NEUROSCIENCE

Speaker-independent auditory attention decoding without access to clean speech sources

Cong Han[1,2]*, James O'Sullivan[1,2]*, Yi Luo[1,2], Jose Herrero[3], Ashesh D. Mehta[3], Nima Mesgarani[1,2]†

Speech perception in crowded environments is challenging for hearing-impaired listeners. Assistive hearing devices cannot lower interfering speakers without knowing which speaker the listener is focusing on. One possible solution is auditory attention decoding in which the brainwaves of listeners are compared with sound sources to determine the attended source, which can then be amplified to facilitate hearing. In realistic situations, however, only mixed audio is available. We utilize a novel speech separation algorithm to automatically separate speakers in mixed audio, with no need for the speakers to have prior training. Our results show that auditory attention decoding with automatically separated speakers is as accurate and fast as using clean speech sounds. The proposed method significantly improves the subjective and objective quality of the attended speaker. Our study addresses a major obstacle in actualization of auditory attention decoding that can assist hearing-impaired listeners and reduce listening effort for normal-hearing subjects.

INTRODUCTION

Speech communication in acoustic environments with more than one speaker can be extremely challenging for hearing-impaired listeners (*1*). Assistive hearing devices have seen substantial progress in suppressing background noises that are acoustically different from speech (*2*, *3*), but they cannot enhance a target speaker without knowing which speaker the listener is conversing with (*4*). Recent discoveries of the properties of speech representation in the human auditory cortex have shown an enhanced representation of the attended speaker relative to unattended sources (*5*). These findings have motivated the prospect of a brain-controlled assistive hearing device to constantly monitor the brainwaves of a listener and compare them with sound sources in the environment to determine the most likely talker that a subject is attending to (*6*). Then, this device can amplify the attended speaker relative to others to facilitate hearing that speaker in a crowd. This process is termed auditory attention decoding (AAD), a research area that has seen considerable growth in recent years.

Multiple challenging problems, including nonintrusive methods for neural data acquisition and optimal decoding methods for accurate and rapid detection of attentional focus, must be resolved to realize a brain-controlled assistive hearing device. In addition, we have only a mixture of sound sources in realistic situations that can be recorded with one or more microphones. Because the attentional focus of the subject is determined by comparing the brainwaves of the listener with each sound source, a practical AAD system needs to automatically separate the sound sources in the environment to detect the attended source and subsequently amplify it. One solution that has been proposed to address this problem is beamforming (*7*); in this process, neural signals are used to steer a beamformer to amplify the sounds arriving from the location of the target speaker (*8*, *9*). However, this approach requires multiple microphones and can be beneficial only when ample spatial separation exists between the target and interfering speakers. An alternative and possibly complementary method is to leverage the recent success in automatic speech separation algorithms that use deep neural network models (*10*, *11*). In one such approach, neural networks were trained to separate a pretrained, closed set of speakers from mixed audio (*12*). Next, separated speakers were compared with neural responses to determine the attended speaker, who was then amplified and added to the mixture. Although this method can help a subject interact with known speakers, such as family members, this approach is limited in generalization to new, unseen speakers, making it ineffective if the subject converses with a new person, in addition to the difficulty of scaling up to a large number of speakers.

To alleviate this limitation, we propose a causal, speaker-independent automatic speech separation algorithm that can generalize to unseen speakers, meaning that the separation of speakers can be performed without any prior training on target speakers. Speaker-independent speech separation has been one of the most difficult speech processing problems to solve (*13*). In recent years, several solutions have been proposed to address this problem (*11*, *14*, *15*). One such approach is the deep attractor network [DAN; (*10*, *11*)]. DAN performs source separation by projecting the time-frequency (T-F) (spectrogram) representation of a mixed audio signal into a high-dimensional space in which the representation of the speakers becomes more separable. Compared with the alternative speaker-independent approaches (*14*), DAN is advantageous in that it performs an end-to-end separation, meaning the entire process of speaker separation is learned together. However, DAN (*10*, *11*) was proposed for noncausal speech separation, meaning that the algorithm required an entire utterance to perform the separation. In real-time applications, such as in a hearing device, a causal, low-latency algorithm is required to prevent perceivable distortion of the signal (*16*).

In this study, we address the problem of speaker-independent AAD without clean sources using a novel online implementation of DAN [online DAN (ODAN)] to automatically separate unseen sources. Because this system can generalize to new speakers, it overcomes a major limitation of the previous AAD approach that required training on the target speakers (*14*). The proposed AAD framework enhances the subjective and objective quality of perceiving the attended speaker in a multi-talker (M-T) mixture. By combining recent advances in automatic speech processing and brain-computer interfaces, this study represents a major advancement toward solving one of the most

[1]Department of Electrical Engineering, Columbia University, New York, NY, USA. [2]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA. [3]Department of Neurosurgery, Hofstra-Northwell School of Medicine and Feinstein Institute for Medical Research, Manhasset, New York, NY, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: nima@ee.columbia.edu

difficult barriers in actualizing AAD. This solution can help people with hearing impairment communicate more easily.

RESULTS

Figure 1 shows a schematic of the proposed speaker-independent AAD framework. A speaker separation algorithm first separates the speakers in M-T mixed audio. Next, the spectrograms of the separated speakers are compared with the spectrogram that is reconstructed from the evoked neural responses in the auditory cortex of the listener to determine the attended speaker. Then, the attended speaker is amplified relative to other speakers in the mixture before it is delivered to the listener. We describe each of these processing stages in detail below.

Speaker-independent speech separation using the ODAN

Defining the problem of source separation

The problem of speech separation is formulated as estimating C sources, $s_1(t), \ldots, s_c(t) \in \mathcal{R}^{1 \times T}$ from the mixture waveform $x(t) \in \mathcal{R}^{1 \times T}$

$$x(t) = \sum_{i=1}^{C} s_i(t) \tag{1}$$

Taking the short-time Fourier transform (STFT) of both sides formulates the source separation problem in the T-F domain where the complex mixture spectrogram is the sum of the complex source spectrograms

$$X(f,t) = \sum_{i=1}^{C} S_i(f,t) \tag{2}$$

where $X(f,t)$ and $S_i(f,t) \in \mathbb{C}^{F \times T}$. One common approach for recovering the individual sources, S_i, is to estimate a real-valued T-F mask for each source, $M_i \in \mathcal{R}^{F \times T}$, such that

$$|\hat{S}_i(f,t)| = |X(f,t)|\, M_i(f,t) \tag{3}$$

The waveforms of the separated sources are then approximated using the inverse STFT of $|\hat{S}_i(f,t)|$ using the phase of the mixture audio

$$\hat{s}_i(t) = \mathrm{IFFT}(|\hat{S}_i(f,t)| \angle X(f,t)) \tag{4}$$

The mask for each source needs to be estimated directly from the mixture spectrogram

$$M_i = \mathcal{H}(|X(f,t)|; \theta) \tag{5}$$

where $\mathcal{H}(\cdot)$ is the mask estimation model defined by parameter θ.

