Brain-Controlled Augmented Hearing for Spatially Moving Conversations in Multi-Talker Environments

*Vishal Choudhari, Cong Han, Stephan Bickel, Ashesh D. Mehta, Catherine Schevon, Guy M. McKhann, and Nima Mesgarani**

Focusing on a specific conversation amidst multiple interfering talkers is challenging, especially for those with hearing loss. Brain-controlled assistive hearing devices aim to alleviate this problem by enhancing the attended speech based on the listener's neural signals using auditory attention decoding (AAD). Departing from conventional AAD studies that relied on oversimplified scenarios with stationary talkers, a realistic AAD task that involves multiple talkers taking turns as they continuously move in space in background noise is presented. Invasive electroencephalography (iEEG) data are collected from three neurosurgical patients as they focused on one of the two moving conversations. An enhanced brain-controlled assistive hearing system that combines AAD and a binaural speaker-independent speech separation model is presented. The separation model unmixes talkers while preserving their spatial location and provides talker trajectories to the neural decoder to improve AAD accuracy. Subjective and objective evaluations show that the proposed system enhances speech intelligibility and facilitates conversation tracking while maintaining spatial cues and voice quality in challenging acoustic environments. This research demonstrates the potential of this approach in real-world scenarios and marks a significant step toward developing assistive hearing technologies that adapt to the intricate dynamics of everyday auditory experiences.

Takeaways

- Developed a brain-controlled hearing algorithm for dynamic multitalker settings with moving conversations, closely mimicking real-world listening environments
- Introduced a binaural speech separation model that isolates speech from moving talkers while preserving spatial cues, enhancing auditory perception and attention decoding
- Proposed system improves speech intelligibility and reduces listening effort in realistic acoustic scenes

1. Introduction

Speech communication in multi-talker environments is challenging, particularly for the hearing impaired.[1] Modern hearing aids, though proficient at suppressing general background noises,[2,3] fall short in a critical aspect: they cannot selectively enhance the attended talker's speech without first knowing which talker is the target.[4] This limitation underscores the need for a brain-controlled approach, in which the listener's neural responses are used to decode and enhance the talker to whom attention is directed,[5] a technique known as auditory attention decoding (AAD).[6] In parallel, the field of automatic speech separation has seen significant progress in the recent years.[7–9] Speech separation aims to isolate individual talkers from a mixture (captured by one or more microphones). Auditory attention decoding can be combined with automatic speech separation to enable a brain-controlled hearing device[6,10–13] by isolating and amplifying the speech of the attended talker,

V. Choudhari, C. Han, N. Mesgarani
Department of Electrical Engineering
Columbia University
New York, NY 10027, USA
E-mail: nima@ee.columbia.edu

V. Choudhari, C. Han, N. Mesgarani
Mortimer B. Zuckerman Mind Brain Behavior Institute
New York, NY 10027, USA

S. Bickel, A. D. Mehta
Hofstra Northwell School of Medicine
Uniondale, NY 11549, USA

S. Bickel, A. D. Mehta
The Feinstein Institutes for Medical Research
Manhasset, NY 11030, USA

C. Schevon
Department of Neurology
Columbia University
New York, NY 10027, USA

G. M. McKhann
Department of Neurological Surgery, Vagelos College of Physicians and Surgeons
Columbia University, New York
New York, NY 10027, USA

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/advs.202401379

DOI: 10.1002/advs.202401379

therefore mitigating the challenges posed by multi-talker environments.

Past studies have established the feasibility of decoding auditory attention from both invasive[5,6,10,11] and non-invasive[13–16] neural recordings. Despite these advancements, existing studies predominantly employ overly simplistic acoustic scenes that do not mimic the real world scenarios.[6,10–14,17] These experimental setups have been limited to stationary talkers without background noise, and primarily focus on distinguishing between two concurrent talkers. Many commonly used datasets[16,17] use such a simplistic setting. This lack of realism in experimental design is a significant barrier to the generalization of these technologies to everyday life scenarios. Real-world listening involves dynamic conversation involving multiple talkers, often engaged in turn-taking while moving in space, all amidst varying background noises. As these elements have not been incorporated in prior AAD research, our study aims to bridge this gap by simulating a more realistic experimental paradigm and proposing a framework to deal with such challenging listening scenes, therefore advancing the field of AAD toward practical applications.

Another important factor that past research has often overlooked is the listeners' desire to track moving talkers in space. This aspect is crucial for natural listening[18] and, thus, for the effectiveness of brain-controlled hearing devices. A successful brain-controlled hearing device must separate speech streams as they move in space while preserving the perceived spatial location of each talker.[19] Previous studies of AAD have been based on decoding only the spectro-temporal features of speech.[16,17,19] However, recent scientific studies have shown that the human auditory cortex also encodes the location of the attended talker[20–22] which can potentially lead to the ability to decode the spatial trajectory of attended talkers. Our study takes a crucial step by investigating whether adding talker trajectories can improve the AAD performance.

Another persistent challenge in AAD model fitting and evaluation is the difficulty in accurately determining the attentional focus of the listener, especially with high temporal resolution. Previous methods often assume that subjects correctly followed the task instructions and focused on the to-be-attended (cued) talker throughout the experiment, overlooking the possibility of inadvertent attention shifts.[23] This assumption can lead to mislabeling in data and biasing the performance evaluation of AAD algorithms. In contrast to previous research, our study addresses this issue by integrating a behavior measure into our experimental design to ascertain the ongoing focus of the subject more precisely, thereby enhancing the reliability of our data and the validity of our evaluation metrics.

In this work, we present a comprehensive and novel approach to AAD that uses complex, dynamic stimuli that more closely resemble real-world acoustic environments. Specifically, we use two concurrent conversations that feature moving talkers and natural background noise, alongside speaker turn-taking among attended and unattended conversations. Furthermore, we introduce a novel task for determining the ground truth labels in attention-focused conversation by requiring the subject to detect deliberately placed repeated words (1-back task).[24,25] Lastly, we propose a refined brain-controlled hearing system equipped with a real-time, speaker-independent binaural speech separation model[6,10] that preserves the spatial location of the talkers and outputs real-time speaker trajectories that are used to increase the decoding accuracy of the attended talker. We demonstrate that the system improves speech intelligibility and conversation tracking and preserves the spatial characteristics and voice quality essential for realistic and immersive auditory experiences. This represents a significant advancement toward brain-controlled hearing devices in real-world listening environments.

