG; sEEG; hearing aid; DNN; LSTM; attention; stimulus-reconstruction

1. Introduction

In a natural auditory environment, people with normal hearing can easily attend to a single speaker amongst many, and can switch their attention from one speaker to another with ease [1]. However, this task is extremely challenging for people who suffer from hearing impairments, which has been attributed to an increase in listening effort and a reliance on higher-level compensatory cognitive processes [2–5]. Assistive hearing devices can suppress certain types of background noise [6], but they cannot help a user attend to a single conversation amongst many without knowing which speaker is being attending to.

Several studies have revealed a dynamic and selective representation of the acoustic features of an attended speaker in human auditory cortex [7–12]. These findings have led to the idea of auditory attention decoding (AAD): the ability to decode the identity of an attended speaker over short enough time-scales so as to be useful for an assistive hearing device. AAD has been successfully implemented using various neural signal acquisition methods including non-invasive magnetoencephalography (MEG; [8]), electroencephalography (EEG; [13, 14]), and invasive electrocorticography (ECoG; [15]). However, many challenges must be overcome before AAD can be practically implemented [16–28]: from determining the optimal positioning of electrodes on the scalp [22], around the ear [23], and in the ear canal [25, 28], to developing state-space models to improve accuracy and speed at decoding attention [27].

In a typical implementation of AAD, the neural responses recorded from the subject's brain are compared with the acoustic properties of the individual speakers. The speaker who produces the maximum similarity with the neural data is determined to be the target. This technique assumes that there is access to the clean sound sources; a difficult requirement to satisfy for real-world implementations of this idea when only the mixed audio signal is available. One way to address the issue of the lack of access to the clean sources is beamforming [29, 30]. Specifically, the use of multichannel microphone recordings combined with non-negative independent component analysis [21] or multi-channel weiner filters (MWF) [31] has been proposed to create a neuro-steered spatial audio filter. While the quality of the separation in these methods is not as good as an actual clean source, AAD algorithms can tolerate moderate amounts of noise and cross-talk [19]. However, these methods are limited to scenarios where the target and interfering sources are separated in space with respect to the recording microphones: a condition that is not consistently guaranteed. In addition, the limited amplification factor achievable with beamforming may be insufficient when a greater suppression of the interfering speakers is desired to improve a user's experience [32]. An approach that addresses these shortcomings is needed to maximize the benefits of AAD and produce a cleaner, more robust amplification of a target speaker; a result that can be incorporated into assistive hearing devices.

We propose a novel framework that combines advances in single-channel speech separation methods with AAD to alleviate the requirement of a spatial separation between the target speaker and interfering speakers. This method can be used in place of, or in tandem with beamforming to create a more realistic solution to the AAD problem. Our method requires prior training on target speakers, meaning that its use is restricted to a known set of speakers with whom the user interacts. However, new speakers can be added to this set using a small amount of training data (~20 min).

Our system is based on deep neural network (DNN) audio source separation algorithms [33, 34]. A schematic of our proposed system is shown in figure 1: a spectrogram of the mixture is fed to several DNNs, each trained to separate a specific speaker from a mixture (DNN Spk1 to DNN SpkN). Simultaneously, a user is attending to one of the speakers (in this case, Spk1; red), and a spectrogram of this speaker is reconstructed from the neural recordings of the user [12]. This reconstruction is then compared with the outputs of each of the DNNs using a correlation analysis in order to determine which spectrogram is most similar to the neural reconstruction. Once identified, the selected spectrogram is converted into an acoustic waveform and added to the mixture so as to amplify the attended speaker.

We tested the efficacy of our system using invasive ECoG recordings from neurological subjects undergoing epilepsy surgery [12]. We used ECoG recordings because it allows us to determine the varying contributions of different auditory cortical areas to AAD. Given electrodes in the correct areas, ECoG also allows us to examine the upper bound of decoding speed and accuracy. Our system, while tested using invasive neural recordings, can also be applied to non-invasive methodologies [13, 22]. This framework for AAD systems bridges the gap between the latest developments in speech separation algorithms and speech prostheses to help a user attend to a single speaker in a multi-speaker environment.

2. Methods

2.1. Participants

A total of six subjects who were undergoing clinical treatment for epilepsy took part in this study. All subjects gave their written informed consent to participate in research; five subjects were situated at North Shore University Hospital (NSUH), and one subject was situated at the Columbia University Medical Center (CUMC). Subjects one and two were implanted with high-density subdural electrode arrays over the left (language dominant) temporal lobe with coverage over superior temporal gyrus (STG). The remaining four subjects partook in stereotactic EEG in which they were implanted bilaterally with depth electrodes. This resulted in varying amounts of coverage over the left and right auditory cortices for each subject (see figure 5).

2.2. Stimuli and experiments

Each subject participated in two experiments for this study: a single-speaker (S-S) and multi-speaker (M-S) experiment. The S-S experiment was used as a control. Each subject listened to four stories read by a female and male speaker (hereafter referred to as Spk1_f and Spk2_m, respectively), for a total of eight stories (four stories twice). Both Spk1_f and Spk2_m were native American English speakers, and were recorded in-house. In order to ensure the attentional engagement of each subject, the stories were randomly paused and the subject was instructed to repeat the last sentence. For the M-S experiment, subjects were presented with a mixture of the same female and male speakers (Spk1_f and Spk2_m), with no spatial separation between them. The acoustic waveform of each speaker was matched to have the same root mean squared (RMS) intensity. Unlike some studies in this field, we did not remove pauses longer than 0.5 s in order to make the task as natural as possible. All stimuli were presented using a single Bose® SoundLink® Mini 2 speaker situated directly in front of the subject.

The M-S experiment was divided into four blocks. Before each block, the subject was instructed to focus their attention on one speaker and to ignore the other. All subjects began the experiment by attending to the male speaker, and switched their attention to the alternate speaker on each subsequent block. In order to ensure that the subjects were engaged in the task, the story was intermittently paused and the subjects were asked to repeat the last sentence of the attended speaker. The locations of the pauses were predetermined and were the same for all subjects, but the subjects were unaware of when the pauses would occur. In total, there were 11 min and 37 s of audio presented to each subject during the M-S experiment. The S-S experiment lasted twice as long as each subject was required to listen to each story read by each speaker independently.

2.3. Data preprocessing and hardware

The subjects at NSUH were recorded using Tucker Davis Technologies (TDT®) hardware and sampled at 2441 Hz. The subject at CUMC was recorded using Xltek® hardware and sampled at 500 Hz. All further processing steps were performed offline. All filters were designed using Matlab's® Filter Design Toolbox, and were used in both a forwards and backwards direction to remove phase distortion. The TDT data were resampled to 500 Hz. A 1st order Butterworth high-pass filter with a cut-off frequency at 1 Hz was used to remove DC drift. Data were subsequently re-referenced using a common average scheme, with the effect of noisy electrodes mitigated by using a trimmed mean approach (i.e. removing the top and bottom 10% of electrodes at each sample point). To clarify, these electrodes were not removed from further analyses; rather, they are excluded when obtaining the mean of the data at each sample point. Line noise at 60 Hz and its harmonics (up to 240 Hz) were removed using 2nd order IIR notch filters with a bandwidth of 1 Hz. A period of silence was recorded before each experiment, and the corresponding data were normalized by subtracting the mean and dividing by the standard deviation of this pre-stimulus period.

Data were then filtered into the high-gamma band (70–150 Hz), the power of which is known to be modulated by speech [11, 12]. In order to obtain the power of this broad band, the data were first filtered into eight frequency bands between 70 and 150 Hz, each with a bandwidth of 10 Hz, using Chebyshev Type 2 filters. The power (analytic amplitude) of each band was then obtained using a Hilbert transform. We took the average of all eight frequency bands as the total power. This method is commonly used in neuroscience research [35].

2.4. Single-channel speaker separation

In order to automatically separate each speaker from the mixture, we employed a method of single-channel speech separation that utilizes a class of DNNs known as long short-term memory (LSTM) DNNs [34, 36]. Each DNN was trained to separate one specific speaker from two speaker mixtures. Figure 2 shows an example of speech separation using these DNNs. In our experiment, there were only two speakers ($\mathrm{Spk1_f}$ and $\mathrm{Spk2_m}$) presented to each subject. However, we are proposing a system that could work in a real-world situation where a device would contain multiple DNNs, each trained to separate specific speakers, any of whom may or may not be present in the environment. Because of this, we trained four DNNs to separate four speakers (two female and two male), hereafter referred to as $\mathrm{Spk1_f}$, $\mathrm{Spk2_m}$, $\mathrm{Spk3_f}$, and $\mathrm{Spk4_m}$. All speakers were native American English speakers. As stated before, two of the speakers ($\mathrm{Spk1_f}$ and $\mathrm{Spk2_m}$) were recorded in-house. However, $\mathrm{Spk3_f}$ and $\mathrm{Spk4_m}$ were taken from the wall street journal (WSJ) corpus [37].