Speaker-independent speech separation

In real-world scenarios, the identity of speakers in a mixture is usually unknown in advance. Therefore, training separation models using data from target speakers is not possible (*12*). Several recent deep-learning approaches for speaker-independent separation have made significant progress with satisfactory results (*10*, *11*, *14*, *15*). In particular, the DAN aims to directly maximize the reconstruction accuracy of the sources, therefore allowing for end-to-end training of the model (*10*, *11*). However, the DAN method was designed for noncausal speech separation, which means the separation of the speakers at each segment of an incoming audio stream relied on information from the entire mixture utterance. Speech separation in AAD, however, requires real-time implementation, which necessitates a causal algorithm that can separate speakers at each segment using only the current and past inputs. To overcome this challenge, we introduce an online extension of DAN in this study, ODAN. Figure 2A shows the flowchart of the ODAN algorithm. In this novel extension of DAN, source separation is performed by first projecting the mixture spectrogram onto a high-dimensional space where T-F bins belonging to the same source are placed closer together to facilitate their assignment to the corresponding sources. This procedure is performed in multiple steps. First, the mixture magnitude spectrogram, $|X(f,t)|$, is projected onto a tensor, $V(f,t,k)$, where each T-F bin is represented by a vector of length K (Fig. 2B)

$$V = \phi(|X|; \theta) \tag{6}$$

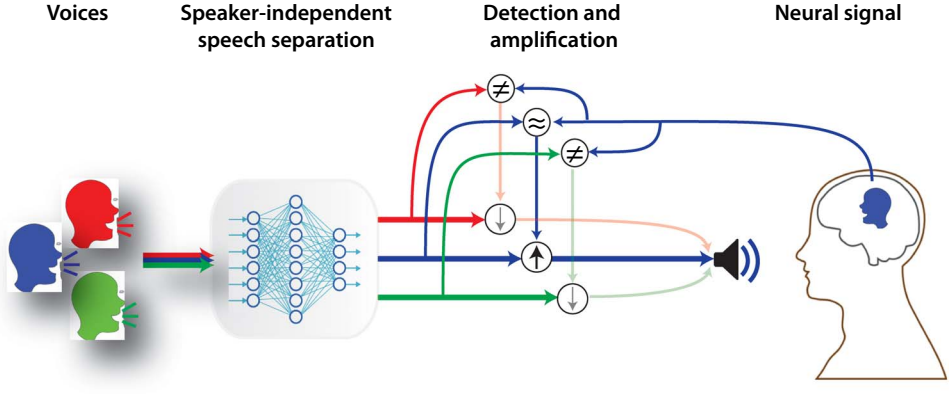

Fig. 1. Schematic of the proposed brain-controlled assistive hearing device. A brain-controlled assistive hearing device can automatically amplify one speaker among many. A deep neural network automatically separates each of the speakers from the mixture and compares each speaker with the neural data from the user's brain to accomplish this goal. Then, the speaker that best matches the neural data is amplified to assist the user.

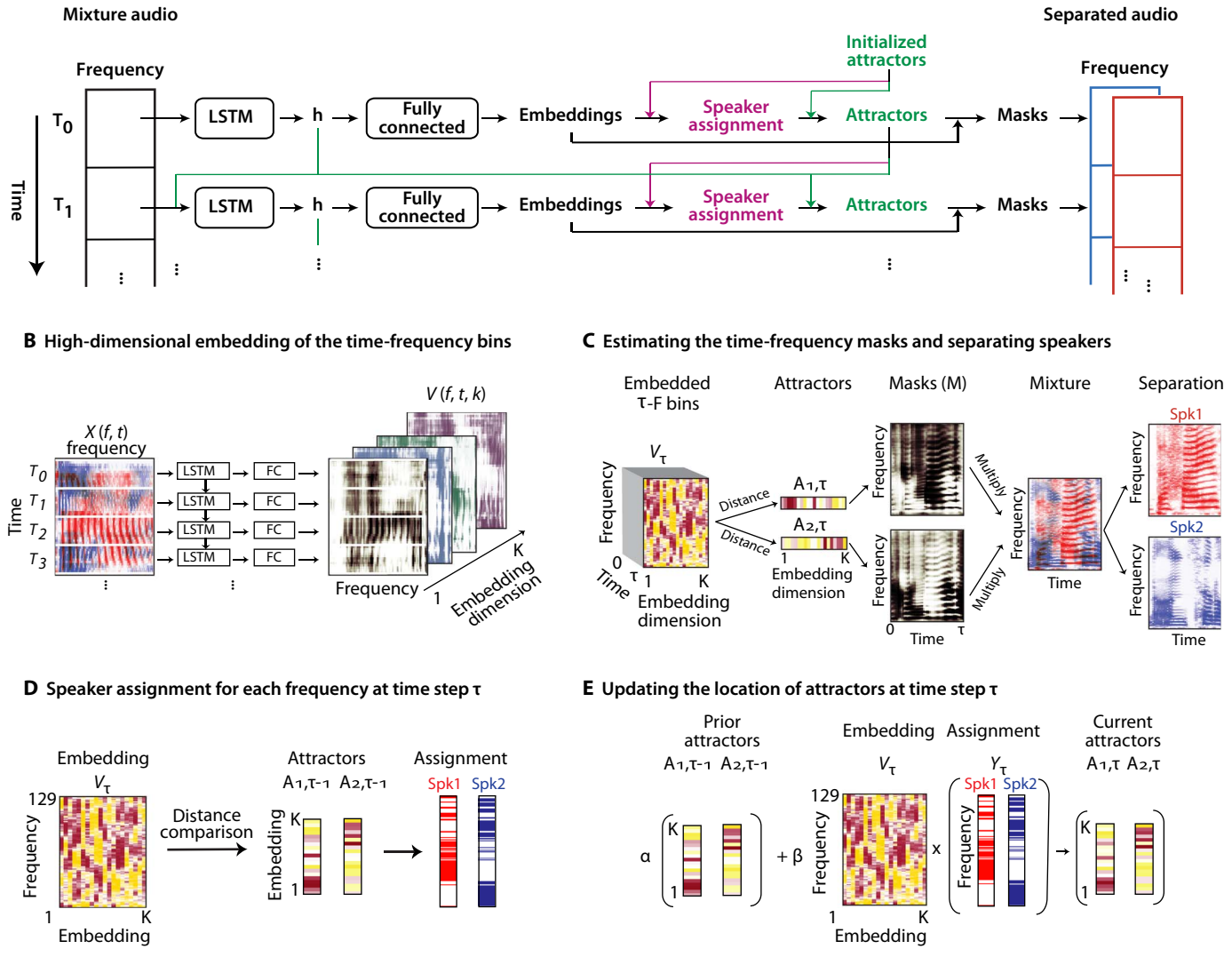

Fig. 2. Speaker-independent speech separation with ODAN. (**A**) The flowchart of the ODAN for speech separation. (**B**) The T-F representation of the mixture sound is projected into a high-dimensional space in which the T-F points that belong to the same speaker are clustered together. (**C**) The center of each speaker representation in the embedding space is referred to as the attractors. The distance between the embedded T-F points and the attractors defines a mask for each speaker that multiplies the T-F representation to extract the speakers. (**D**) The location of the attractors is updated at each time step. First, the previous location of the attractors is used to determine the speaker assignment for the current frame. (**E**) Then, the attractors are updated based on a weighted average of the previous attractors and the center of the current frame defined by the speaker assignments.

where the separation model, $\phi(\cdot)$, is implemented using a deep neural network with parameter θ. We refer to this representation as the embedding space. The neural network that embeds the spectrogram consists of a four-layer long short-term memory (LSTM) network, followed by a fully connected layer (FC) (see Materials and Methods for the details of the network architecture). To assign each embedded T-F bin to one of the speakers in the mixture, we track the centroid of the speakers in the embedding space along time. We refer to the centroids of the source i and at time step τ as the attractor points, $A_{\tau,i}(k)$, because they pull together and attract all the embedded T-F bins that belong to the same source. Therefore, the distance [defined as the dot product (*17*)] between the embedded T-F bins to each of the attractor points determines the source assignment for that T-F bin, which is then used to construct a mask to recover that source (Fig. 2C)

$$M_{\tau,i}(f) = \text{Softmax}\left(\sum_k A_{\tau,i}(k)\, V_\tau(f,k)\right) \quad (7)$$

where the Softmax function is defined as

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_{i=1}^{C} e^{x_i}}$$

The masks subsequently multiply by the mixture magnitude spectrogram to estimate the magnitude spectrograms of each source (Fig. 2C and Eq. 3). All the parameters of the ODAN are found jointly during the

training phase by minimizing the source reconstruction error of the entire utterance,

$$\ell = \sum_{f,t,i} \| \, |S_i(f,t)| - |X(f,t)| \, M_i(f,t) \|_2^2 \quad (8)$$

Online tracking of the attractor points

While DAN uses the embedding of the entire mixture utterance to calculate the attractor points (*10*, *11*), ODAN estimates the attractor locations at each time step using only the current and past inputs. The initial location of the attractor points in the embedding space (at $\tau = 0$) is chosen from a fixed, pretrained set of points in the embedding space (see "Initializing attractor points" section in Material and Methods). Updating the attractor points in each time step is performed using a one-step generalized expectation maximization (EM) algorithm (*18*). At time step τ, we first calculate the source assignment vectors for each speaker, $Y_{\tau,i}(f)$, from the embedded frequency channels $V_\tau(f,k)$ by comparing the distance of each embedded T-F bin to each attractor from the previous time step, $A_{\tau-1,i}(k)$ (Fig. 2D)

$$Y_{\tau,i}(f) = \text{Softmax}\left(\sum_k A_{\tau-1,i}(k)\, V_\tau(f,k)\right) \quad (9)$$