2. Results

2.1. Subjects and Neural Data

Neural responses from three patients undergoing epilepsy treatment were collected as they performed the task with intracranial electroencephalography (iEEG). Two patients (Subjects 1 and 2) had stereo-electroencephalography (sEEG) depth as well as subdural electrocorticography (ECoG) grid electrodes implanted over the left hemispheres of their brains. The other patient (Subject 3) only had sEEG depth electrodes implanted over their left-brain hemisphere. All subjects had electrode coverage over their left temporal lobe, spanning the auditory cortex. The neural data was processed to extract the envelope of the high gamma band (70–150 Hz), which was used for the rest of the analysis. Speech-responsive electrodes were determined using t-tests on neural data samples collected during speech v/s silence (see Experimental Section). S1, S2 and S3 had 17, 34 and 42 speech-responsive electrodes respectively as shown in Figure S1 (Supporting Information).

2.2. Experiment Design

The experiment had a total of 28 multi-talker trials, with the average duration of 44.2 s (standard deviation = 2.0 s) each. As shown in **Figure 1a**, the trials consisted of two concurrent and independent conversations that were spatially separated and continuously moving in the frontal half of the horizontal plane of the subject (azimuthal range of −90 to +90 degrees). The distances of these conversations from the subject (i.e., their loudness) were equal and constant throughout the experiment. Both conversations were of equal power (RMS). Talkers were all native American English speakers. Diotic background noise[26,27] (either "pedestrian" or "speech babble") was also mixed along with the conversations at power either 9 or 12 dB below the power of a conversation stream.

Different talkers took turns in these conversations. As shown in Figure 1b, in the to-be-attended conversation, a talker switch took place at ≈50% trial time mark whereas for the to-be-unattended conversation, two talker switches took place, one at ≈25% trial time mark and the other nearly at the 75% trial time mark. Talkers in each conversation occasionally repeated words which were deliberately inserted in both the conversation streams (1-back detection task). The selection of the repeated words was done strategically to ensure that the repeats across the two conversations were non-overlapping in time (see Figure 1, Methods and video demo in supporting information). A repeated word was simulated by simply replicating the talker's voice waveform associated with the word. The average time interval between the onsets of two repeated words in a conversation was 7.0 s

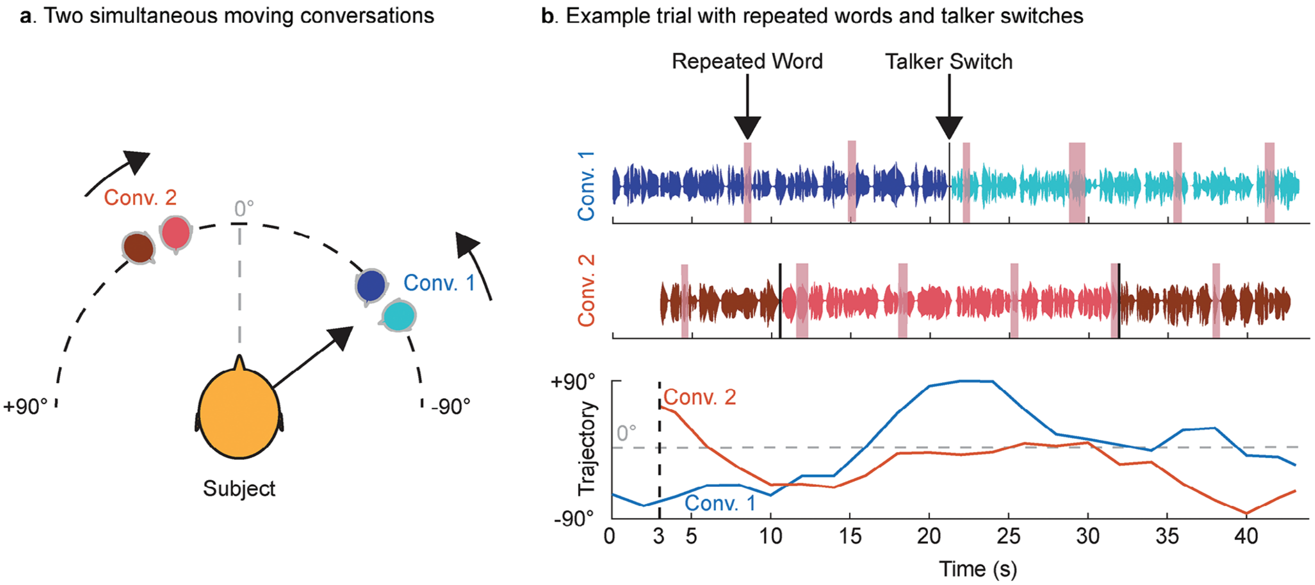

Figure 1. Experiment design. a) Every trial consisted of two concurrent conversations moving independently in the front hemifield of the subject. Each conversation had two distinct talkers taking turns. b) Talkers in each conversation repeated a word at random intervals (1-back detection task), as highlighted in pink. The cued (to-be-attended) conversation had a talker switch at ≈50% trial time mark whereas the uncued (to-be-unattended) conversation had two talker switches, at ≈25% and 75% trial time marks.

(standard deviation = 1.0 s). The assignment of male and female talkers to various segments of the conversations was counterbalanced across trials to ensure equal durations of concurrent conversations with the same and different genders.

The subjects were instructed to follow (attend to) the conversation that started first and press a push button upon hearing a repeated word in the followed conversation. The uncued (to-be-unattended) conversation started 3 seconds after the onset of the cued (to-be-attended) conversation. The trials were spatialized using head-related transfer functions (HRTFs) and delivered to the subjects via earphones.

2.3. System Proposal

Brain-controlled hearing devices need to combine a speech separation model along with auditory attention decoding to determine and enhance the attended talker. Performing AAD requires having access to individual speech streams and trajectories of every talker in the acoustic scene with which neural representations can be compared to determine the attended talker. Our proposed framework for a binaural brain-controlled hearing device assumes that there are two single-channel microphones, one on the left ear and the other on the right, as shown in **Figure** 2a. These microphones capture the left and right components of the sounds arriving at the ears of the wearer. The system framework, shown in Figure 2b, makes use of a deep learning-based speaker-independent binaural speech separation model that separates a binaural mixture of speech streams of two moving talkers (recorded by the binaural microphones) into their individual speech streams while also preserving their spatial cues. As spatial cues are preserved in the separated speech streams of the talkers, the model is also able to estimate the trajectories of the moving talkers in the acoustic scene. Auditory attention decoding is enabled by performing canonical correlation analysis (CCA) which uses the wearer's neural data and the talkers' separated speech and estimated trajectory streams to determine and enhance the attended talker by suppressing the unattended talker. With CCA, the neural responses and the attended talker's speech and trajectory streams are both linearly transformed to maximize information correlation.[28]

2.4. Speaker-Independent Binaural Speech Separation

In this section, we present an automatic speaker-independent speech separation model that separates moving sound sources and preserves spatial cues for all directional sources, enabling listeners to accurately locate each of the moving sources in space. The proposed model, as shown in Figure 2c comprises of two main modules, namely the binaural separation module and the binaural post-enhancement module. Both modules adopt TasNet which has demonstrated exceptional performance in separating audio sources.[8,29,30] Furthermore, the causal configuration enables low-latency processing, making it well-suited for real-time applications.