The speech waveforms were converted into 100D Mel-frequency spectrograms. The goal was then to obtain an estimate $\hat{S}$ of a clean target spectrogram S from a mixture M. To do so, a soft mask $\hat{Y}$ was learnt and applied to the mixture to mask the interfering speech. The squared Euclidian distance between the masked spectrogram and the clean target spectrogram was treated as the error in order to generate the gradient that was back propagated through the DNN to update the parameters. The objective function was therefore:

$$E\left(\hat{Y}\right)=\|\hat{Y}M-S\|_2^2.$$

The inputs to the DNNs were log spectrograms, normalized so that each frequency band had zero mean and unit variance. Each DNN contained four layers with 300 nodes each, followed by a single layer containing 100 nodes with a logistic nonlinearity. An acoustic waveform was generated by inverting this spectrogram using the phase of the original mixture. See [34] for further information. Because our current method is proposed as a single output system (monaural), we did not attempt to address any binaural effects of this approach.

For training, we used ~5 h of speech from 103 interfering speakers from the WSJ corpus, and 20 min of speech from the target speakers. The target speaker was always mixed with one interfering speaker, and both were mixed into the same channel and with the same RMS intensity. Unseen utterances were used for testing (for both the target and interfering speakers). To ensure generalization, each of the DNNs never saw any of the other target speakers during training. E.g. the DNN trained to separate $\mathrm{Spk1_f}$ never saw $\mathrm{Spk2_m}$, $\mathrm{Spk3_f}$, or $\mathrm{Spk4_m}$.

2.5. Objective measure of speech separation quality

In order to determine the efficacy of the speech-separation algorithm, we used a commonly used objective measure of speech quality (MOS) known as the perceptual evaluation of speech quality (PESQ) score [38]. The PESQ algorithm produces a score between 1.0 and 4.5, where high values indicate better quality. This score is known to correlate well with subjective listening tests and has proven to be a reliable objective measure for the assessment of speech quality [39], and to some degree speech intelligibility [40].

2.6. DNN output correlation analysis

In our experiment, only two speakers were present in the mixture (Spk1$_f$ and Spk2$_m$). To ensure generalization, we trained two additional DNNs to separate two additional speakers, one male and one female (Spk3$_f$ and Spk4$_m$). We wanted to test how each DNN behaved when given various mixtures. To do this, we created a data set consisting of 103 random speakers (taken from the WSJ corpus) mixed with target speakers Spk1$_f$, Spk2$_m$, Spk3$_f$, and Spk4$_m$. We created 200 mixtures for each target speaker, resulting in 800 mixtures in total. We fed every mixture through each of the four DNNs and tested how well each DNN separated the target speaker in each case by obtaining a correlation coefficient (Pearson's *r*-value) between the output of the DNN and the spectrogram of the clean target speaker. Reported *r*-values are averaged across frequency, i.e. we calculated a single correlation value for each frequency band, and then averaged these to obtain a single correlation value.

2.7. Stimulus-reconstruction

In order to determine the attended speaker, we employed a method known as stimulus-reconstruction [8, 12, 13, 41]. This method applies a spatiotemporal filter (decoder) to neural recordings to reconstruct an estimate of the spectrogram of the attended speaker. The 100D Mel-frequency spectrograms were downsampled by a factor of ten, resulting in ten frequency bands. Each decoder was trained using the data from the single-speaker (S-S) experiment only. This was done to minimize any potential bias that may result from training the decoders on the M-S data, and to ensure generalization to new unseen speakers. Electrodes were chosen if they were significantly more responsive to speech than to silence. To perform statistical analyses, the neural data were segmented into 500 ms chunks and divided into two categories: speech and silence. Significance was determined using an unpaired *t*-test (false discovery rate (FDR) corrected, $q < 0.05$). This resulted in varying numbers of electrodes for each subject (see figure 5). Rejecting electrodes that were unresponsive to speech was done so as to reduce the number of parameters required to fit each decoder, as only electrodes that are responsive to speech could contribute to a linear reconstruction of an acoustic spectrogram. The decoders were trained using all electrodes simultaneously, and with time-lags from −400 to 0 ms. In addition, we also trained decoders using single-electrodes so as to determine the anatomical locations that contributed most to AAD. See [41] for further information on the stimulus-reconstruction algorithm.

2.8. Neural correlation analysis

As previously stated, we trained decoders using the data from the S-S experiment. These same decoders could then be used to reconstruct spectrograms from the M-S experiment [12]. Determining to whom the subject is attending requires a correlation analysis, commonly using Pearson's *r*-value [12]. Typically, whichever spectrogram has the largest correlation with the reconstructed spectrogram is taken to be the attended speaker [13, 22]. However, because we are using four DNNs, each trained to separate a speaker that may or may not be present in the mixture, the analysis becomes slightly more complex. Crucially, because the DNNs that do not see their designated speakers in the mixture are likely to output a signal very close to the mixture, we found that it was necessary to normalize the correlation values with respect to the mixture. This is because the correlation between the reconstructed spectrograms and the mixture can be very large (see Results; figure 4).

For clarity, we will first define some terminology: a spectrogram outputted from the *k*th DNN will be referred to as S_{DNN_k}, the spectrogram of the mixture as S_{MIX}, and the reconstructed spectrogram (from the neural responses) as S_{RECON}. In order to emphasize large correlations, we applied a Fisher transformation (inverse hyperbolic tangent) to each *r*-value.

The normalization procedure involved five steps. First, we obtained the correlation between S_{RECON} and each S_{DNN_k}, which we will refer to as ρ_{1_k}:

$$\rho_{1_k}=\tanh^{-1}\left\{r\left(S_{\mathrm{RECON}},S_{\mathrm{DNN}_k}\right)\right\} \quad (1)$$

where $r(x, y)$ is Pearson's correlation between the variables x and y, and $\tanh^{-1}$ is the inverse hyperbolic tangent function.

Next, we obtained the correlation between S_{RECON} and the difference between S_{DNN_k} and S_{MIX}, which we will refer to as ρ_{2_k}:

$$\rho_{2_k}=\tanh^{-1}\left\{r\left(S_{\mathrm{RECON}},S_{\mathrm{MIX}}-S_{\mathrm{DNN}_k}\right)\right\}. \quad (2)$$

Intuitively, this value should be close to zero if a DNN is outputting the mixture, small if it is outputting the attended speaker (because subtracting the attended spectrogram from the mixture will only leave behind portions of the unattended spectrogram), and large if it outputs the unattended speaker (similarly, because only portions of the attended spectrogram will be left). Therefore, taking the difference of ρ_{1_k} and ρ_{2_k}, and dividing by their sum, should produce a score (α_k) that can differentiate between each of these cases:

$$\alpha_k=\frac{\rho_{1_k}-\rho_{2_k}}{\rho_{1_k}+\rho_{2_k}}. \quad (3)$$

This was followed by a test-normalization (t-norm; [42]), in which the α score for each DNN was normalized relative to the distribution of α scores from all DNNs:

$$\beta_k=\frac{\alpha_k-\mu_\alpha}{\sigma_\alpha} \quad (4)$$

where μ_α and σ_α are the mean and standard deviation of the distribution of α scores from all DNNs. Finally, we subtracted the correlation between S_{DNN_k} and S_{MIX}, and added the constant 1, resulting in the final normalized correlation value (P*k*) for each DNN:

$$\mathrm{P}k{=}\beta_k{-}\tanh^{-1}\left\{r\left(S_{\mathrm{DNN}_k},S_{\mathrm{MIX}}\right)\right\}{+}1. \quad (5)$$

This last normalization step will further penalize a DNN that is simply outputting the mixture rather than separating the speakers. This could occur if a DNN's trained speaker was not in the mixture. The addition of the constant 1 is used to make the final result more intuitive, as otherwise the values would typically be less than zero.

Of the ten frequency bands in the downsampled spectrograms, the lowest two frequency bands (~50–200 Hz) were excluded to avoid bias towards the male speaker whose fundamental frequency occupied this region. All correlation values reported are the average of the r-values obtained across the remaining eight frequency bands.

2.9. Attention decoding index (ADI)

In order to obtain a measure of our ability to determine the attended speaker from neural recordings, we first segmented the reconstructed spectrogram from the M-S experiment into 20 s bins, resulting in 34 segments (17 where the subjects attended to male speaker, and 17 to the female speaker). As mentioned, we trained four DNNs to separate two female (Spk1_{f} and Spk3_{f}) and two male (Spk2_{m} and Spk4_{m}) speakers from random mixtures. Therefore, we obtained four normalized correlation values for each segment: P1_{f}, $mP2_{\mathrm{m}}$, $mP3_{\mathrm{f}}$, and $mP4_{\mathrm{m}}$. Because Spk1_{f} and Spk2_{m} were the only speakers that were actually presented to the subject, we would expect that $mP1_{\mathrm{f}}$ and $mP2_{\mathrm{m}}$ would be the largest, depending on whom the subject was attending to.