A Softmax function is applied to enhance the source assignment contrast. Next, we update the location of the attractors based on the centroid of the current frame, the previous location of the attractors, and the current input (Fig. 2E)

$$A_{\tau,i}(k) = (1-\alpha_{\tau,i}(k))A_{\tau-1,i}(k) + \alpha_{\tau,i}(k)C_{\tau,i}(k)$$

$$C_{\tau,i}(k) = \frac{\sum_f V_\tau(f,k)Y_{\tau,i}(f)}{\sum_f Y_{\tau,i}(f)} \quad (10)$$

where $C_{\tau,i}(k)$ is the centroid of the embeddings of source i at time step τ, and parameter α determines the rate of the update at time τ by controlling the trade-off between the previous location of the attractors and the centroid of the sources in the current frame

$$\alpha_{\tau,i}(k) = \frac{\sum_f Y_{\tau,i}(f)}{Q_{\tau,i}(k)\sum_{t=0}^{\tau-1}\sum_f Y_{t,i}(f) + \sum_f Y_{\tau,i}(f)}$$

where $Q_{\tau,i}(k)$ determines the contribution of attractor history and the current attractor estimates at time step τ. Parameter $Q_{\tau,i}(k)$ at each time frame is calculated by a neural network from the current frequency vector ($|X(f,\tau)|$, the output of the LSTM layer in the last time step, and the previous location of attractors (green lines in Fig. 2A; see "Calculating the updated rate of attractors" section in Materials and Methods). Once the attractors for the current frame are updated, the masks for separating the current frame are calculated using the similarity of the T-F embeddings and each attractor (Fig. 2C).

Evaluating speech separation accuracy

As shown in Eq. 6, ODAN projects T-F bins into a high-dimensional embedding space that is optimal for source separation (Eq. 8), meaning that T-F bins belonging to the same source should be placed closer to each other in the embedding space. To confirm that this situation is the case, we projected the representation of the two speakers in both the spectrogram domain and embedding domain onto a two-dimensional space using principal components analysis (*19*) to allow visualization. This improved separability of the speakers is shown in Fig. 3A, where the representations are visualized using the first two principal components of the spectrogram and embedding space. Accordingly, T-F bins with more power for each speaker are shown in red and blue, and the improved separation in the embedding space is evident from the decreased overlap between red and blue dots in the embedding space (Fig. 3A).

We evaluated the ODAN model on single-channel, two-speaker and three-speaker separation tasks. We used the WSJ0-2mix and WSJ0-3mix datasets generated from the *Wall Street Journal* (WSJ0) because it is commonly used for comparison with state-of-the-art speaker separation systems. This dataset contains 30 hours of training data, 10 hours of validation data, and 5 hours of test data. The mixed sounds are generated by randomly selecting utterances from different speakers in the WSJ0 training set and mixing them at various signal-to-noise ratios (SNRs), randomly chosen between −2.5 and 2.5 dB. Table 1 shows the comparison of the ODAN method with other state-of-the-art speaker-independent speech separation methods on two-speaker and three-speaker mixtures. The evaluation is conducted using the signal-to-distortion ratio (SDR), scale-invariant SNR (SI-SNR) (*11*), perceptual evaluation of speech quality (PESQ) score (*20*), and extended short-term objective intelligibility (ESTOI) score (*21*) (Materials and Methods). As seen in Table 1, the ODAN method performs well in separating speakers in the mixture and even performs on par with the noncausal DAN method, which computes the separation from the entire utterance using a global clustering of the embeddings. We also tested the ability of the ODAN in dealing with an unknown number of speakers in the mixture. This was done by assuming the maximum number of speakers to be three and training the algorithm on both two-speaker (WSJ0-2mix) and three-speaker (WSJ0-3mix) datasets. During the test phase, no information about the number of speakers was provided, and the outputs that have low power (less than 20 dB relative to the other outputs) were discarded. As seen in Table 2, the same ODAN network can successfully separate one-, two-, or three-speaker mixtures without any prior information on the number of sources in the mixture during the test phase. In addition, we tested whether ODAN can adapt and perform separation even when speakers in the mixture change over time, which frequently occurs in real-world situations. We concatenated mixtures of different speakers where the speakers in the mixture change every 4 s. Figure 3B shows the mean squared error (MSE) between the separated speech and actual speaker spectrograms over time, where the line at 4 s indicates the time of speaker change. Figure 3B shows that ODAN converges to new mixtures in less than 1.2 s (*t* test, $P < 0.05$) by adapting to new speakers to correctly separate them. This ability to track the speakers is important and enables it to work in real-world acoustic scenarios.

Behavioral AAD experiment and neural measurements

Neural recordings

To test the feasibility of using the ODAN speech separation network in a brain-controlled hearing device, we used invasive electrophysiology to measure neural activity from three neurosurgical patients undergoing treatment for epilepsy. Two subjects (subjects 1 and 2) were implanted with high-density subdural electrocorticography (ECoG) arrays over their language dominant temporal lobe, providing coverage of the superior temporal gyrus (STG), which selectively represents attended