The binaural separation module takes binaural mixed signals as input and simultaneously separates speech for both left and right channels. Specifically, two linear encoders transform the two channels of mixed signals y^L, $y^R \in \mathbb{R}^T$ into 2-D representations E^L, $E^R \in \mathbb{R}^{N\times H}$, respectively, where T represents the waveform length, N represents the number of encoder bases, and H represents the number of time frames. To explicitly exploit spatial information, we concatenated the encoder outputs and inter-channel phase differences (IPDs) and inter-channel level differences (ILDs) between y^L and y^R, forming spectro-temporal and

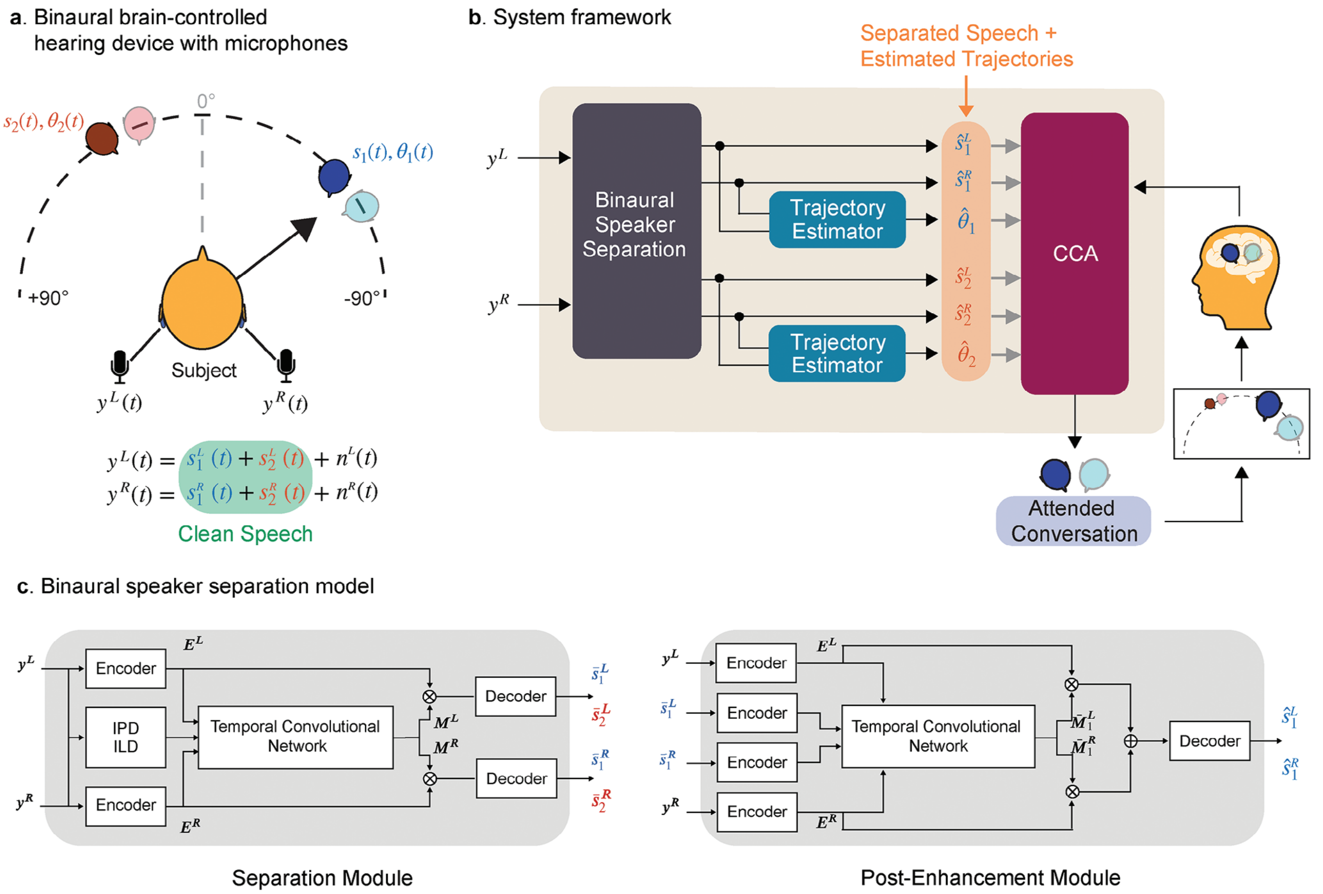

Figure 2. Proposed framework for a binaural brain-controlled hearing device. a) The framework requires two microphones, one each on both the left and the right ear. The microphones separately capture the left and the right mixtures of sound sources arriving at the ears. b) The speaker separation works with these microphone recordings to binaurally separate the speech streams while also estimating the trajectories of the talkers. These outputs are used in combination with the wearer's neural data to decode and enhance the attended talker. c) The binaural speaker separation model consists of an initial separation module whose outputs are further improved by a post-enhancement module.

spatial-temporal features (as detailed in the Experimental Section). We then pass this feature through a series of temporal convolutional network (TCN) blocks to estimate multiplicative masks, M^L, M^R $\in \mathbb{R}^{C\times N\times H}$, where C is the number of speakers. In this study, C is prefixed to be two. These multiplicative masks are then applied to E^L and E^R, respectively, and a linear decoder transforms the masked representations back to the waveform of individual speaker, $\{s_i^L, s_i^R\}_{i=1}^{C}$. The speaker-independent binaural speech separation module was trained using permutation invariant training.[31] Additionally, we imposed the constraint that the speaker order is the same for both channels, allowing the left- and right-channel signals of each individual speaker to be paired directly. The average signal-to-noise ratio (SNR) improvement of the separated speech over the raw mixture was 14.05 ± 4.79 dB.

The binaural post-enhancement module aims to enhance performance in noisy and reverberant environments because post processing stages have shown effectiveness in improving the signal quality.[32] The module takes each pair of the separated stereo sounds (e.g., s_i^L and s_i^R) and the mixed signals (y^L and y^R) as input. Similarly, all the encoder outputs are passed through the TCN blocks to estimate multiplicative masks for separating sources. Unlike the speech separation module that only applies multiplicative masks, which is equivalent to spectral filtering, the speech enhancement module performs both multiplication and summation, equivalent to both spectral and spatial filtering. This is similar to multichannel Wiener filtering.[33] Because the input stereo sound (s_i^L and s_i^R) contains both spectral and spatial information of the speaker i, the enhancement module essentially performs informed speaker extraction without the need for permutation invariant training. The average SNR improvement of the enhanced speech over the raw mixture was 16.77 ± 4.92 dB.