If there were only two possible correlation values to choose from, a segment could be considered correctly decoded if the attended speaker produced the largest correlation with the reconstructed spectrogram. However, when there are multiple values to choose from, it is important to take into account any bias for a particular speaker. This is of particular importance when using intracranial data, because it is possible that some electrodes could be tuned to speaker-specific features, and respond to those features regardless of the attentional focus of the subject. To take into account any such potential bias, we define the ADI as the proportion of the number of correct hits minus the number of false positives, for both target speakers.

$$\mathrm{ADI}{=}(\mathrm{CH}_{\mathrm{Spk1}}{+}\mathrm{CH}_{\mathrm{Spk2}}{-}\mathrm{FP}_{\mathrm{Spk1}}{-}\mathrm{FP}_{\mathrm{Spk2}})/n$$

where $\mathrm{CH}_{\mathrm{SpkN}}$ and $\mathrm{FP}_{\mathrm{SpkN}}$ are the number of correct hits and false positives for speaker N, respectively, and n is the number of segments. While similar to the sensitivity index d' [43], this approach is bounded between [−1,1].

Chance and significant ADI were determined by randomly shuffling the reconstructed spectrograms with respect to the DNN outputs 100 times for each subject. A null distribution of the ADI was then obtained using the same normalized correlation analysis described previously. The resulting mean ± SD performance was 0 ± 0.15. Significant performance was therefore determined to be 0.45 (three times the standard deviation). For comparison, we also calculated the ADI that would be achieved using the clean (ideal) spectrograms of $\mathrm{Spk1_f}$ and $\mathrm{Spk2_m}$. In this ideal situation, we assumed that the DNNs trained on $\mathrm{Spk3_f}$ and $\mathrm{Spk4_m}$ outputted the mixture.

2.10. Dynamic switching of attention

In order to simulate a dynamic scenario in which a subject was switching attention, we divided and concatenated the neural data into consecutive segments in which the subjects were attending to either speaker. Specifically, we divided the data into ten segments, each lasting 60 s. The subjects attended to the male speaker for the first segment. To assess our ability to track the attentional focus of each subject, we used a sliding window approach whereby we obtained correlation values every second over a specified window. We used window sizes ranging from 5 to 30 s (in 5 s increments for 6 window sizes in total). Larger windows should lead to more consistent (less noisy) correlation values and provide a better estimate of the attended speaker. However, they should also be slower at detecting a switch in attention.

2.11. Psychoacoustic experiment

Although the PESQ score is a reliable MOS [39], we still wanted to test whether users would actually prefer to use our proposed system in a multi-speaker scenario. To do so, we performed a psychoacoustic experiment on healthy controls. Twelve subjects (seven female), aged between 20 and 28 years (mean ± SD, 22 ± 2.5) took part. All subjects reported normal hearing and provided written informed consent. The stimuli used for this experiment were the same as those used for the neural experiment, i.e. subjects were always presented with a mixture of $\mathrm{Spk1_f}$ and $\mathrm{Spk2_m}$. However, the way the stimuli were presented was altered to obtain as much information as possible about the subjects' perception. The experiment was divided into four blocks, each containing 15 trials. Each trial consisted of a single sentence. Before each block, the subjects were instructed to pay attention to one of the speakers (starting with the male) and to switch attention on each successive block. In order to test the intelligibility of the speech, after each trial (sentence) the subjects were presented with a transcription of the sentence of the attended speaker with one word missing. Subjects were instructed to type the missing word (the intelligibility task). They were also asked to indicate the difficulty they had understanding the attended speaker on a scale from 1 to 5: very difficult (1), difficult, not difficult, easy, and very easy (5). From these responses, we calculated the mean opinion score (MOS; [44]). This allowed us to obtain both an objective measure of intelligibility and a subjective MOS. For half of the experiment, both speakers were presented at the same RMS power. For the other half, we attempted to amplify the attended speaker. Block order was counterbalanced across subjects. In total, the experiment lasted approximately 15 min, during which subjects were presented with 4 min and 11 s of audio.

In order to amplify the attended speaker, we decided not to perform a simulation; i.e. not to choose when to perform the amplification. Instead, we decided to use real neural data to demonstrate how the overall system could be implemented. We elected to use the neural data from subject two. To dynamically track the attentional focus of the subject, we implemented a strategy similar to the artificial switching of attention discussed earlier; i.e. we used a sliding window approach, attempting to decode the attention of the subject every second. We also chose to use a window size of 20 s to be consistent with our decoding strategy discussed earlier. Whenever we could correctly classify the attended speaker from the neural data for that subject, we added the output from the correct DNN added to the mixture. However, if a mistake was made, and we misclassified the subject's attentional focus, we would present the output from whichever DNN produced the largest normalized correlation. The DNN output was added at a level of +12 dB relative to the mixture.

In addition to obtaining measures of intelligibility and speech quality, we also wanted to determine the participant's overall preference. Subjects were informed before the experiment that they would have to report which half of the experiment required less effort to understand the attended speaker, and they were reminded half way through. Finally, we wanted to know if users would prefer to use the system if they knew that it wasn't perfect. To test this, we asked one final question at the end of the experiment (after they had reported their initial preference): 'For the (1st/2nd) half of this experiment, a 'system' was turned on that tried to amplify the attended speaker. It is not perfect, and may have sometimes amplified the incorrect speaker. Which would you prefer: to have this system turned on or off?'

3. Results

3.1. DNN output correlation analysis

To examine the ability of the DNNs to separate their designated speakers from the mixtures, we measured the correlation between the output of each DNN and the clean target speaker spectrograms for that DNN (figure 3). We indicate the mean ± SD of all networks that were trained to separate a female (left) or male (right) speaker. The gray bars show performance when a network is presented with a mixture containing the speaker that it was trained to separate, and the red/blue bars show performance when a mixture contains an unknown female/male target speaker. As expected, the networks could not separate undesignated speakers from the mixture, but their outputs were slightly more similar to the target speaker when that speaker was the same gender as the network's designated speaker. This result is likely because of the characteristic differences between male and female speakers (e.g. pitch, spectral envelope). The dotted line illustrates the average correlation between the raw mixtures and the clean target spectrograms.

3.2. Neural correlation analysis

To determine which speaker a subject was attending to, we performed a neural correlation analysis where we compared the reconstructed spectrograms (from the neural data) with the output of each DNN (figure 4). The left of the figure (raw *r*-values) shows the average correlations between the reconstructed spectrograms and the outputs of the DNNs for each subject (where each subject is represented by a colored dot). Because the subjects alternated their attention between two speakers, the *r*-values labeled as *attended* and *unattended* come from the DNNs trained on Spk1_f and Spk2_m. The *r*-values labeled as *undesignated* come from the DNNs trained on Spk3_f and Spk4_m.

Although the attended *r*-values are typically higher than the unattended *r*-values, there is also a high correlation between the reconstructed spectrograms and the mixture. Because the DNNs that do not see their designated speaker usually output spectrograms similar to the mixture, the undesignated correlation values were also relatively high. Therefore, it was crucial to normalize the *r*-values with respect to the mixture. This is an important problem that arises when designing any algorithm that does not use the clean sources. Our solution to this problem involved three key steps: (i) incorporating the mixture into the correlation analysis (equations (2) and (3); see Methods), (ii) performing a test-normalization (*t*-norm) to equalize the outputs of each DNN with respect to each other (equation (4)), and (iii) subtracting the correlation between the DNN output and the mixture (equation (5)). The figure on the right (normalized *r*-values) shows the results of this analysis. After applying the normalization method, the attended correlation values were far higher than either the unattended or undesignated *r*-values, which enabled the decoding of the attended speaker.

3.3. Attention decoding index (ADI)

Given that the normalized correlation values differentiated between attended, unattended, and undesignated *r*-values, it was possible to decode the attentional focus of the subjects using the DNN outputs. After segmenting the data into 20 s chunks, a segment was labeled as a correct hit if the normalized correlation between the reconstructed spectrogram and the output of the DNN that was trained to separate the attended speaker was the highest of all four DNN outputs. A segment was considered to be a false positive if the correlation with the unattended speaker was higher than all others. We define the ADI as the proportion of segments correctly decoded minus the proportion of false positives. For comparison, we also calculated the ADI obtained using the ideal (clean) spectrograms of Spk1_f and Spk2_m (see Methods).

Figure 5(A) shows the results of this analysis. We considered the attentional focus of a subject to be successfully decoded if the ADI was greater than the 0.45 threshold calculated using a random shuffle of the data (gray line; see Methods). Of the six subjects who participated in this study, the attentional focus of subjects one, two and three could be decoded.

Because different regions in the auditory cortex are differentially modulated by attention [11], we sought to explain the variability in ADI across subjects by identifying the anatomical locations of each electrode. For each subject, the pie charts illustrate the proportion of electrodes that were responsive to speech from two anatomical regions: Heschl’s gyrus (HG; red) and superior temporal gyrus (STG; blue). Electrodes that were responsive to speech, but that were not in either of these locations, are collectively referred to as Other (green). These Other electrodes were located in various anatomical regions including the middle and inferior temporal gyri, and planum temporale. The single number displayed above each pie chart refers to the total number of electrodes that were responsive to speech for that subject. To determine which anatomical regions produced the highest ADI, we performed a single electrode analysis (figure 5(B)). Similar to figure 5(A), electrodes are colored according to anatomical location: Heschl’s gyrus (HG; red), superior temporal gyrus (STG; blue), and Other (Green). Electrodes in STG and Other produced ADIs significantly greater than zero (Wilcoxon signed-rank test; $p < 0.001$), compared to the use of electrodes in HG ($p = 0.02$). These results show that electrodes placed in STG are important for successfully decoding attention.