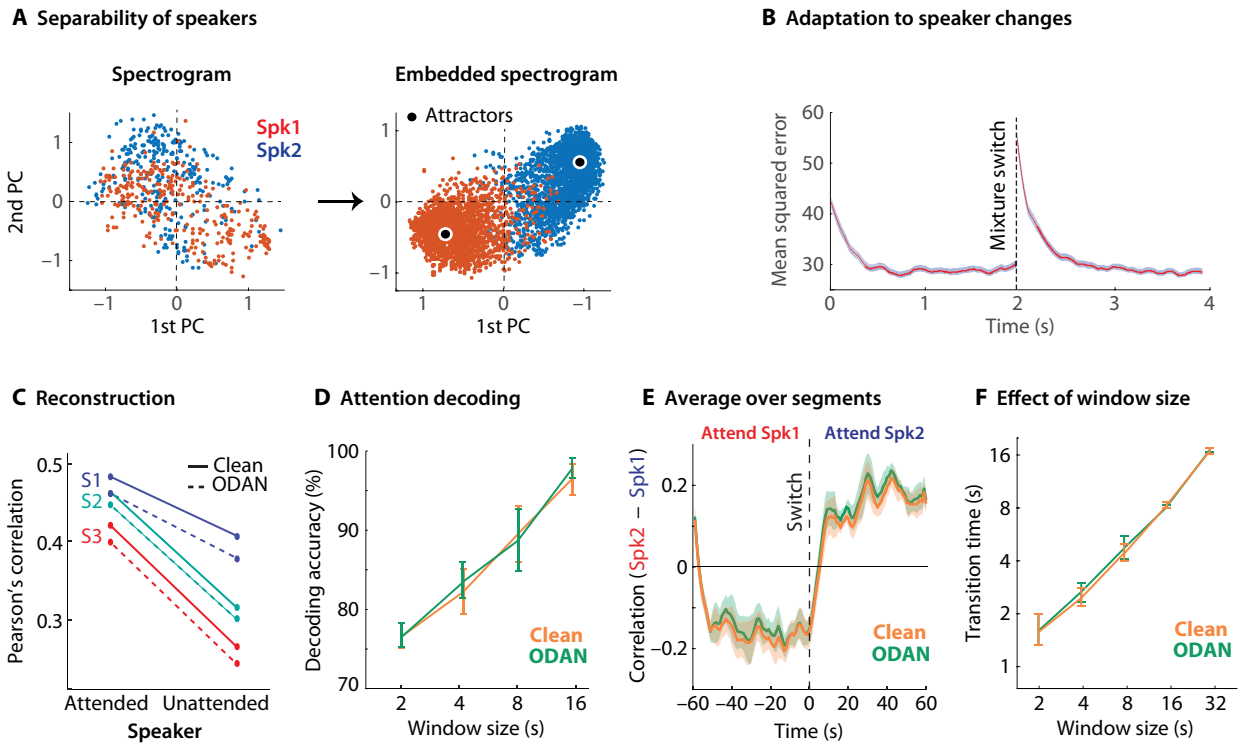

Fig. 3. Evaluating the accuracy of speech separation and attention decoding methods. (**A**) Comparison of separation between the representation of the two speakers in the T-F (left) and embedding space (right). The axis represents the first two principal components of the data that are used to allow visualization. Each dot represents one T-F bin (left) or one embedded T-F bin (right), which are colored based on the relative power of the two speakers in that bin. (**B**) Separation accuracy as a function of time. The dashed line shows the time at which the speakers in the mixture are switched. (**C**) Correlation values between the reconstructed spectrograms (from neural data) and the attended/unattended spectrograms. Correlation values were significantly higher for the attended speaker (paired *t* test, $P < 0.001$; Cohen's $D = 0.8$), thus confirming the effect of attention in the neural data. The correlation with the clean spectrograms was slightly higher than that with the ODAN outputs, but the differences between the attended and unattended speakers were the same for both clean and ODAN outputs. (**D**) Attention decoding: The percentage of segments in which the attended speaker was correctly identified for a varying number of correlation window lengths when using ODAN and the actual clean spectrograms. There was no significant difference between using the clean and the ODAN spectrograms (Wilcoxon rank sum test, $P = 0.9$). (**E**) Dynamic switching of attention was simulated by segmenting and concatenating the neural data into alternating 60-s bins. The dashed line indicates switching attention. The average correlation values from one subject are shown using a 4-s window size for both ODAN and the actual clean spectrograms. The shaded regions denote SE. (**F**) The transition time in detecting a switch of attention was calculated as the time at which the correlation difference between the two speakers crossed zero. The average transition time across subjects increased with larger window sizes; however, there was no significant difference between the transition time of ODAN and the actual clean spectrograms (Wilcoxon rank sum test, $P > 0.6$).

speech (*5*). The third subject was implanted with bilateral stereoelectroencephalography (sEEG), with depth electrodes in Heschl's gyrus (containing primary auditory cortex) and STG. This implantation resulted in varying amounts of coverage over the left and right auditory cortices of each subject (fig. S1). All subjects had self-reported normal hearing and consented to participate in the experiment.

Each subject participated in the following experiments for this study: single-talker (S-T) and M-T experiments. In the S-T experiment, each subject listened to four continuous speech stories (each story was 3 min long), for a total of 12 min of speech material. The stories were uttered once by a female and once by a male speaker (hereafter referred to as Spk1 and Spk2, respectively). For the M-T experiment, the subjects were presented with a mixture of the same speech stories as those in the S-T experiment, where both speakers were combined at a 0-dB target-to-masker ratio. The M-T experiment was divided into four behavioral blocks, each containing a mixture of two different stories spoken by Spk1 and Spk2. Before each experimental block, the subjects were instructed to focus their attention on one speaker and to ignore the other. All the subjects began the experiment by attending to the male speaker and switched their attention to the alternate speaker on each subsequent block. To ensure that the subjects were engaged in the task, we intermittently paused the stories and asked the subjects to repeat the last sentence of the attended speaker before the pause. All the subjects performed the task with high behavioral accuracy and were able to report the sentence before the pause with an average accuracy of 90.5% (S1, 94%; S2, 87.%; and S3, 90%). Speech sounds were presented using a single loudspeaker placed in front of the subject at a comfortable hearing level, with no spatial separation between the competing speakers.

Reconstruction of the attended speaker from evoked neural activity

The reconstructed spectrogram from the auditory cortical responses of a listener in an M-T speech perception task is more similar to the spectrogram of the attended speaker than that of the unattended speaker (*5*). Therefore, the comparison of the neurally reconstructed spectrogram with the spectrograms of individual speakers in a mixture can determine the attentional focus of the listener (*6*). We used a linear reconstruction method (*22*) to convert neural responses back

Table 1. Comparison of speech separation accuracy of ODAN with two other methods for separating two-speaker mixtures (WSJ0-mix2 dataset) and three-speaker mixtures (WSJ0-mix3 dataset). The separation accuracy of ODAN, which is the causal system, is slightly worse but comparable to the other noncausal methods.

Number of Speakers	Model	Causal	SI-SNRi (dB)	SDRi (dB)	PESQ	ESTOI
Two speakers	Original mixture	–	0	0	2.02	0.56
	DAN-LSTM (*11*)	No	9.1	9.5	2.73	0.77
	uPIT-LSTM (*15*)	Yes	–	7.0	–	–
	ODAN	Yes	9.0	9.4	2.70	0.77
Three speakers	Original mixture	–	0	0	1.66	0.39
	DAN-LSTM (*11*)	No	7.0	7.4	2.13	0.56
	uPIT-BLSTM (*15*)	No	–	7.4	–	–
	DPCL++ (*50*)	No	7.1	–	–	–
	ODAN	Yes	6.7	7.2	2.03	0.55

Table 2. Speech separation accuracy of ODAN in separating one-, two-, and three-speaker mixtures (WSJ0-mix2 and WSJ0-mix3 datasets). The ODAN was trained on both the WSJ0-mix2 and WSJ0-mix3 datasets and used in all cases.

Number of speakers	Causal	SI-SNRi (dB)	SDRi (dB)	PESQ	ESTOI
3	Yes	7.0	7.5	2.08	0.56
2	Yes	8.9	9.3	2.63	0.76
1	Yes	**SI-SNR (dB)** 24.4	**SDR (dB)** 25.0	4.14	0.98

to the spectrogram of the sound. This method calculates a linear mapping between the response of a population of neurons to the T-F representation of the stimulus (*22*). This mapping is performed by assigning a spatiotemporal filter to the set of electrodes, which is estimated by minimizing the MSE between the original and the reconstructed spectrograms. We estimated the reconstruction filters using only the neural responses to speech in the S-T experiment. Then, we fixed the filters and used them to reconstruct the spectrogram in the M-T experiments under different attention focuses.

To examine the similarity of the reconstructed spectrograms from the neural responses to the spectrograms of the attended and unattended speakers, we measured the correlation coefficient (Pearson's r) between the reconstructed spectrograms with both ODAN and the actual clean spectrograms of the two speakers. The correlation values were estimated over the entire duration of the M-T experiment. As shown in Fig. 3C, the correlation between the reconstructed and clean spectrograms was significantly higher for the attended speaker than for the unattended speaker (paired t test, $P < 0.001$; Cohen's $D = 0.8$). This observation shows the expected attentional modulation of the auditory cortical responses (*5*). The comparison of the correlation values of ODAN and the actual clean spectrograms (Fig. 3C) shows a similar difference value between the attended and unattended spectrograms (average correlation difference for clean = 0.125 and for ODAN = 0.128), suggesting that ODAN spectrograms can be equally effective for attention decoding. Figure 3C also shows a small but significant decrease in the correlation values of the reconstructed spectrograms with ODAN compared with those of the actual clean spectrograms. This decrease is caused by the imperfect speech separation performed by the ODAN algorithm. Nevertheless, this difference is small and equally present for both attended and unattended speakers. Therefore, this difference did not significantly affect the decoding accuracy as shown below.