The key ingredient of the training is using the signal-to-noise ratio (SNR) as the objective function for both the speech separation and enhancement modules. $SNR\,(x, \hat{x}) = 10 log_{10}\left(\frac{\|x\|_2^2}{\|x-\hat{x}\|_2^2}\right)$, where x and x are the ground truth and estimated waveform, respectively. Because SNR is sensitive to both time shift and power scale of the estimated waveform, it can force the interaural level difference and interaural time difference to be preserved in the estimated waveform. Moreover, we performed utterance-level training on a moving speaker dataset, which encourages the model to leverage spectral and spatial features of speakers in a large context and forces the model to track speakers within the utterance without the need for explicit tracking modules.[19] This approach enables the model to handle moving sources effectively.

We also trained a speaker localizer module using a similar architecture to the enhancement module. The module performs classification of the direction of arrival (DOA) every 80 milliseconds. So, the localizer module estimates a moving trajectory for each moving source which can be utilized to improve the accuracy of attentional decoding. The average DOA error of the estimated trajectories was 4.20 ± 5.76 degrees (the chance level is 60 degrees).

2.5. Behavioral Data Analysis

The push button responses of subjects to repeated words in the conversation being followed help in determining to which conversation a subject was attending. A repeated word in a conversation was considered as correctly detected only if a button press was captured within two seconds of its onset. As shown in Figure S2 (Supporting Information), all subjects tracked more than 65% of the repeated words in the cued (to-be-attended) conversation. We assign these as hits. However, we see that subjects also tracked a non-zero fraction of repeated words in the uncued (to-be-unattended) conversation (false alarms) indicating that there might have been occasions when the subjects were attending to the uncued (to-be-unattended) conversation. We combined the hit rate and false alarm rate for each subject to generate a sensitivity index (d') inspired by signal detection theory[24,34] (SDT). Sensitivity index for each subject was calculated as: d' = z(False Alarm Rate) – z(Hit Rate), where z(x) is the z-score corresponding to the right-tail p-value of x.[34] Subjects were ranked based on their sensitivity indices (S1: 2.8, S2: 2.3, S3: 1.9).

2.6. Auditory Attention Decoding

In order to decode the attended talker, neural signals were compared with speech spectrograms and trajectories of talkers using canonical correlation analysis[28] (CCA) (see Experimental Section). For certain trials where it was evident that the subject was following the uncued conversation (greater than or equal to two repeated words detected in the uncued stream), the "attend" and "unattend" labels were swapped for the conversations in that portion. Subject-wise CCA models were trained, and their performance was evaluated using leave-one-trial-out cross validation, i.e., training on N – 1 trials and testing on the windows from the Nth trial. During training, the CCA models simultaneously learn forward filters on attended talker's clean speech spectrogram and trajectory and backward filters on the neural data such that upon projection with these filters, the neural data and the attended talker stimuli would be maximally correlated. During testing, these learnt filters are applied to the neural data as well as to every talker's speech spectrogram and trajectory. The talker which yields the highest correlation score (based on voting of the top three canonical correlations) was determined as the attended talker. We chose a receptive field of 500 ms for neural data and 200 ms for stimuli spectrograms and trajectories (see Experimental Section). These receptive field durations were chosen to maximize the correlation between continuous speech stimuli and its evoked neural response, as previous studies found that neural responses to speech can occur as late as 500 ms.[35–37] The starting sample of the receptive field windows were aligned in time for both neural data and stimuli.

We evaluated auditory attention decoding accuracies for all subjects for a range of window sizes from 0.5 s to 32 s for the following two stimuli versions:

1) Clean Stimuli: Using the clean (before mixing) ground truth speech spectrograms and trajectories of individual talkers in the acoustic scene.
2) Automatically Separated Stimuli: Using the speech spectrograms and estimated trajectories of talkers yielded by the binaural speech separation model.

Figure 3a shows the attended talker decoding accuracies averaged across subjects as a function of window size for both clean and separated versions after correcting for behavior. For both versions, the attended talker decoding accuracies increase as a function of window size. This is expected since with larger window sizes, more information is available to determine the attended talker. Extended window durations help overcome the combined limitations of neural noise, suboptimal brain coverage and linear decoding techniques. Stimuli version had a very small effect on the AAD accuracies across subjects and window sizes (two-sided Wilcoxon signed-rank test, z = 1.50, p-val = 0.13). This indicates that the AAD performance with automatically separated stimuli is as good as the performance with original clean stimuli (Figure 3a), confirming the efficacy of the proposed speaker-independent speech separation module.

We studied the improvement in AAD performance when talker trajectories are included in addition to talker spectrograms. For this comparison, we trained and tested CCA models (post behavior correction) with only talker spectrograms without trajectories. As shown in Figure 3b, we found that trial-wise AAD performance improved when talker trajectories were also incorporated in addition to talker spectrograms for both clean (two-sided paired t-test, t = 3.2235, df = 80, p-val = 0.002, 95% CI: 0.7534 to 3.1845) and automatically separated (two-sided paired t-test, t = 2.6316, df = 80, p-val = 0.010, 95% CI: 0.3470 to 2.4995) versions of the stimuli. The mean improvement observed was 1.97% for the clean version (from 87.61% to 89.58%) and 1.41% for the automatically separated version (from 86.86% to 88.27%).

Lack of having a behavioral measure and not correcting for the same can lead to underreporting of AAD performance. To study this, we also trained a set of CCA models assuming that the subjects always paid attention to the cued (to-be-attended) conversation. **Figure** 4a compares the AAD performance for clean stimuli when correcting and not correcting for behavior. Not correcting for behavior significantly hurts AAD performance (two-sided Wilcoxon signed-rank test, signed-rank = 0, p-val < 0.001). This is also true when evaluating with the automatically separated version of the stimuli (two-sided Wilcoxon signed-rank test, signed-rank = 0, p-val < 0.001). The mean improvement observed across all window sizes was 2.26% for the clean version (from 84.28% to 86.54%) and 2.82% for the automatically separated version (from 83.17% to 85.98%).