3.4. Dynamic switching of attention

In order to simulate a dynamic situation where subjects alternated their attention between the two speakers, we segmented and concatenated the data into 60 s bins, with a subject's attention switching at the beginning of each section. Figure 6 shows the results of this analysis for an example subject (subject one) using a 20 s window size. Each solid black line marks a switch in attention. The blue bar at the top indicates segments when the subject was attending to Spk2_m and the red bar indicates segments when the subject was attending to Spk1_f. Normalized correlation values for each of the four DNNs are plotted at the bottom of the figure. Ideally, the blue ($mP2_m$) and red ($mP1_f$) lines would alternate being the largest, and the cyan ($mP4_m$) and magenta ($mP3_f$) lines would be close to zero.

From this analysis, we obtained a measure of decoding-accuracy, which we define as the percentage of samples in which the correct target speaker produced the highest normalized correlation value. Unlike the ADI method discussed earlier, this method does not formally take into account false-positives. This is because the sliding window approach inevitably produces false-positives around transitions in attention that are not due to a bias towards a particular speaker, but are instead an inherent property of the method. However, these false-positives will still reduce the decoding-accuracy achievable because the correct target speaker will not produce the highest correlation values in these regions. We observed a decoding-accuracy of 75% for the data displayed in figure 6.

Figure 7(A) displays the same results as shown in figure 6, but averaged over all sections when the subject was attending to Spk2_m (−60 s–0 s) and Spk1_f (0 s–60 s). Shaded regions denote standard error. This analysis was used in order to determine a measure of the transition time (how long it takes to detect a switch in attention for a subject). Transition times were calculated as the time at which the blue (Spk2_m) and red (Spk1_f) lines intersected in the averaged data. This calculation was performed for each subject whose attentional focus we could decode and for all window-sizes (between 5 and 30 s; figure 7(B)). For all subjects, the transition-time increases monotonically with larger window-sizes. We also display the decoding-accuracy that was achieved for each window-size (figure 7(B); bottom panel). These results suggest that window-sizes between 15 and 20 s achieve the best trade-off between speed and accuracy. Longer window-sizes produce a reduction in decoding-accuracy because they also result in longer transition times, and therefore more false-positives after a switch in attention.

3.5. Objective and subjective measures of speech separation quality

While the decoding-accuracy score presented in figure 7 is an important objective measure of performance, the goal of our study was to also enhance the quality of the audio by amplifying the attended speaker relative to the other sound sources. The quality of the output depends on a combination of factors including the accuracy of the DNN speech separation, the neural reconstruction, the attention decoding, and the amplification scheme. We tested the improved quality of the final output audio using both subjective and objective tests. Demos of the final audio output as a subject switches attention are provided online [45].

The output of the system proposed in this study produced an objectively cleaner speech signal with a significant increase in the PESQ score (Wilcoxon signed-rank test, $p < 0.001$; figure 8(A)). To subjectively test the preference of users on the quality of the audio output of our system in a multi-speaker scenario, we performed a psychoacoustic experiment (see Methods). Because it is known that listening effort is increased for those with hearing impairments in multi-speaker scenarios [4, 5], we measured the subjective quality of the output of our system by asking listeners to rate (from 1 to 5) the difficulty of attending to the target speaker (Mean Opinion Score, MOS). All but one subject reported a larger MOS when the system was on, with a median MOS of 3.87 (25th percentile, 3.73; 75th percentile, 4.12) versus a median MOS of 3.5 (25th percentile, 2.9; 75th percentile, 3.8) when the system was off (figure 8(B)), which was a significant increase in MOS (Wilcoxon signed-rank test, $p < 0.001$). The one subject with no increase in MOS had almost identical scores for both system on (3.7) and system off (3.8). The majority of subjects also reported a preference for the segments where the system was turned on (9 out of 12), and a majority reported a preference for using the system once they were informed of its nature (10 out of 12). In the *intelligibility task*, there was no significant difference in the numbers of words that were correctly reported when the system was on versus off. This lack of improved intelligibility is a well-known phenomenon in speech enhancement research where noise suppression does not typically improve intelligibly scores, even though listening effort is reduced [46, 47]. However, it is difficult to conclude whether or not our system changes speech intelligibility, as it is possible that our task was unable to reveal any potential improvements.

4. Discussion

We have developed an end-to-end system that incorporates the latest single-channel automatic speech-separation algorithms into the auditory attention-decoding (AAD) platform. Our proposed system is an important step towards developing realistic cognitively controlled hearing aids that use only the mixed audio of multiple speakers. This approach alleviates the spatial separation requirements of multi-channel approaches, but can also be used in tandem with beamforming methods for optimal source separation [21, 31]. In addition to successfully identifying the attended speaker, our system also amplifies that speaker, resulting in a significant increase in the subjective quality of the listening experience. Combined with the latest developments in AAD research, this work will move the field toward realistic hearing aid devices that can automatically and dynamically track a user’s direction of attention, and amplify an attended speaker.

4.1. Invasive and non-invasive methodologies

We used invasive ECoG recordings from neurological patients [12] because we wanted to determine the varying contributions of different auditory cortical areas to AAD. In addition, ECoG allowed us to determine the achievable decoding speed and accuracy using a neural signal that has a higher signal-to-noise ratio than non-invasive recordings. However, our use of invasive recordings does not limit our approach to only this methodology, as previous work has established the feasibility of applying AAD techniques to non-invasive neural recordings [13, 22]. While non-invasive approaches would be preferable in future hearing aids, there are an increasing number of implantable devices being used to treat many neurological disorders such as vagus nerve stimulation ([48, 49]), responsive neurostimulation [50], cochlear implants [51], and deep brain stimulation [52]. Minimally invasive devices are also being developed that can enter the brain via a blood vessel, thus negating the need for open brain surgery [53]. Because invasive implantable devices can only provide a very limited sampling of brain areas, our finding that implicates STG in AAD is of significant interest.

4.2. System generalization

While the subjects in our study were only presented with mixtures of two speakers ($\mathrm{Spk1}_f$ and $\mathrm{Spk2}_m$), our system can be extended to a more general case of multiple speakers because none of the DNNs were presented with mixtures containing the other target speaker during training (e.g. the DNN trained to separate $\mathrm{Spk1}_f$ never saw $\mathrm{Spk2}_m$ during training). Each DNN is therefore able to separate a trained target speaker from other unseen speakers.

4.3. Limitations

Our proposed system is limited to a closed set of speakers and helps a hearing-impaired user interact with those specific speakers in social settings. The device could switch to default operation when no target speaker is detected. Importantly, a new speaker can be easily added to the system, requiring only a small amount of clean speech from the new target speaker. For our experiments, we found that 20 min of clean speech was sufficient to achieve a satisfactory separation of the target speakers (mean Source-to-Distortion Ratio = 5.52). Because our reconstruction method from neural data is trained using single-speaker data, adding a new speaker does not involve recording new neural responses for that speaker.

A practical limitation for all algorithms intended for hearing aids is that hardware constraints could limit the number of DNNs that could be housed inside a portable device. However, modern hearing aids are able to perform off-board computing by interfacing with a cell phone [6], and specialized hardware is also becoming available for low-power neural network implementations [54–56]. Another consideration is the fact that DNNs rely heavily on the data used to train them. Therefore, additional training would be required to separate speakers under different environmental conditions [57]. Also, because people tend to involuntarily speak louder in noisy situations, which affects acoustic features such as pitch, rate and syllable duration (the Lombard effect [58]), this would also need to be taken into account during training of the DNNs.

4.4. Decoding-accuracy

We discovered that the reconstructed spectrograms in our data had a high correlation with the raw mixture. This is a problem that we needed to address, because a DNN is likely to output the mixture when its designated speaker is not present. The reason why the reconstructed spectrograms had such a high correlation with the mixture can be understood by looking at the single-electrode analysis (figure 5(B)). We found that many speech responsive electrodes are not modulated by attention, and instead encode both the attended and unattended speakers. Therefore, training decoders using all speech-responsive electrodes will lead to reconstructions that are similar to the mixture. This analysis also revealed that the location of electrodes in the brain played an important role in decoding attention. This result is likely due to the variable degree of attentional modulation in various parts of the auditory system [11]. In particular, STG appears to be the cortical area that best encodes attended speech, whereas primary auditory cortex (HG) does not. As a result of this, we were unable to decode the attention of half of the subjects because they had insufficient coverage over STG (the correlation between ADI and the number of electrodes in STG is 0.58). This insight could be useful when using source localization methods in EEG and MEG that can emphasize the neural signals coming from particular brain regions [59]. However, we focused our analysis on the high-gamma band of the neural data (70–150 Hz), which is difficult to measure using non-invasive methods, although not impossible [60–62]. Therefore, the contribution of brain areas to AAD in lower frequency bands of the neural signal needs to be further investigated.