Decoding the attentional focus of the listener

To study how the observed reconstruction accuracy with attended and unattended speakers (Fig. 3C) translates into attention decoding accuracy, we used a simple classification scheme in which we computed the correlation between the reconstructed spectrograms with both clean attended and unattended speaker spectrograms over a specified duration. Next, the attended speaker is determined as the speaker with a higher correlation value. The duration of the signal used for the calculation of the correlation is an important parameter and affects both the decoding accuracy and speed. Longer durations increase the reliability of the correlation values, hence improving the decoding accuracy. This phenomenon is shown in Fig. 3D, where the varying duration of the temporal window was used to determine the attended speaker. The accuracy in Fig. 3D indicates the percentage of segments for which the attended speaker was correctly decoded. The accuracy was calculated for the following cases: when using ODAN spectrograms and when using the actual clean spectrograms. We found no significant difference in decoding accuracy with ODAN or the clean spectrograms when different time windows were used (Wilcoxon rank sum test, $P = 0.9$). This finding confirms that automatically separated sources by the ODAN algorithm result in the same attention decoding accuracy as that with the actual clean spectrograms. As expected, increasing the correlation window resulted in improved decoding accuracy for both ODAN and actual clean sources (Fig. 3D).

Next, we examined the temporal properties of attention decoding when ODAN and the actual clean spectrograms were used. We simulated a dynamic switching of attention where the neural responses were concatenated from different attention experiment blocks such that the neural data alternated between attending to the two speakers. To accomplish this, we first divided the neural data in each experiment block into 60-s segments (total of 12 segments) and interleaved segments from the two attention conditions (see Materials and Methods). We compared the correlation values between the reconstructed spectrograms with both ODAN and the actual clean spectrograms using a sliding window of 4 s. Then, we averaged the correlation values over the segments by aligning them according to the time of the attention switch. Figure 3E shows the average correlation for one example subject over all the segments where the subject was attending to Spk1 in the first 60 s and switched to Spk2 afterward. The overlap between the correlation plots calculated from ODAN and the actual clean spectrograms shows that the temporal properties of attention decoding are the same in both cases; hence, ODAN outputs can replace the clean spectrograms without any significant decrease in decoding speed. We quantified the decoding speed using the transition time, which is the time it takes to detect a switch in the listener's attention. Transition times were calculated as the time at which the average correlation crossed the zero line. Figure 3F shows the average transition times for the three subjects

for five different sliding window durations. As expected, the transition times increase for longer window lengths, but there were no significant differences between ODAN and the clean spectrograms (paired t test, $P > 0.7$; Fig. 3F).

Increased subjective and objective perceived quality of the attended speaker

To test if the difficulty of attending to the target speaker is reduced using the ODAN-AAD system, we performed a psychoacoustic experiment comparing the original mixture and sounds in which the decoded target speaker was amplified by 12 dB (Materials and Methods) (see movie S1 and online at naplab.ee.columbia.edu/NNAAD for a demo of the end-to-end ODAN-AAD system). This particular amplification level has been shown to significantly increase the intelligibility of the attended speaker while keeping the unattended speakers audible enough to enable attention switching (*23*). Subjects were asked to rate the difficulty of attending to the target speaker in three conditions when listening to the following: (i) the raw mixture, (ii) the enhanced target speech using the output of the ODAN-AAD, and (iii) the enhanced target speech using the output of the clean-AAD system. Twenty listeners with normal hearing participated in the psychoacoustic experiment, where they each heard 20 sentences in each of the three experimental conditions in random order. Subjects were instructed to attend to one of the speakers and report the difficulty of focusing on that speaker. Subjects were asked to rate the difficulty on a scale of 1 to 5 using the mean opinion score [MOS; (*24*)]. The bar plots in Fig. 4A show the median MOS ± standard error (SE) for each of the three conditions. The average subjective score for the ODAN-AAD shows a significant improvement over the mixture (56% improvement; paired t test, $P < 0.001$), demonstrating that the listeners had a stronger preference for the modified audio than for the original mixture. Figure 4A also shows a small but significant difference between the average MOS score with the actual clean sources and that with ODAN separated sources (78% versus 56% improvement over the mixture). The MOS values using the clean sources show the upper bound of AAD improvement if the speaker separation algorithm was perfect. Therefore, this analysis illustrates the maximum extra gain that can be achieved by improving the accuracy of the speech separation algorithm (14% over the current system). Figure 4B shows a similar analysis when an objective perceptual speech quality measure is used [PESQ; (*20*)], showing a result similar to what we observed in the subjective tests. Together, Fig. 4 demonstrates the benefit of using the ODAN-AAD system in improving the perceived quality of the target speaker.

DISCUSSION

We present a framework for AAD that addresses the lack of access to clean speech sources in real-world applications. Our method uses a novel, real-time, speaker-independent speech separation algorithm that uses deep-learning methods to separate the speakers from a single channel of audio. Then, the separated sources are compared with the reconstructed spectrogram from the auditory cortical responses of the listener to determine and amplify the attended source. The integration of speaker-independent speech separation in the AAD framework is also a novel contribution. We tested a system on two unseen speakers and showed improved subjective and objective perception of the attended speaker when using the ODAN-AAD framework.

A major advantage of our system over previous work (*12*) is the ability to generalize to unseen speakers, which enables a user to communicate more easily with new people. Because ECoG electrodes reflect the summed activity of thousands of neurons in the proximity of the electrodes (*25*), the spectral tuning resolution of the electrodes is relatively low (*26*). As a result, the reconstruction filters that map the neural responses to the stimulus spectrogram do not have to be trained on specific speakers and can generalize to novel speakers, as we have shown previously (*5*, *27*). Nonetheless, generalization to various noisy, reverberant acoustic conditions is still a challenging problem and requires training on a large amount of data recorded from as many noisy environments as possible (*3*). Recent studies have shown the feasibility of using neural network models in joint speech separation and denoising (*28*), which will be needed in a real-world implementation of AAD. Moreover, similar speech processing approaches, such as automatic speech recognition, have seen great benefit from large-scale training whenever possible (*29*, *30*). Therefore, speech separation is also expected to obtain a similar benefit in robustness to adverse acoustic conditions. In addition to increasing the amount of training data and training conditions, separation accuracy can be significantly improved when more than one microphone can be used to record mixed audio. The advantage of enhancing speech with multiple microphones has been previously demonstrated (*31*, *32*), particularly in severely noisy environments or when the number of competing speakers is large (e.g., more than three).

One major limitation in advanced signal processing approaches for hearing technologies is the limited computation and power resources that are available in wearable devices. Nevertheless, designing specialized hardware that can efficiently implement deep neural network models is an active research area that has recently seen substantial progress (*33*–*35*). Specialized hardware also significantly reduces the power consumption needed for computation. In addition, hearing aid devices can already perform off-board computation by interfacing with a remote device, such as a mobile phone, which provides another possibility for extending the computational power of these devices (*2*).

Although we used invasive neural recordings to test our system, previous research has already shown that attention decoding is also possible with noninvasive neural recordings, including scalp EEG with different or the same gender mixtures (*6*), around the ear EEG electrodes (*36*), and in-ear EEG recordings (*37*). The SNR of these recordings is not as high as that of invasive methods, but they can provide enough information needed to decode the attentional focus (*6*, *36*, *37*), although this may come at the expense of reducing the decoding speed of the AAD. Alternatively, several recent studies have examined the efficacy of minimally invasive neural recording techniques where the electrodes are placed under the skin without penetrating the bone (*38*). Further advances in noninvasive neural recording from the human brain can further increase the fidelity of the neural recording to improve both the accuracy and speed of attention decoding.

The accuracy of AAD also critically depends on the decoding algorithm being used (*39*, *40*). For example, the accuracy and speed of decoding can be improved when stochastic models are used to estimate the attention focus using a state-space model (*41*) instead of the moving average that we used in this paper. In addition, while we used fixed reconstruction filters derived from the S-T responses, this experimental condition may not always be available. In these scenarios, it is possible to circumvent the need for S-T responses by online estimation of the encoding/decoding coefficients from the responses to the mixture (*41*, *42*), which may lead to more flexible and robust estimation of the decoding filters. Last, decoding methods that factor in the head-related filtering of the sound can also improve the attention decoding accuracy (*43*).