Next, for the models trained without correcting for behavior, we examined whether the behavioral performance on the

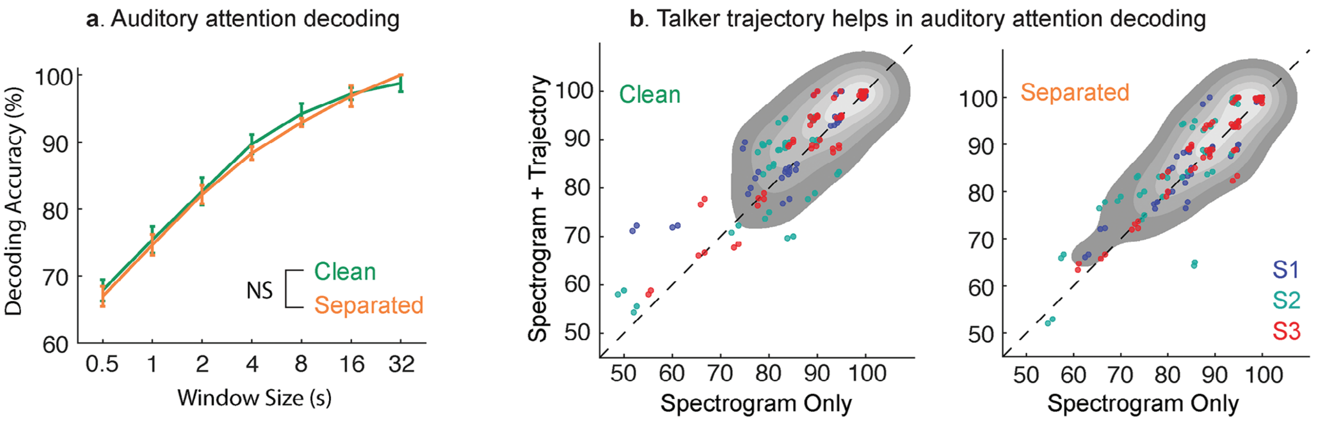

Figure 3. Evaluating auditory attention decoding (AAD) performance and the contribution of talker trajectory. a) AAD accuracies averaged across subjects as a function of window size. The decoding accuracies are comparable between the clean and separated versions (two-sided Wilcoxon signed-rank test, z = 1.50, p-val = 0.13). Error bars indicate the standard error of mean. b) Scatter plots comparing trial-wise AAD accuracies for a window size of 4 s when using only spectrogram versus spectrogram + trajectory. Each point represents a trial. AAD accuracies improved significantly when talker trajectories were also incorporated in addition to their speech spectrograms for both clean (two-sided paired t-test, t = 3.2235, df = 80, p-val = 0.002, 95% CI: 0.7534 to 3.1845) and separated (two-sided paired t-test, t = 2.6316, df = 80, p-val = 0.010, 95% CI: 0.3470 to 2.4995) versions.

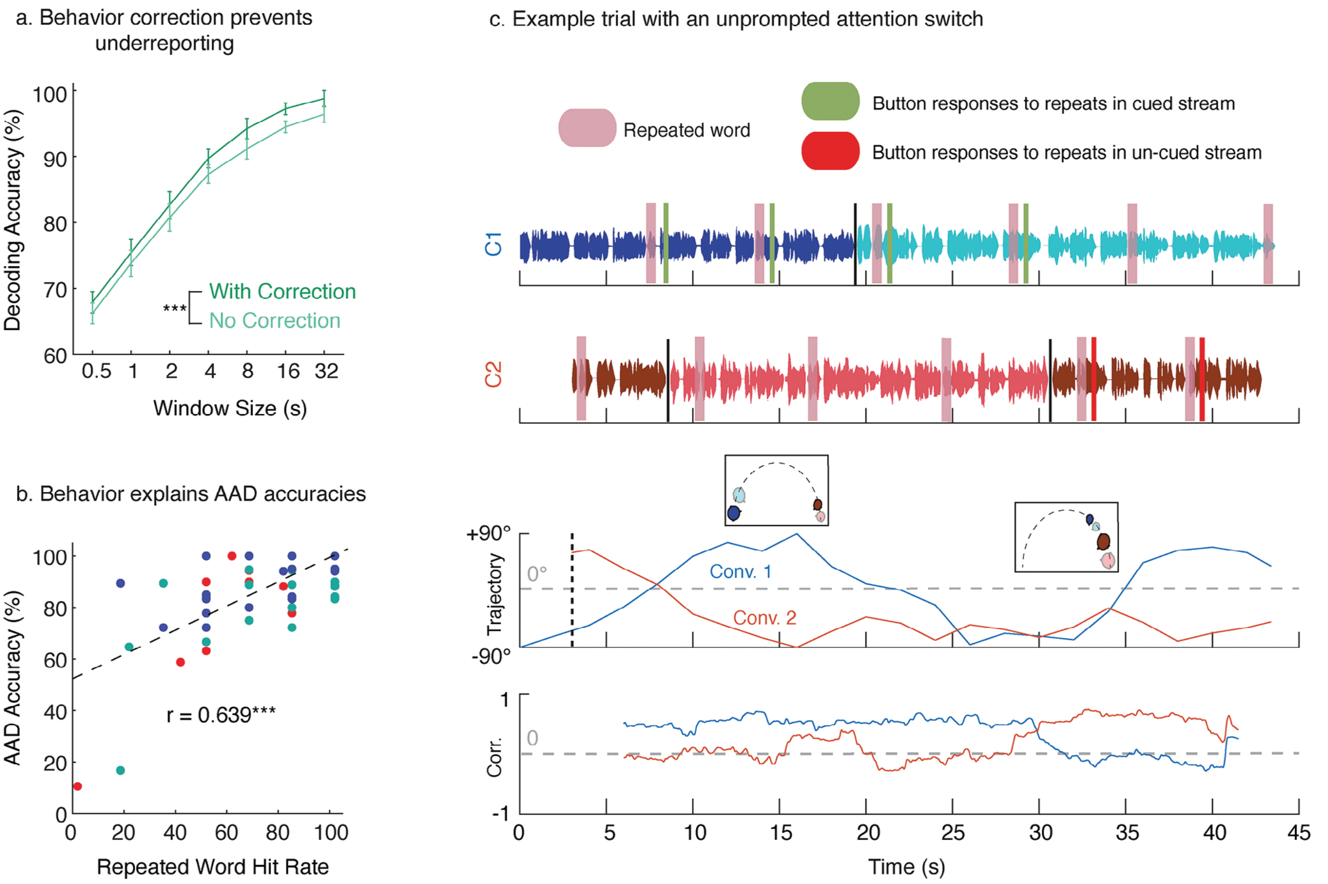

Figure 4. Correcting for behavior prevents underreporting AAD performance. a) A separate set of models were trained without correcting for behavior. The decoding accuracies are plotted for the clean version of speech for both with and without behavior correction. Not correcting for behavior can lead to significant underreporting of AAD performance (two-sided Wilcoxon signed-rank test, signed-rank = 0, p-val < 0.001) b) For models trained without correcting for behavior, trial-wise behavioral performance and AAD accuracies are significantly correlated (Pearson's r = 0.639, p-val < 0.001). c) An example trial from one of the subjects who shifts attention from the cued conversation (Conv. 1) to the uncued conversation (Conv. 2) in the middle of the trial. Repeated words in the conversation streams are shaded in pink. Button press responses to the repeated words are shown in green (red) for the cued (uncued) conversation. The last plot shows the first canonical correlation for both the conversation streams obtained by continuously sliding a 4 s window. Behavior is well correlated with the canonical correlations.