4.5. Dynamic switching of attention

In a real-world situation, it is likely that users would want to dynamically switch their attention between multiple speakers as a conversation progresses. Although the subjects in our study alternated their attention between two speakers, they did not do so in a dynamic fashion; rather, there was a substantial break between each block of the experiment. When simulating switching of attention (figure 6), the window size used for estimation has an effect on both decoding- accuracy and transition-time (the time it takes to detect a switch in attention; figure 7). Our findings indicate that there is an optimal window-size for decoding-accuracy: shorter window sizes produce *r*-values that are too noisy, and longer window sizes prohibit the rapid detection of switches in attention. This problem is particularly important when using neural signals with a lower signal to noise ratio (such as around the ear [23], or in ear EEG [28]). It is possible that more elaborate decoding algorithms can be used to speed up decoding and provide a better trade-off between decoding-accuracy and transition-time [27].

4.6. Psychoacoustics

One important requirement of speech enhancement techniques is to ensure that the resulting speech is not distorted or corrupted, as users tend to prefer no enhancement over an amplified but distorted signal [47]. In our experiments, the automatically separated target speaker was amplified by +12 dB relative to the mixture. This level has been shown to significantly increase the intelligibility of an attended speaker in a two-talker scenario (from ≈88% to 98% [63]). Importantly, an unattended speaker should still be audible so that users can switch their attention should they choose to do so. It is still possible to understand speakers when they are attenuated by 12 dB, although intelligibility drops to ≈78% [63]. These parameters need to be further optimized when our type of system is tested in hearing impaired listeners and in closed-loop setups, where the decoding algorithm and the subject's brain can co-adapt to converge to a suitable solution [64]. We observed a significant increase in the MOS when using our system, and almost all subjects reported that they would prefer to have the system turned on; a finding that supports our systems potential as a useful and effective way to identify and amplify an attended speaker.

Acknowledgments

This research was funded by the National Institute on Deafness and Other Communication Disorders: DC014279.

References

1. Bregman AS, McAdams S. Auditory scene analysis: the perceptual organization of sound. J Acoust Soc Am. 1994; 95:1177–8.
2. Alain, C., Dyson, BJ., Snyder, JS. Handbook of Models for Human Aging. New York: Academic; 2006. Aging and the perceptual organization of sounds: a change of scene; p. 759-69.
3. Mackersie CL. Talker separation and sequential stream segregation in listeners with hearing losspatterns associated with talker gender. J Speech Lang Hear Res. 2003; 46:912–8. [PubMed: 12959469]
4. Peelle JE, Wingfield A. Listening effort in age-related hearing loss. Hear J. 2016; 69:10–12.
5. Peelle JE, Wingfield A. The neural consequences of age-related hearing loss. Trends Neurosci. 2016; 39:486–97. [PubMed: 27262177]
6. Clark JL, Swanepoel DW. Technology for hearing loss—as we know it, and as we dream it. Disabil Rehabil Assist Technol. 2014; 9:408–13. [PubMed: 24712413]
7. Ding N, Simon JZ. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol. 2012; 107:78–89. [PubMed: 21975452]
8. Ding N, Simon JZ. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci USA. 2012; 109:11854–9. [PubMed: 22753470]
9. Horton C, D'Zmura M, Srinivasan R. Suppression of competing speech through entrainment of cortical oscillations. J Neurophysiol. 2013; 109:3082–93. [PubMed: 23515789]
10. Power AJ, Foxe JJ, Forde EJ, Reilly RB, Lalor EC. At what time is the cocktail party? A late locus of selective attention to natural speech. Eur J Neurosci. 2012; 35:1497–503. [PubMed: 22462504]
11. Zion Golumbic EM, et al. Mechanisms underlying selective neuronal tracking of attended speech at a 'cocktail party'. Neuron. 2013; 77:980–91. [PubMed: 23473326]
12. Mesgarani N, Chang EF. Selective cortical representation of attended speaker in multi-talker speech perception. Nature. 2012; 485:233–6. [PubMed: 22522927]
13. O'Sullivan JA, et al. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cereb Cortex. 2015; 25:1697–706. [PubMed: 24429136]
14. Horton C, Srinivasan R, D'Zmura M. Envelope responses in single-trial EEG indicate attended speaker in a 'cocktail party'. J Neural Eng. 2014; 11:046015. [PubMed: 24963838]
15. Dijkstra K, et al. Identifying the attended speaker using electrocorticographic (ECoG) signals. Brain-Comput Interfaces. 2015; 2:161–73.
16. Biesmans W, et al. Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. IEEE Trans Neur Sys Rehab Eng. 2017; 25:402–12.
17. Alickovic, E., Lunner, T., Gustafsson, F. A system identification approach to determining listening attention from EEG signals. 24th European Signal Processing Conf. (EUSIPCO); 2016. p. 31-5.
18. O'Sullivan, JA., Reilly, RB., Lalor, EC. Improved decoding of attentional selection in a cocktail party environment with EEG via automatic selection of relevant independent components. 37th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC); 2015. p. 5740-3.
19. Aroudi, A., Mirkovic, B., De Vos, M., Doclo, S. Auditory attention decoding with EEG recordings using noisy acoustic reference signals. 2016 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP); 2016. p. 694-8.

20. Ekin, B., Atlas, L., Mirbagheri, M., Lee, AK. An alternative approach for auditory attention tracking using single-trial EEG. 2016 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP); 2016. p. 729-33.

21. Van Eyndhoven S, Francart T, Bertrand A. EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses. IEEE Trans Biomed Engin. 2016; 64:1045–56.

22. Mirkovic B, Debener S, Jaeger M, De Vos M. Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. J Neural Eng. 2015; 12:046007. [PubMed: 26035345]

23. Bleichner MG, Mirkovic B, Debener S. Identifying auditory attention with ear-EEG: cEEGrid versus high-density cap-EEG comparison. J Neural Eng. 2016; 13:066004. [PubMed: 27705963]

24. Haghighi, M., Moghadamfalahi, M., Nezamfar, H., Akcakaya, M., Erdogmus, D. Toward a brain interface for tracking attended auditory sources. 2016 IEEE 26th Int. Workshop on Machine Learning for Signal Processing (MLSP); 2016. p. 1-5.

25. Fiedler, L., Obleser, J., Lunner, T., Graversen, C. Ear-EEG allows extraction of neural responses in challenging listening scenarios—a future technology for hearing aids?. 2016 IEEE 38th Annual Int. Conf. of the Engineering in Medicine and Biology Society (EMBC); 2016. p. 5697-700.

26. Das N, Biesmans W, Bertrand A, Francart T. The effect of head-related filtering and ear-specific decoding bias on auditory attention detection. J Neural Eng. 2016; 13:056014. [PubMed: 27618842]

27. Akram S, Presacco A, Simon JZ, Shamma SA, Babadi B. Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. NeuroImage. 2016; 124:906–17. [PubMed: 26436490]

28. Fiedler L, Woestmann M, Graversen C, Brandmeyer A, Lunner T, Obleser J. Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. J Neural Eng. 1988; 14:036020.

29. Van Veen BD, Buckley KM. Beamforming: a versatile approach to spatial filtering. IEEE ASSP Mag. 1988; 5:4–24.

30. Markovich S, Gannot S, Cohen I. Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. IEEE Trans Audio Speech Lang Process. 2009; 17:1071–86.

31. Das, N., Van Eyndhoven, S., Francart, T., Bertrand, A. Adaptive attention-driven speech enhancement for EEG-informed hearing prostheses. 2016 IEEE 38th Annual Int. Conf. of the Engineering in Medicine and Biology Society (EMBC); 2016. p. 77-80.

32. Nordholm SE, Claesson I, Grbic N. Performance limits in subband beamforming. IEEE Trans Speech Audio Process. 2003; 11:193–203.

33. Huang, P-S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P. Deep learning for monaural speech separation. 2014 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP); 2014. p. 1562-6.

34. Weninger, F., Hershey, JR., Le Roux, J., Schuller, B. Discriminatively trained recurrent neural networks for single-channel speech separation. 2014 IEEE Global Conf. on Signal and Information Processing (GlobalSIP); 2014. p. 577-81.

35. Bouchard KE, Mesgarani N, Johnson K, Chang EF. Functional organization of human sensorimotor cortex for speech articulation. Nature. 2013; 495:327–32. [PubMed: 23426266]

36. Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, JR., Schuller, B. Int Conf Latent Variable Analysis and Signal Separation. Berlin: Springer; 2015. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR; p. 91-99.

37. Paul, DB., Baker, JM. The design for the wall street journal-based CSR corpus. HLT '91 Proc. of the workshop on Speech and Natural Language; 1992. p. 357-62.

38. Rix, AW., Beerends, JG., Hollier, MP., Hekstra, AP. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. 2001 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing Proc. (ICASSP′01); 2001. p. 749-52.

39. Grancharov, V., Kleijn, WB. Springer Handbook of Speech Processing. Berlin: Springer; 2008. Speech quality assessment; p. 83-100.

40. Ma J, Hu Y, Loizou PC. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J Acoust Soc Am. 2009; 125:3387–405. [PubMed: 19425678]

41. Mesgarani N, David SV, Fritz JB, Shamma SA. Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. J Neurophysiol. 2009; 102:3329–39. [PubMed: 19759321]

42. Lotia P, Khan MR. A review of various score normalization techniques for speaker identification system. Int J Adv Eng Technol. 2012; 3:650.