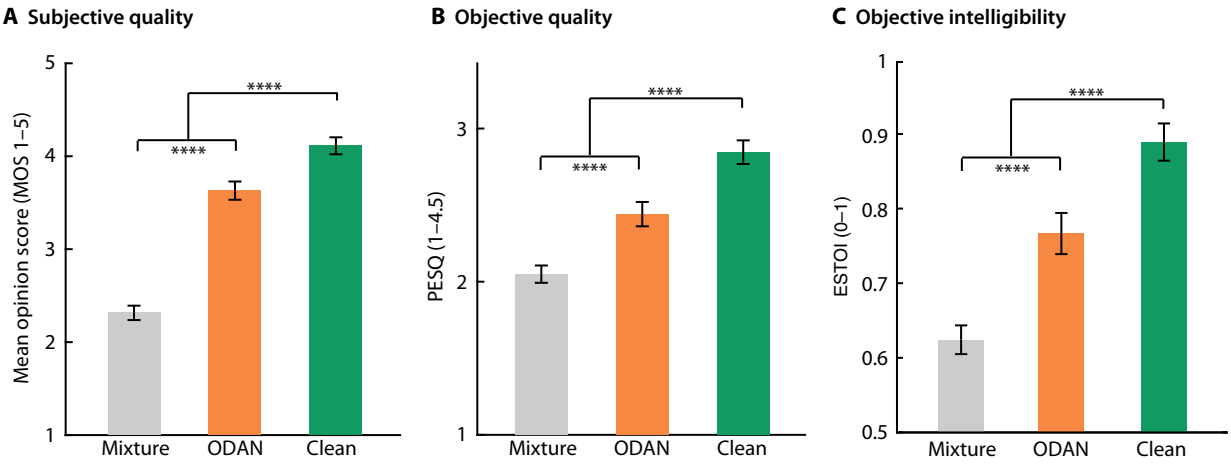

Fig. 4. Improved subjective quality and objective quality and intelligibility of the ODAN-AAD system. (**A**) Subjective listening test to determine the ease of attending to the target speaker. Twenty healthy subjects were asked to rate the difficulty of attending to the target speaker when listening to (i) the raw mixture, (ii) the ODAN-AAD amplified target speaker, and (iii) the clean-AAD amplified target speaker. The detected target speakers in (ii) and (iii) were amplified by 12 dB relative to the interfering speakers. Subjects were asked to rate the difficulty on a scale of 1 to 5 (MOS). The bar plots show the median MOS ± SE for each condition. The enhancement of the target speaker for the ODAN-AAD and clean-AAD systems was 100 and 118%, respectively ($P < 0.001$). (**B** and **C**) Objective quality (PESQ) and intelligibility (ESTOI) improvement of the target speech in the same three conditions as in (A). ****$P < 0.0001$, *t* test.

In summary, our proposed speaker-independent AAD system represents a feasible solution for a major obstacle in creating a brain-controlled hearing device, therefore bringing this technology a step closer to reality. Such a device can help hearing-impaired listeners more easily communicate in crowded environments and reduce the listening effort for normal-hearing subjects, therefore reducing listening fatigue.

MATERIALS AND METHODS

Participants

Three subjects who were undergoing clinical treatment for epilepsy at North Shore University Hospital participated in this study. All patients provided informed consent as monitored by the local institutional review board and in accordance with the ethical standards of the Declaration of Helsinki. The decision to implant the electrode targets and the duration of implantation were made entirely on clinical grounds without reference to this investigation. Patients were informed that participation in this study would not alter their clinical treatment and that they could withdraw at any time without jeopardizing their clinical care. Two subjects (subjects 1 and 2) were implanted with high-density subdural electrode arrays over their left (language dominant) temporal lobe with coverage over the STG. The remaining subject partook in sEEG, in which he or she was implanted bilaterally with depth electrodes. The coverage over the left and right auditory cortices for each subject is shown in fig. S1.

Stimuli and experimental design

Each subject participated in the following experiments for this study: S-T and M-T experiments. Each subject listened to four stories read by a female and male speaker (denoted by Spk1 and Spk2). Both Spk1 and Spk2 were native American English speakers and were recorded in-house. For the M-T experiment, subjects were presented with a mixture of the same female and male speakers (Spk1 and Spk2), with no spatial separation between them. The acoustic waveform of each speaker was matched to obtain the same root mean squared intensity. All stimuli were presented using a single Bose SoundLink Mini 2 speaker situated directly in front of the subject. The M-T experiment was divided into four blocks by mixing different stories of Spk1 and Spk2. In total, there were 11 min and 37 s of audio presented to each subject during the M-T experiment. The S-T experiment lasted twice as long as each subject was required to listen to each story once read by Spk1 and once by Spk2.

Data preprocessing and hardware

Data were recorded using Tucker Davis Technologies hardware and sampled at 2441 Hz. The data were resampled to 500 Hz. A first-order Butterworth high-pass filter with a cutoff frequency at 1 Hz was used to remove DC drift. Data were subsequently re-referenced using a local scheme, whereby the average voltage from the nearest neighbors was subtracted from each electrode. Line noise at 60 Hz and its harmonics (up to 240 Hz) were removed using second-order infinite impulse response (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 SD of this prestimulus period.

Next, data were filtered into the high-gamma band (70 to 150 Hz); the power of this band is modulated by auditory stimuli (*5*, *44*, *45*). To obtain the power of this broad band, we first filtered the data into eight frequency bands between 70 and 150 Hz with increasing bandwidth using Chebyshev type 2 filters. Then, the power (analytic amplitude) of each band was obtained using a Hilbert transform. We took the average of all eight frequency bands as the total power of the high-gamma band.

Transformation of electrode locations onto an average brain

The electrodes were first mapped onto the brain of each subject using coregistration, followed by their identification on the postimplantation computed tomography scan using BioImage Suite. To get the anatomical location labels of these electrodes, we used the Freesurfer's automated cortical parcellation by Destrieux brain atlas (*46*). These labels were closely inspected by the neurosurgeons using the subject's coregistered postimplant magnetic resonance imaging. We plotted the electrodes on the average Freesurfer brain template.

Stimulus reconstruction

To determine the attended speaker, we used a method known as stimulus reconstruction (*22*, *47*). This method applies a spatiotemporal filter (decoder) to neural recordings to reconstruct an estimate of the spectrogram that a user is listening to. The decoder is trained by performing linear regression to find a mapping between the neural recordings and spectrogram. Training on single-speaker data was performed to minimize any potential bias that may result from training the decoders on the M-T data. Electrodes were chosen if they were significantly more responsive to speech than to silence. To perform these statistical analyses, we segmented the neural data into 500-ms chunks and divided them into the following categories: speech and silence. Significance was determined using unpaired *t* test (false discovery rate corrected, $q < 0.05$). This electrode selection resulted in varying numbers of electrodes for each subject (see fig. S1). The decoders were trained using all electrodes simultaneously and with time lags from −400 to 0 ms. See (*22*) for further information on the stimulus reconstruction algorithm.

Decoding accuracy

As previously stated, we trained the decoders using single-speaker data. These same decoders could then be used to reconstruct spectrograms from the M-T experiment (*5*). Determining to whom the subject is attending requires correlation analysis, commonly using Pearson's *r* value (*6*, *36*). Typically, the spectrogram that has the largest correlation with the reconstructed spectrogram is considered the attended speaker. We used window sizes ranging from 2 to 32 s to calculate correlations (in logarithmically increasing sizes). We defined decoding accuracy as the percentage of the segments in which the reconstructions had a larger correlation with the attended spectrogram than with the unattended spectrogram.

Dynamic switching of attention

To simulate a dynamic scenario in which a subject was switching attention between two speakers, we divided and concatenated the neural data into consecutive segments in which subjects were attending to either speaker. Specifically, we divided the data into 10 segments, each lasting 60 s. 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 2 to 32 s (in logarithmically increasing sizes). Larger windows should lead to more consistent (less noisy) correlation values, thus providing a better estimate of the attended speaker. However, this approach should also be slower at detecting a switch in attention, therefore leading to a reduction in decoding speed.