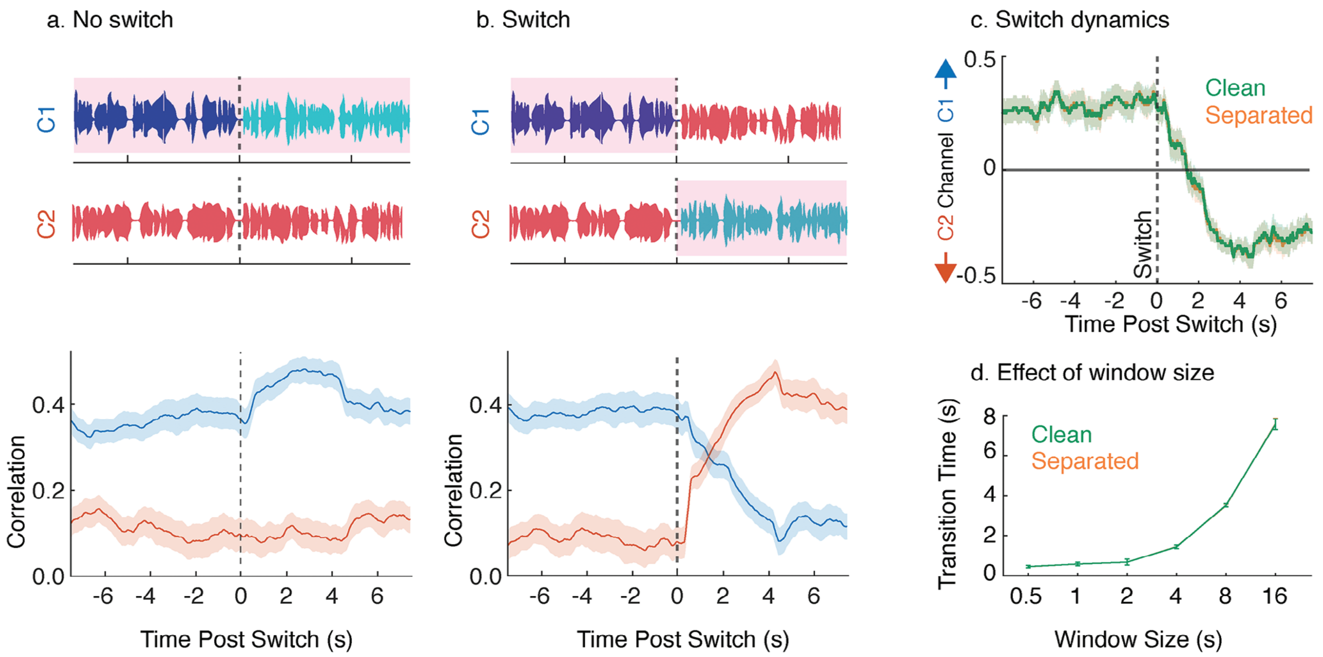

Figure 5. The proposed system seamlessly tracks turn-takings. This is facilitated by the speaker separation module which places talkers in a conversation on the same output channel by relying on location and talker continuity cues. a) Attended conversation is highlighted with a pink shade. Correlations shown are the average of the top three canonical correlations for separated version of the stimuli. b) Attention switch from one conversation to another can be simulated by swapping the output channels of the binaural separation system. c) Channel preference dynamics after simulated attention switch for a decoding window size of 4 s. d) Transition times as a function of decoding window size. No significant differences were observed between the clean and separated versions (two-sided Wilcoxon signed-rank test, signed-rank = 17, p-val = 0.70). Error bars in all plots indicate the standard error of mean.

repeated word detection task could explain the AAD performance on a trial-by-trial basis. We first computed the proportion of repeated words detected in the cued conversation (hit rate) for each trial and for each subject. We also computed corresponding trial-wise AAD accuracies for a window size of 4 s. As shown in Figure 4b, we found that hit rate on the repeated word detection task was significantly correlated with the trial-wise AAD accuracies (Pearson's r = 0.639, p-val < 0.001). Figure 4c shows an example trial from one of the subjects who, based on behavioral responses, was initially attending to the cued (to-be-attended) conversation and then later attends to the uncued (to-be-unattended) conversation after the conversations cross in space. The canonical correlations mapping the neural data with both the cued and uncued stimuli also capture this shift of attention from one conversation to the other. Thus, the repeated word detection task helps explain AAD performance on a trial-by-trial basis.

2.7. System Dynamics During Talker Transitions

Turn-takings during conversations create talker switches in the attended conversation. For good user experience, it is important that the system tracks the talker switch and seamlessly enhances the new attended talker. Our experiment paradigm, inspired by real-world settings, had asynchronous talker switches in both to-be-attended and to-be-unattended conversations. The new talker continued at the same location as the previous talker in the conversation. The speaker separation model was able to put talkers of a conversation on the same output channel using location and talker continuity. As a result, the system was able to seamlessly track turn-takings in conversations, as shown in **Figure 5a**.

In some cases, the wearer of the hearing device might switch attention from a conversation at a particular location to another conversation at a different location. To study how our system responds in such cases, we artificially swapped the outputs of the binaural speech separation system at the point of talker switch in the cued conversation, as shown in Figure 5b. Since we combine the results of the top three canonical correlations based on voting to determine the attended talker or channel, we define a metric, channel preference index (CPI), i.e.,

$$\text{CPI} = \frac{\text{\# of votes favoring Channel 1}}{3} - 0.5 \tag{1}$$

Thus, a positive CPI would indicate a preference to Channel 1 whereas a negative CPI would indicate a preference to Channel 2. In Figure 5c, we show the CPI averaged across trials for one of the subjects (S3) when attention switch is simulated. We define the transition time as the time point where the average CPI crosses 0. Figure 5d shows the transition times (averaged across subjects) as a function of window size for both clean and separated versions. No significant difference was found in the transition times across subjects and window sizes between the clean and separated versions (two-sided Wilcoxon signed-rank test, signed-rank = 17, p-val = 0.70).

2.8. Evaluation of System Performance

2.8.1. Part A: Subjective

To evaluate the performance of the proposed system, an online Amazon MTurk experiment was conducted with 24 native speakers of American English with self-reported normal hearing. The participants listened to simulated output of the proposed system using the neural signatures obtained by concatenating the channels from all the three subjects, for a total of 15 trials, five for each of the following conditions, in a blind fashion:

1) System Off: The raw mixture stimuli that was played to the subjects from whom neural data was recorded.
2) System On (Separated): Mixture in which the attended talker, as determined by the neural signatures, was enhanced using the output of the binaural speaker-separation model.
3) System On (Clean): Mixture in which the attended talker was enhanced using clean ground truth speech.