43. Macmillan, N., Creelman, C. Detection Theory: a User's Guide. Mahwah, NJ: Lawrence Erlbaum; 2005.

44. Rothauser E, et al. IEEE recommended practice for speech quality measurements. IEEE Trans Audio Electroacoust. 1969; 17:225–46.

45. Mesgarani, N., O'Sullivan, J., Zhuo, Chen. Neural decoding of attentional selection in multi-speaker environments without access to clean sources. Provisional Patent filed June 2016. 2016. (http://naplab.ee.columbia.edu/nnaad.html)

46. Loizou PC, Kim G. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Trans Audio Speech Lang Process. 2011; 19:47–56. [PubMed: 21909285]

47. Loizou, PC. Speech Enhancement: Theory and Practice. Boca Raton, FL: CRC Press; 2013.

48. Schachter SC, Saper CB. Vagus nerve stimulation. Epilepsia. 1998; 39:677–86. [PubMed: 9670894]

49. Amar AP, Levy ML, Liu CY, Apuzzo ML. Vagus nerve stimulation. Proc IEEE. 2008; 96:1142–51.

50. Kossoff EH, et al. Effect of an external responsive neurostimulator on seizures and electrographic discharges during subdural electrode monitoring. Epilepsia. 2004; 45:1560–7. [PubMed: 15571514]

51. Peterson NR, Pisoni DB, Miyamoto RT. Cochlear implants and spoken language processing abilities: review and assessment of the literature. Restor Neurol Neurosci. 2010; 28:237–50. [PubMed: 20404411]

52. Perlmutter JS, Mink JW. Deep brain stimulation. Annu Rev Neurosci. 2006; 29:229–57. [PubMed: 16776585]

53. Synchron. Introducing the Stentrode, the first endovascular neural interface. www.synchronmed.com/

54. Ovtcharov K, Ruwase O, Kim J-Y, Fowers J, Strauss K, Chung ES. Accelerating deep convolutional neural networks using specialized hardware. Microsoft Research Whitepaper. 2015; 2

55. Andri, R., Cavigelli, L., Rossi, D., Benini, L. YodaNN: an ultra-low power convolutional neural network accelerator based on binary weights. 2016 IEEE Computer Society Annual Symp. on VLSI (ISVLSI); 2016. p. 236-41.

56. Lacey G, Taylor GW, Areibi S. Deep learning on fpgas: past, present, and future. 2016 (arXiv: 1602.04283).

57. Li J, Deng L, Gong Y, Haeb-Umbach R. An overview of noise-robust automatic speech recognition. IEEE/ACM Trans Audio Speech Lang Process. 2014; 22:745–77.

58. Brumm H, Zollinger SA. The evolution of the Lombard effect: 100 years of psychoacoustic research. Behaviour. 2011; 148:1173–98.

59. Grosse-Wentrup M, Liefhold C, Gramann K, Buss M. Beamforming in noninvasive brain–computer interfaces. IEEE Trans Biomed Eng. 2009; 56:1209–19. [PubMed: 19423426]

60. Darvas F, Scherer R, Ojemann JG, Rao RP, Miller KJ, Sorensen LB. High gamma mapping using EEG. NeuroImage. 2010; 49:930–8. [PubMed: 19715762]

61. Ball T, et al. Movement related activity in the high gamma range of the human EEG. Neuroimage. 2008; 41:302–10. [PubMed: 18424182]

62. Lenz D, Jeschke M, Schadow J, Naue N, Ohl FW, Herrmann CS. Human EEG very high frequency oscillations reflect the number of matches with a template in auditory short-term memory. Brain Res. 2008; 1220:81–92. [PubMed: 18036577]

63. Brungart DS, Simpson BD, Ericson MA, Scott KR. Informational and energetic masking effects in the perception of multiple simultaneous talkers. J Acoust Soc Am. 2001; 110:2527–38. [PubMed: 11757942]

64. McFarland DJ, Sarnacki WA, Wolpaw JR. Should the parameters of a BCI translation algorithm be continually adapted? J Neurosci Methods. 2011; 199:103–7. [PubMed: 21571004]

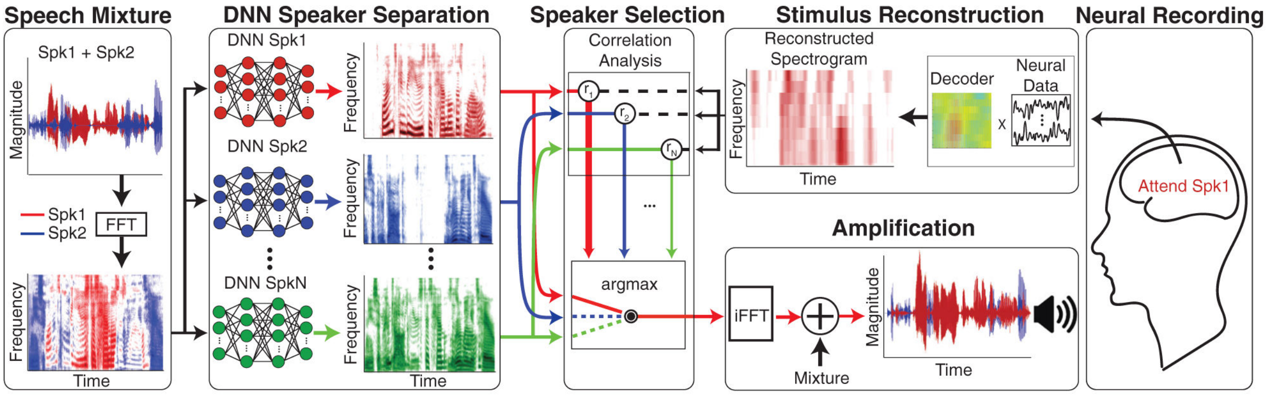

Figure 1.

A schematic of our proposed system. Two speakers, Spk1 (red) and Spk2 (blue), are mixed together into a single acoustic channel. In order to separate the speakers, a spectrogram of the mixture is first obtained (the two speakers have been marked red and blue for visualization purposes only). The spectrogram is then input to each of several DNNs, each trained to separate a specific speaker from a mixture. Simultaneously, a user is attending to one of the speakers (in this case, Spk1; red). A spectrogram of this speaker is reconstructed from the neural recordings of the user. This reconstruction is then compared with the outputs of each of the DNNs using a correlation analysis in order to select the appropriate spectrogram, which is then converted into an acoustic waveform and added to the mixture so as to amplify the attended speaker.

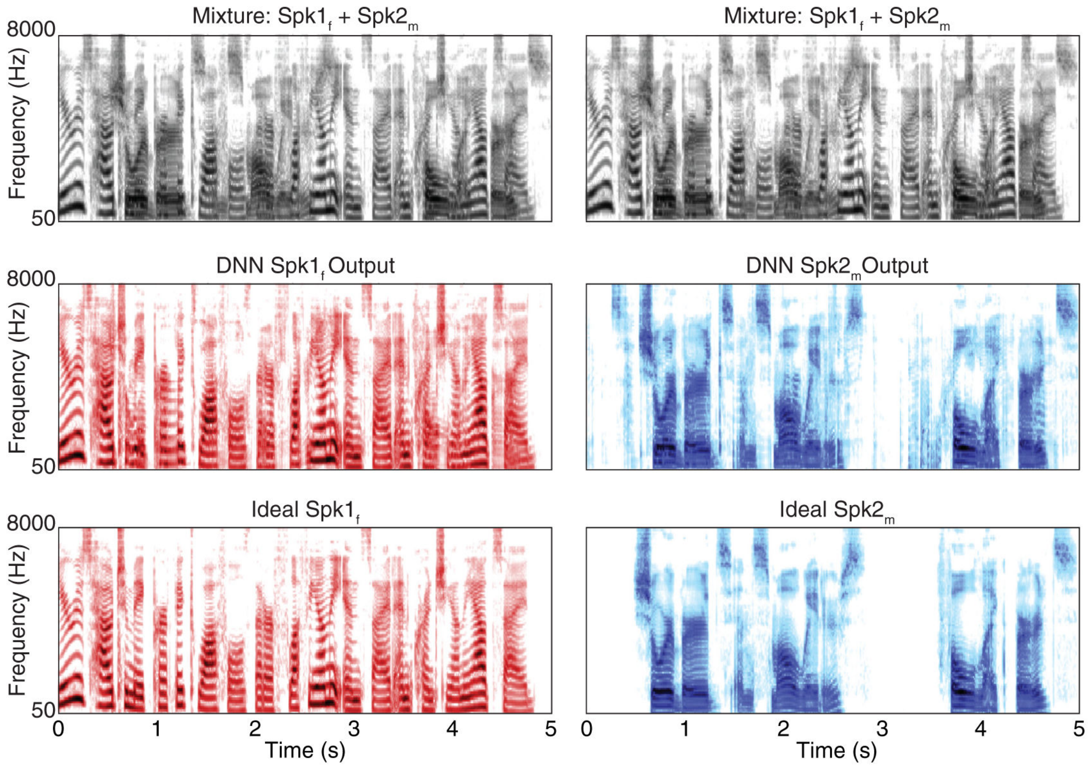

Figure 2.