Psychoacoustic experiment

We tested the perceived quality of the modified speech by performing a psychoacoustic experiment on 20 healthy controls using Amazon Mechanical Turk (www.MTurk.com). 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 Spk1 and Spk2. However, we altered the presentation of the stimuli to obtain as much information as possible about the subjects' perception. The experiment was divided into six blocks, each containing nine trials. Each trial consisted of a single sentence. One-third of the trials consisted of the raw mixture, another third contained modified audio using the ODAN-AAD framework, and the remaining third contained modified audio using the original clean sources with the AAD framework. The trial order was randomized. Before each block, the subjects were instructed to pay attention to one of the speakers. To test the difficulty of attending to the target speaker, after each trial (sentence), we asked the subjects to indicate the difficulty they had in understanding the attended speaker on a scale of 1 to 5 as follows: very difficult (1), difficult, not difficult, easy, and very easy (5). From these responses, we calculated the MOS (*24*). In total, the experiment lasted approximately 15 min.

Initializing attractor points

The initial position of the attractor points at $\tau = 0$ in the embedding space was chosen from a set of N predetermined points, which we refer to as anchor points (*11*). During the training phase, we created N randomly initialized, trainable anchor points in the embedding space V, which are denoted by $B_j = 1, \ldots, N$. During the training of the network, the position of the anchor points was jointly optimized to maximize the separability of the mixture sounds. After the training was performed, the anchor points were fixed. To separate a mixture that contains C speakers during the test phase, we first chose all possible C combinations of the N anchor points, resulting in $\binom{N}{C}$ subsets of the N anchors. Next, we found the distance of the embedded T-F bins at $\tau = 0$ from the anchor points in each of the $\binom{N}{C}$ subsets. The C initial attractors for a particular mixture are the ones in the subset that minimize in-set similarity between the attractors (i.e., maximizing the in-set distance between the chosen attractor points).

Calculating the updated rate of attractors

The location of the attractors at each time step was updated on the basis of their previous position, the centroid of the embeddings for the current frame, and the current input frame

$$A_{\tau,i}(k) = (1 - \alpha_{\tau,i}(k))A_{\tau-1,i}(k) + \alpha_{\tau,i}(k)C_{\tau,i}(k)$$

$$C_{\tau,i}(k) = \frac{\sum_f V_\tau(f,k)Y_{\tau,i}(f)}{\sum_f Y_{\tau,i}(f)}$$

where $C_{\tau,i}(k)$ is the centroid of the embeddings of source i at time step τ, and parameter α determines the rate of the update at time τ by controlling the trade-off between the previous location of the attractors and the centroid of the sources in the current frame. If α is too small, the attractor changes position too quickly from one frame to the next, which may result in a noisy estimate and unstable separation. If α is too large, the attractor will be too slow to track the changes in the mixture condition, which could be problematic if the speakers in the mixture change over time. To optimally estimate α, we calculated a dynamic weighting function to control the relative weight of previous and current estimates using a parameter, Q, for each source i at time step τ

$$Q_{\tau,i}(k) = \sigma\left(h_{\tau-1}W + X_\tau U + A_{\tau-1,i}(k)J + b\right)$$

where $\sigma(\cdot)$ is the sigmoid activation function, $h_{\tau-1}$ is the previous output of the last LSTM layer, X_τ is the current mixture feature, and W, U, J, and b are parameters that are jointly learned during the training of the network. Given parameter $Q_{\tau,i}(k)$, the update parameter α is estimated using the following equation

$$\alpha_{\tau,i}(k) = \frac{\sum_f Y_{\tau,i}(f)}{Q_{\tau,i}(k)\sum_{t=0}^{\tau-1}\sum_f Y_{t,i}(f) + \sum_f Y_{\tau,i}(f)}$$

where $Q_{\tau,i}(k)$ adjusts the contribution of previous and current attractor estimates at time step τ. We found that parameter Q correctly tracks a change in the speakers in the mixture because the change creates a discrepancy between the previous output of the LSTM network and the current input, as shown in fig. S2.

ODAN network architecture

The network consisted of four unidirectional LSTM layers with 600 units in each layer. The embedding dimension was set to 20 based on the observations reported in (*11*), which resulted in a fully connected layer of 2580 hidden units (20 embedding dimensions times 129 frequency channels) after the LSTM layers. The number of anchors was set to 6 (*11*). We trained the models using curriculum training (*11*), in which we first trained the models on 100-frame-long input segments (0.8 s) and continued training thereafter on 400-frame input segments (3.2 s). The batch size was set to 128. Adam (*48*) was used as the optimizer with an initial learning rate of $1e^{-4}$, which was halved if validation error does not decrease after three epochs. The total number of epochs was set to 150, and early stopping was applied if validation error is not decreased after 10 consecutive epochs. All models were initialized using a pretrained LSTM DAN model. A gradient clip with a maximum norm of 0.5 was applied to accelerate training.

ODAN training data

The neural network models were trained by mixing speech utterances from the *Wall Street Journal* corpus (*49*). We used the WSJ0-2mix and WSJ0-3mix datasets, which contain 30 hours of training, 10 hours of validation, and 5 hours of test data. The test set contained 3000 mixtures generated by combining utterances from 16 unseen speakers from the si_dt_05 and si_et_05 subsets. All sounds were resampled to 8 kHz to simplify the models and to reduce computational costs. The input feature is the log magnitude spectrogram computed using a STFT, with 32-ms window length (256 samples) and 8-ms hop size (64 samples), and weighted by the square root of a hamming window. Wiener filter–like masks (*14*) were used as the training objective.

Evaluation metrics

We evaluated and compared the separation performance on the test set using the following metrics: SDR, SI-SNR, (*11*), and PESQ score (*20*), as well as ESTOI (*21*) for the evaluation of speech quality and intelligibility.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/5/5/eaav6134/DC1

Fig. S1. Electrode coverage and speech responsiveness for each subject.

Fig. S2. The change in the update parameter of attractors (parameter *q* in methods) when the speakers in the mixture switch.

Movie S1. The full demo of the proposed ODAN-AAD system.