Enhanced mixtures were generated by suppressing the unattended talker and the background noise in the mixture by the same scale factor such that the resulting power difference between the attended and the unattended talker was 9 dB (see Experimental Section). Like the iEEG participants, the MTurk participants were also instructed to follow the cued conversation and press space bar on their keyboards upon hearing the repeated words in the conversation being followed. After each trial, the participants were asked to rate the difficulty of following the cued conversation on a scale from 1 (very difficult) to 5 (very easy) and the quality of voices in the conversation on a scale from 1 (bad) to 5 (excellent). To also test the intelligibility of the conversations, the participants were also asked to respond to a multiple-choice question based on the content of the cued conversation after each trial. A short localization task was also included at the end of each trial to determine if the attended talker can be localized post enhancement. In the localization task, participants indicated the perceived location of the attended talker, choosing between five spatial regions of left, front left, center, front right, and right.

Figure 6 summarizes the results. As shown in Figure 6a, under both the “system on” conditions, the repeated word detection accuracy in the cued conversation is enhanced when compared to the “system off” condition (two-sided paired t-test, p-val < 0.001), whereas for the uncued conversation (Figure 6b), the detection accuracy is reduced (two-sided paired t-test, p-val < 0.01). This means that the system helps track the cued conversation and prevents unintentional tracking of the uncued conversation. We also find that intelligibility of the cued conversation is significantly enhanced under the “system on” conditions (Figure 6c, two-sided paired t-test, p-val < 0.05). No significant differences are observed between the clean and separated versions of the “system on” condition. Ease of attending to the cued conversation increases from “system off” condition to “system on with separated speech” condition (two-sided paired t-test, p-val < 0.0001) to “system on with clean speech” condition (two-sided paired t-test, p-val < 0.01), as shown in Figure 6d. Surprisingly, no differences in voice quality of the talkers in the cued conversation were observed between the “system off” and the “system on with separated speech” condition (Figure 6f). However, participants rated the voice quality in the “system off with clean speech” condition higher than the other two (two-sided paired t-test, p-val < 0.05). These results indicate that a scope for improvement exists for the speaker-independent binaural speech separation model and its upper bounds (when there is ideal separation) are captured by the “system on with clean speech” condition. The ability to localize talkers in space, as shown in Figure 6e, was comparable across all the three conditions highlighting retention of the attended talker spatial cues when the system is turned on. In summary, the system helps follow the conversation of interest, increases its intelligibility and the ease of attending to it while also preserving spatial cues.

2.8.2. Part B: Objective

In addition to subjective evaluation, we also performed an objective evaluation where the same system simulated outputs in the subjective evaluation were compared with their corresponding clean to-be-attended conversation waveforms (as reference) to calculate narrowband MOS-mapped Perceptual Evaluation of Speech Quality[38] (PESQ) and Extended Short-Time Objective Intelligibility[39] (ESTOI) scores. As expected, in Figure 6g,h, we see a significant improvement in these scores as we progress from “system off” condition to “system on with separated speech” condition to “system on with clean speech” condition (two-sided paired t-tests, p-val < 0.0001).

3. Discussion

We introduced a novel AAD experimental paradigm that diverges from existing studies by incorporating concurrent conversations with natural turn-takings where talkers move in space amidst background noise. This approach represents a substantial advancement in creating realistic auditory scenarios for AAD research. Our binaural speaker separation system successfully separated these dynamic conversations into individual streams while preserving talker spatial cues. Additionally, the speech separation system provides real-time talker trajectories to the AAD algorithm, enhancing its decoding accuracy. The use of the repeated word detection task across the conversations provided a robust ground truth label for the attended conversation with a high temporal resolution and explained AAD performance on a trial-by-trial basis. Evaluations of the proposed system revealed improved tracking of the attended conversation and increased intelligibility while preserving the perceived location of each talker in space.

The primary aim of this study was to address the limitations of previous AAD research that predominantly assumed two stationary talkers,[6,10,11,13,14] thereby restricting the applicability of such research to real-world scenarios. In realistic acoustic scenes, we normally listen to simultaneous conversations which can involve multiple talkers. Our research extends previous work by replacing concurrent talkers with concurrent conversations involving natural turn-taking. By introducing a speaker-independent speech separation model that leverages both spatial and spectro-temporal information, our research marks a significant step toward creating an immersive listening experience that closely mimics natural environments. This model not only separates the speech of moving talkers but also allows listeners to accurately