An example of single-channel speech separation using DNNs. The top two panels display the spectrogram of a mixture of two speakers: Spk1_f (female) and Spk2_m (male). Both panels are the same. The left middle panel displays the output of a DNN that was trained to separate Spk1_f from arbitrary mixtures, and the right middle panel displays the output of a DNN that was trained to separate Spk2_m. The bottom two panels display the ideal clean spectrograms of each speaker, thus representing ideal separation. It is important to note that the DNNs never saw the interfering speakers during training or the utterance of either speaker during testing.

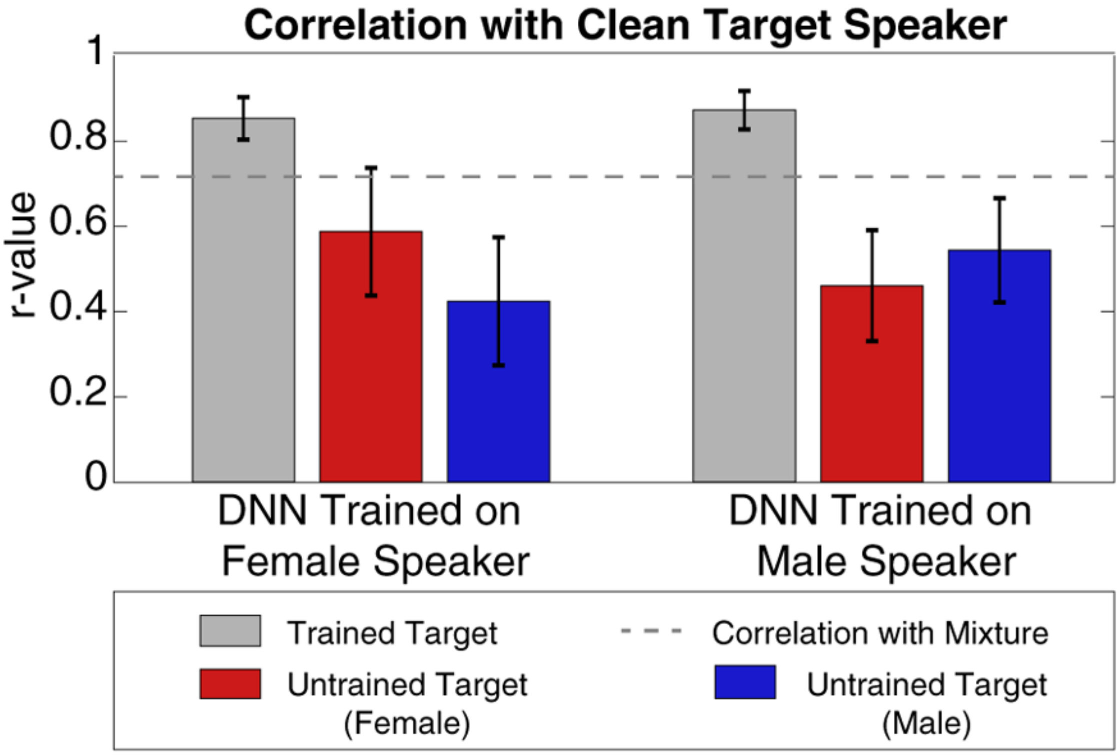

Figure 3.

DNN Output Correlation Analysis. To test the performance of the DNNs at single-channel speaker-separation, we created multiple mixtures of random speakers with four target speakers (two male and two female) and passed every mixture through four DNNs, each pre-trained to separate one of the target speakers. Performance was measured by obtaining the correlation (*r*-value) between the output of each DNN and the spectrogram of the clean target speaker. The results from the four networks are split into two by averaging the *r*-values obtained from the networks that were trained to separate a female (left of figure) or male (right of figure) speaker. The gray bars show the results when a DNN was presented with a mixture containing its designated pre-trained speaker (trained target), and the red and blue bars when the mixture contained an undesignated speaker (untrained target) that was female (red) or male (blue). The dotted line shows the average correlation between the raw mixture and the clean target speaker.

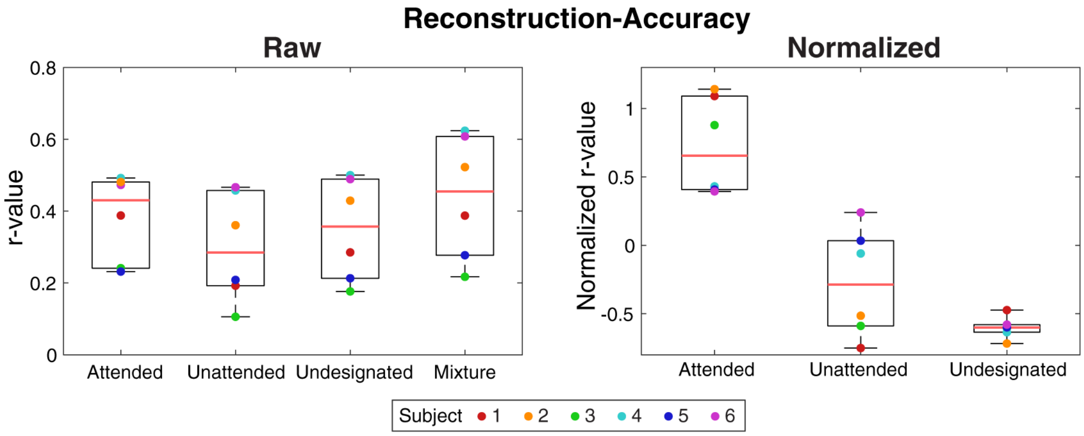

Figure 4.

Reconstruction accuracy. The correlations between the reconstructed spectrograms (from the neural data) and the outputs of the DNNs. Left panels show the raw *r*-values, and right panels show normalized *r*-values (see methods). Each subject is represented by a colored dot. Because the subjects alternated their attention between two speakers, the *r*-values labeled as attended and unattended come from the DNNs trained on Spk1$_f$ and Spk2$_m$, whereas the *r*-values labeled as undesignated come from the DNNs that were trained on Spk3$_f$ and Spk4$_m$. Therefore, undesignated in this sense means that these DNNs were not trained to separate either of the speakers in the mixture that the subjects actually listened to.

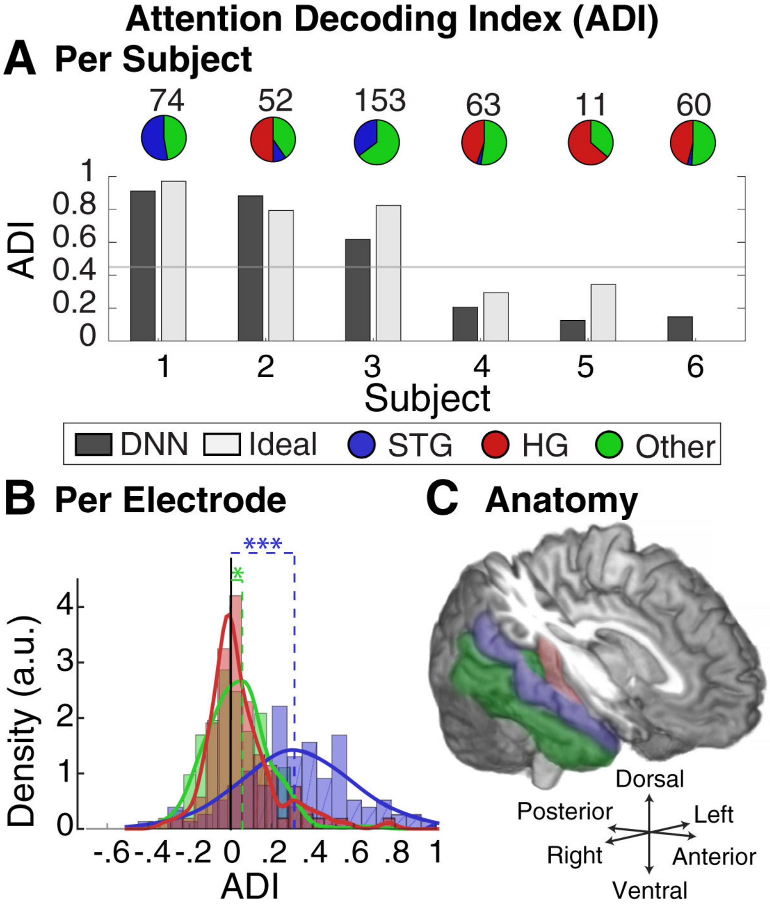

Figure 5.