REFERENCES AND NOTES

1. R. Carhart, T. W. Tillman, Interaction of competing speech signals with hearing losses. *Arch. Otolaryngol.* **91**, 273–279 (1970).
2. V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, U. Rass, Signal processing in high-end hearing aids: State of the art, challenges, and future trends. *EURASIP J. Appl. Signal Process.*, 2915–2929 (2005).
3. J. Chen, Y. Wang, S. E. Yoho, D. Wang, E. W. Healy, Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises. *J. Acoust. Soc. Am.* **139**, 2604–2612 (2016).
4. R. Plomp, Noise, amplification, and compression: Considerations of three main issues in hearing aid design. *Ear Hear.* **15**, 2–12 (1994).
5. N. Mesgarani, E. F. Chang, Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* **485**, 233–236 (2012).
6. J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, E. C. Lalor, Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* **25**, 1697–1706 (2015).
7. B. D. Van Veen, K. M. Buckley, Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Mag.* **5**, 4–24 (1988).
8. S. Van Eyndhoven, T. Francart, A. Bertrand, EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses. *I.E.E.E. Trans. Biomed. Eng.* **64**, 1045–1056 (2017).
9. A. Aroudi, D. Marquardt, S. Doclo, EEG-based auditory attention decoding using steerable binaural superdirective beamformer, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2018), pp. 851–855.
10. Z. Chen, Y. Luo, N. Mesgarani, Deep attractor network for single-microphone speaker separation, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2017), pp. 246–250.
11. Y. Luo, Z. Chen, N. Mesgarani, Speaker-independent speech separation with deep attractor network. *IEEE/ACM Trans. Audio Speech Lang. Process.* **26**, 787–796 (2018).
12. J. O'Sullivan, Z. Chen, J. Herrero, G. M. McKhann, S. A. Sheth, A. D. Mehta, N. Mesgarani, Neural decoding of attentional selection in multi-speaker environments without access to clean sources. *J. Neural Eng.* **14**, 056001 (2017).
13. D. Wang, J. Chen, Supervised speech separation based on deep learning: An overview. *IEEE/ACM Trans. Audio Speech Lang. Process.* **26**, 1702–1726 (2018).
14. J. R. Hershey, Z. Chen, J. Le Roux, S. Watanabe, Deep clustering: Discriminative embeddings for segmentation and separation, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2016), pp. 31–35.
15. M. Kolbæk, D. Yu, Z.-H. Tan, J. Jensen, M. Kolbaek, D. Yu, Z.-H. Tan, J. Jensen, Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ ACM Trans. Audio Speech Lang. Process.* **25**, 1901–1913 (2017).
16. K. W. Grant, S. Greenberg, Speech intelligibility derived from asynchronous processing of auditory-visual information, in AVSP 2001-International Conference on Auditory-Visual Speech Processing (AVSP, 2001), Scheelsminde, Denmark, pp. 132–137.
17. G. Strang, *Introduction to Linear Algebra* (Wellesley-Cambridge Press Wellesley, 1993).
18. T. K. Moon, The expectation-maximization algorithm. *IEEE. Sig. Process. Mag.* **13**, 47–60 (1996).
19. I. Jolliffe, Principal component analysis, in *International Encyclopedia Statistical Science* (Springer, 2011), pp. 1094–1096.
20. A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (IEEE, 2001), Salt Lake City, UT, 7 to 11 May, pp. 749–752.
21. J. Jensen, C. H. Taal, An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**, 2009–2022 (2016).
22. N. Mesgarani, S. V. David, J. B. Fritz, S. A. Shamma, Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *J. Neurophysiol.* **102**, 3329–3339 (2009).
23. D. S. Brungart, Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Am.* **109**, 1101–1109 (2001).
24. MOS, Vocabulary for performance and quality of service (ITU-T Recs 10, 2006).
25. S. Ray, J. H. R. Maunsell, Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLOS Biol.* **9**, e1000610 (2011).
26. P. W. Hullett, L. S. Hamilton, N. Mesgarani, C. E. Schreiner, E. F. Chang, Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *J. Neurosci.* **36**, 2014–2026 (2016).
27. A. Hassan, B. Khalighinejad, J. L. Herrero, A. D. Mehta, N. Mesgarani, Reconstructing intelligible speech from the human auditory cortex. *BioRxiv*, 350124 (2018).
28. M. Kolbæk, D. Yu, Z.-H. Tan, J. Jensen, Joint separation and denoising of noisy multi-talker speech using recurrent neural networks and permutation invariant training, in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)* (IEEE, 2017), pp. 1–6.
29. W. Chan, N. Jaitly, Q. Le, O. Vinyals, Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, in *2016 IEEE International Conference Acoustics Speech and Signal Processing (ICASSP)* (IEEE, 2016), pp. 4960–4964.
30. H. Sak, A. W. Senior, F. Beaufays, Long short-term memory recurrent neural network architectures for large scale acoustic modeling, in *15th Annual Conference of the International Speech Communication Association* (Interspeech, 2014), pp. 338–342.
31. Z. Chen, J. Li, X. Xiao, T. Yoshioka, H. Wang, Z. Wang, Y. Gong, Cracking the cocktail party problem by multi-beam deep attractor network, in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (IEEE, 2017), pp. 437–444.
32. J. Heymann, L. Drude, R. Haeb-Umbach, Neural network based spectral mask estimation for acoustic beamforming, in *2016 IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2016), pp. 196–200.

33. K. Ovtcharov, O. Ruwase, J.-Y. Kim, J. Fowers, K. Strauss, E. S. Chung, Accelerating deep convolutional neural networks using specialized hardware. *Microsoft Res.Whitepaper.* **2**, 1–4 (2015).
34. R. Andri, L. Cavigelli, D. Rossi, L. Benini, YodaNN: An Ultra-Low Power Convolutional Neural Network Accelerator Based on Binary Weights, in *2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)* (IEEE, 2016), pp. 236–241.
35. G. Lacey, G.W. Taylor, S. Areibi, Deep learning on fpgas: Past, present, and future. arXiv:1602.04283 (2016).
36. B. Mirkovic, S. Debener, M. Jaeger, M. De Vos, Decoding the attended speech stream with multi-channel EEG: Implications for online, daily-life applications. *J. Neural Eng.* **12**, 046007 (2015).
37. L. Fiedler, M. Wöstmann, C. Graversen, A. Brandmeyer, T. Lunner, J. Obleser, Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. *J. Neural Eng.* **14**, 036020 (2017).
38. Y. B. Benovitski, A. Lai, C. C. McGowan, O. Burns, V. Maxim, D. A. X. Nayagam, R. Millard, G. D. Rathbone, M. A. le Chevoir, R. A. Williams, D. B. Grayden, C. N. May, M. Murphy, W. J. D'Souza, M. J. Cook, C. E. Williams, Ring and peg electrodes for minimally-Invasive and long-term sub-scalp EEG recordings. *Epilepsy Res.* **135**, 29–37 (2017).
39. S. A. Fuglsang, T. Dau, J. Hjortkjær, Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *Neuroimage* **156**, 435–444 (2017).
40. A. de Cheveigné, D. D. E. Wong, G. M. Di Liberto, J. Hjortkjær, M. Slaney, E. Lalor, Decoding the auditory brain with canonical component analysis. *Neuroimage* **172**, 206–216 (2018).
41. S. Akram, A. Presacco, J. Z. Simon, S. A. Shamma, B. Babadi, Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. *Neuroimage* **124**, 906–917 (2016).
42. S. Miran, S. Akram, A. Sheikhattar, J. Z. Simon, T. Zhang, B. Babadi, Real-time tracking of selective auditory attention from M/EEG: A bayesian filtering approach. *Front. Neurosci.* **12**, 262 (2018).
43. P. Patel, L. K. Long, J. L. Herrero, A. D. Mehta, N. Mesgarani, Joint representation of spatial and phonetic features in the human core auditory cortex. *Cell Rep.* **24**, 2051–2062 (2018).
44. E. M. Z. Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann, R. R. Goodman, R. Emerson, A. D. Mehta, J. Z. Simon, D. Poeppel, C. E. Schroeder, Mechanisms underlying selective neuronal tracking of attended speech at a 'cocktail party'. *Neuron* **77**, 980–991 (2013).
45. N. E. Crone, D. Boatman, B. Gordon, L. Hao, Induced electrocorticographic gamma activity during auditory perception. *Clin. Neurophysiol.* **112**, 565–582 (2001).
46. C. Destrieux, B. Fischl, A. Dale, E. Halgren, Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* **53**, 1–15 (2010).
47. H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, N. Mesgarani, Towards reconstructing intelligible speech from the human auditory cortex. *Sci. Rep.* **9**, 874 (2019).
48. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv:1412.6980 (2014).
49. D. B. Paul, J. M. Baker, The design for the Wall Street Journal-based CSR corpus, *in Proceedings of the Workshop on Speech and Natural Language* (Association for Computational Linguistics, 1992), pp. 357–362.
50. Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, J.R. Hershey, Single-channel multi-speaker separation using deep clustering, in *Proceedings of Interspeech*, pp. 545–549.

Acknowledgments

Funding: This work was funded by a grant from the National Institutes of Health (NIDCD-DC014279), National Institute of Mental Health (R21MH114166), Columbia Technology Ventures, and the Pew Charitable Trusts, Pew Biomedical Scholars Program. **Author contributions:** J.O. and N.M. designed the experiments and analyzed the neural data. J.O., J.H., N.M., and A.D.M. recorded the neural data. C.H., Y.L., and N.M. developed the speech separation algorithm. N.M., J.O., C.H., and Y.L. wrote the manuscript. All authors commented on the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the author.

Submitted 6 October 2018
Accepted 9 April 2019
Published 15 May 2019
10.1126/sciadv.aav6134

Citation: C. Han, J. O'Sullivan, Y. Luo, J. Herrero, A. D. Mehta, N. Mesgarani, Speaker-independent auditory attention decoding without access to clean speech sources. *Sci. Adv.* **5**, eaav6134 (2019).

Science Advances

Speaker-independent auditory attention decoding without access to clean speech sources

Cong Han, James O'Sullivan, Yi Luo, Jose Herrero, Ashesh D. Mehta and Nima Mesgarani

Sci Adv **5** (5), eaav6134.
DOI: 10.1126/sciadv.aav6134