Attention decoding index (ADI). (A) The proportion of segments (20 s) in which the attentional focus of each subject could be correctly decoded. The gray line (0.45) indicates an ADI significantly above chance (see methods). The pie charts illustrate the proportion of electrodes from two anatomical regions: Heschl's gyrus (HG; red) and superior temporal gyrus (STG; blue). Electrodes that were responsive to speech, but that were not in either of these locations, are collectively referred to as Other (green). These electrodes were located in various anatomical regions including the middle and inferior temporal gyri, and planum temporale. The number that is displayed above each pie chart refers to the total number of electrodes that were responsive to speech for that subject. (B) The ADI for individual electrodes, displayed as a histogram. Bars are colored according to anatomical location. (C) Visualization of the anatomical locations HG (red), STG (blue) and Other (green), from an example subject (subject four). All brain regions above the lateral sulcus in the right hemisphere have been removed in order to expose HG.

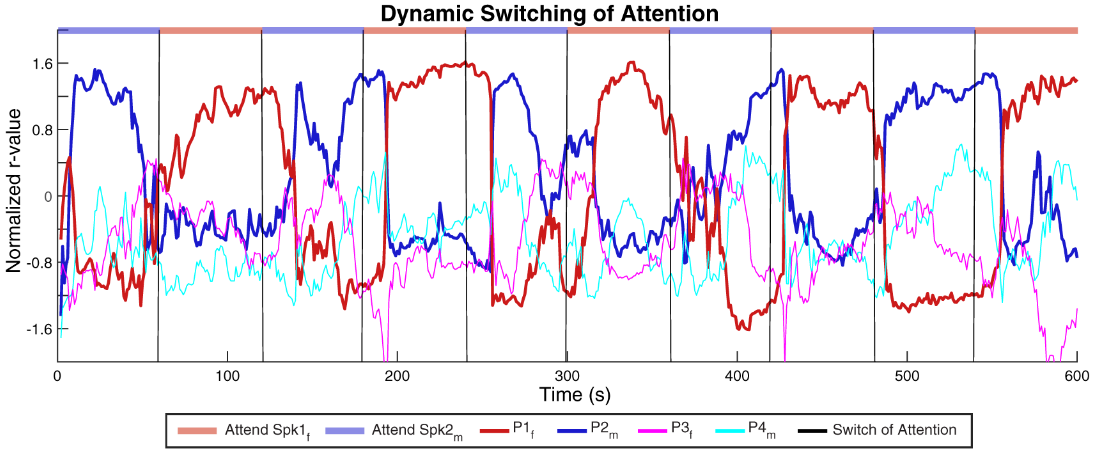

Figure 6.

Dynamic switching of attention. Data were segmented into 60 s chunks and concatenated into consecutive blocks that alternated with regards to the speaker being attending to. The results shown here are for an example subject (subject one). Black lines indicate a switch in attention and the colored bar on top indicates the speaker being attended to (red: Spk1$_f$. Blue: Spk2$_m$). We used a sliding window (20 s width) to obtain normalized *r*-values every second for each of the four DNNs. Ideally, P2$_m$ (blue) and P1$_f$ (red) would alternate in being the largest (corresponding to the speaker being attended), and P3$_f$ (magenta) and P4$_m$ (cyan) would be smallest.

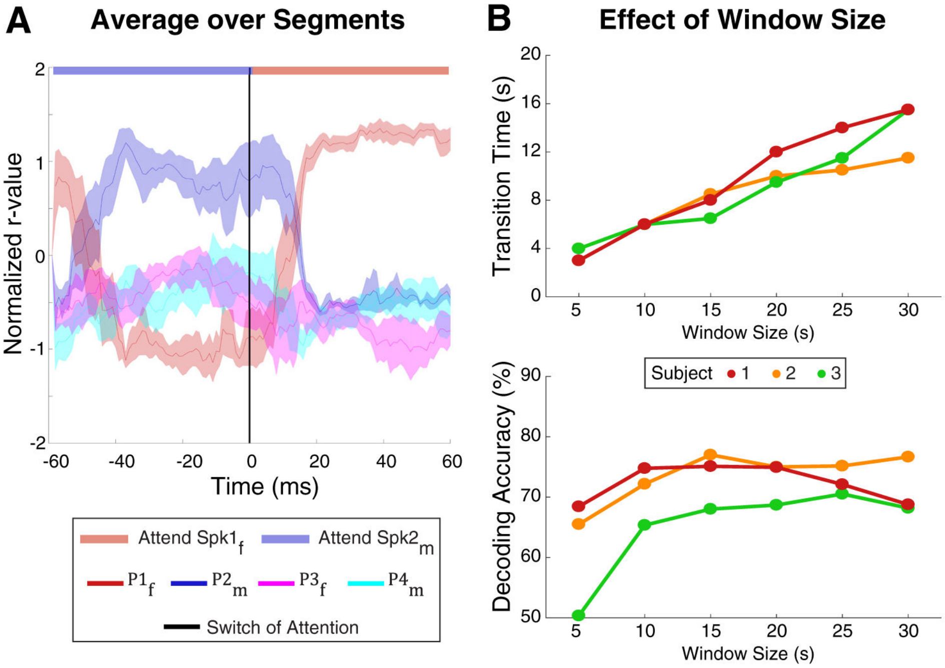

Figure 7.

(A) Average over segments. Here we show the same data as in figure 6, but with the average of all segments aligned to the time of attentional switch (black line) from the male speaker to the female speaker. Data displayed is from the same example subject as in figure 5 (subject one). (B) Effect of window size. The decoding-accuracies and transition-times obtained using a range of window-sizes, and for each subject whose attention could be decoded (subjects one, two and three).

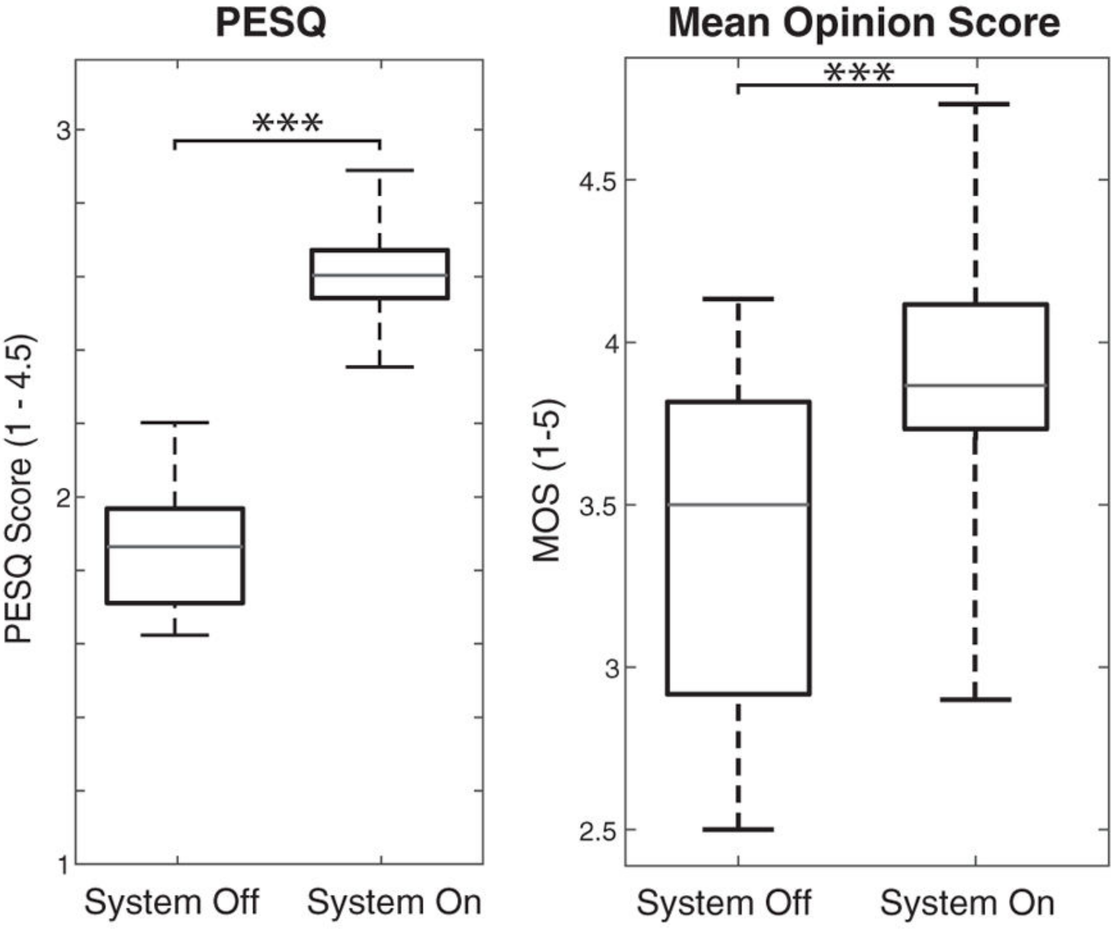

Figure 8.

Speaker separation quality. The left panel displays the PESQ score for the raw mixtures (system off) and the outputs of the DNNs (system on). The PESQ score is an objective measure of speech quality, and ranges between 1 and 4.5 with higher numbers representing better quality. The right panel displays the MOS when the system was off (left) and on (right). The MOS ranges between 1 and 5 with higher numbers representing better quality. There was a significant improvement in both the PESQ and MOS scores when the system was on (right tailed wilcoxon signed-rank test, $p < 0.001$). On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the extrema.