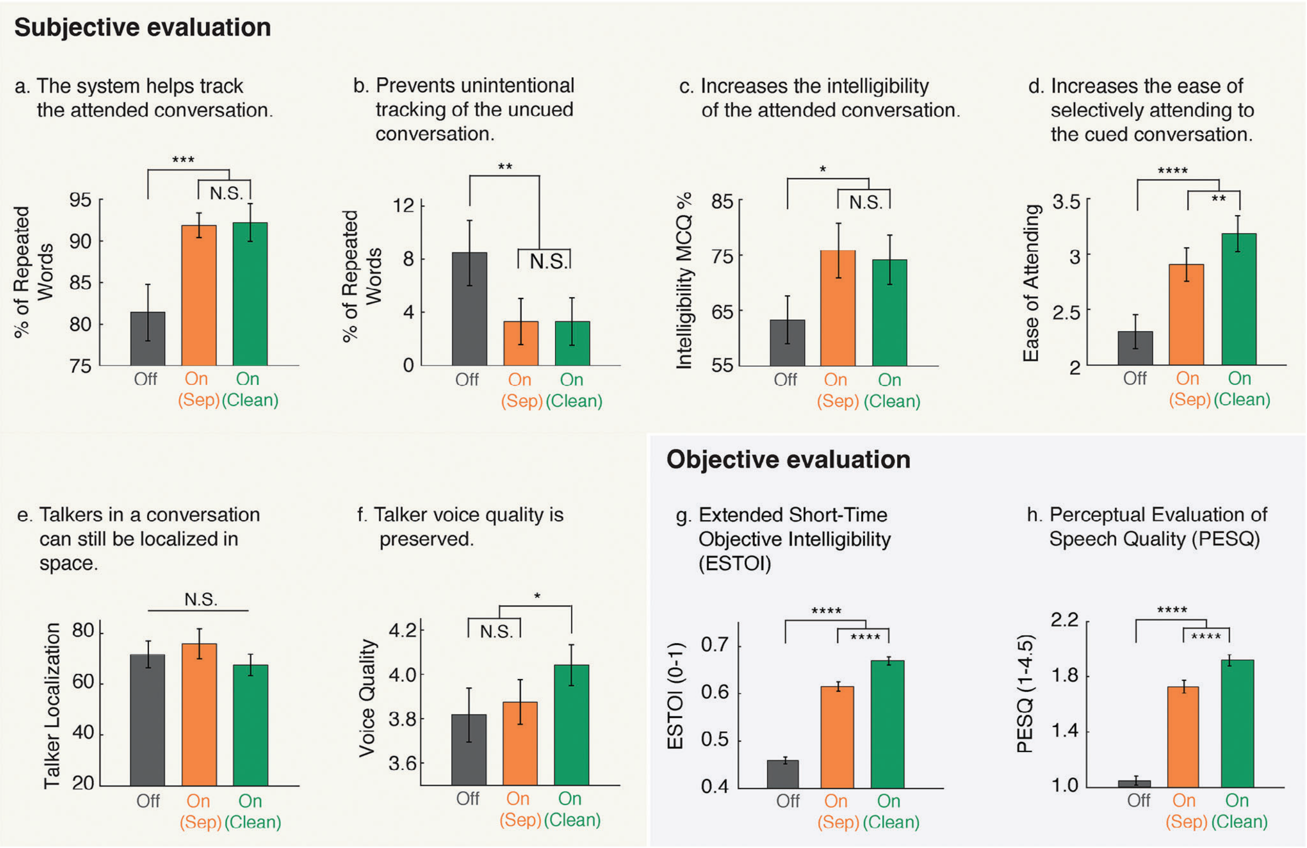

Figure 6. Subjective and objective evaluation of system outputs shows enhanced tracking of the cued conversation, improved intelligibility and retention of talker spatial cues and voice quality. For subjective tests, twenty-four online participants listened to trials from the following conditions: i) System Off: raw original mixture played to iEEG subjects, ii) System On (Separated): attended talker enhanced with the output of the binaural separation system, iii) System On (Clean): attended talker enhanced with clean ground truth speech. a) Repeated word detection accuracy in the cued conversation increases significantly when the system is turned on for both clean as well as separated versions (two-sided paired t-test, p-val < 0.001). b) Repeated word detection accuracy for the uncued conversation drops significantly when the system is turned on (two-sided paired t-test, p-val < 0.01). c) Intelligibility of the cued conversation is significantly increased under the system on conditions (two-sided paired t-test, p-val < 0.05) d) Attending to the cued conversation is easier under the system on conditions (two-sided paired t-test, p-val < 0.0001). e) Participants can localize talkers in space equally well in all conditions (chance = 20%, see Experimental Section). f) No significant difference in voice quality ratings was observed between the system off condition versus the system on with separated speech condition. However, participants rated the voice quality of the system on with clean speech condition to be relatively higher (two-sided paired t-test, p-val < 0.05). g and h) Objective evaluation also shows improved quality and intelligibility. Both PESQ and ESTOI scores increase from "system off" condition to "system on with separated speech" condition to "system on with clean speech" condition (two-sided paired t-tests, p-val < 0.0001). Error bars in all plots indicate the standard error of mean.

track their locations, an aspect crucial for realistic AAD applications. An essential contribution of our study is that incorporating real-time talker trajectories estimated by the speech separation algorithm in addition to spectro-temporal information can improve AAD accuracy.[20,40–42] Further research is needed to distinguish listener motion-induced from talker motion-induced acoustic change and how it could be encoded differently in the human auditory cortex.[43,44]

Another contribution of our study is introducing a behavioral task of repeated word detection across conversations, allowing us to identify the actual attended conversation with high temporal resolution. This method addresses a common issue in previous AAD studies where subjects' attention could inadvertently shift to the unattended stream,[23] leading to mislabeled data and affecting the training and evaluation of AAD models. By incorporating a behavioral measure into our experiment design, we have enhanced the accuracy of determining the attended talker or conversation. In future AAD studies with moving talkers, a higher degree of temporal resolution can be achieved by asking the subjects also to report the spatial trajectory of the conversation followed. Additionally, further research is needed to investigate the difference between endogenous and exogenous auditory attention switches and how they may be decoded differently.[45]

While our study focused on neural activity in the high gamma band, incorporating low-frequency neural activity, which has been shown to track motion and attention, could improve AAD accuracies. Prior invasive[46] and non-invasive[13,47] AAD studies have shown signatures of auditory attention (via tracking of the envelope of the attended speech) in the lower frequencies (1–7 Hz). A recent study[22] also showed that low-frequency neural activity also tracks the location of the attended talker, especially in delta (<2 Hz) phase and alpha (8–12 Hz) power.

Including low-frequency neural signals might provide a more comprehensive understanding of the neural underpinning of auditory attention and enhance the performance of AAD systems.

A critical aspect of future research should involve transitioning to a real-time, closed-loop system. This requires the integration of speech separation and AAD components to work synchronously in a causal, real-time manner. Furthermore, determining how to optimally manipulate the acoustic scene based on the decoded attended talker remains an area for further investigation. Such acoustic modifications should help the listener follow the attended conversation while still maintaining the ability to switch to the unattended one. Our experiment design could be further aligned with real-world scenarios by introducing more complex motion patterns for talkers, such as radial motion and motion pauses. This would add a layer of complexity to the auditory scene, presenting conversations with time-varying power and potentially challenging the current speaker separation model. Addressing this challenge may involve retraining or fine-tuning the model on datasets with these characteristics.

A brain-controlled hearing device that can quickly and accurately adapt to changes in the listener's attention is a challenge that may be more effectively addressed with invasive neural recording techniques. However, a critique of our approach is the reliance on invasive neural recordings which might be perceived as less accessible. Considering the rapid advancements in speech BCI research involving invasive neural recordings,[48–52] these methods are becoming increasingly common and feasible. The precision and speed offered by invasive recordings are currently unmatched by non-invasive techniques, making them essential for exploring the upper limits of AAD performance. While future research continues to explore less invasive or alternative neural recording methods, our current focus on invasive recordings is crucial for advancing the field and setting benchmarks for performance of these systems and establishing minimum required performance for listeners to prefer AAD functionality.

Our study contributes significantly to AAD research and brain-controlled hearing devices by introducing more realistic experimental paradigms and advancing the technology toward practical applications. The insights from this research enhance our understanding of auditory attention in complex environments and pave the way for future innovations in assistive hearing technologies.

4. Experimental Section

Participants: The study had a total of three human participants of which two (Subjects 1 and 2) were from North Shore University Hospital (NSUH) and one (Subject 3) was from Columbia University Irving Medical Center (CUIMC). All participants were undergoing clinical treatment for epilepsy. Subjects 1 and 2 were both implanted with subdural electrocorticography (ECoG) grid and stereo-electroencephalography (sEEG) depth electrodes on their left-brain hemispheres. Subject 3 only had sEEG depth electrodes implanted over their left-brain hemisphere. The electrode targets for these participants were determined purely based on clinical requirements. The participants provided informed consent as per the local Institutional Review Board (IRB) regulations. IRB protocol number: AAAD5482 (M00Y18)

Neural Data Pre-Processing + Hardware: The neural data of participants from NSUH (Subjects 1 and 2) were recorded using Tucker-Davis Technologies (TDT) hardware using a sampling rate of 1526 Hz. The neural data of the participant from CUIMC (Subject 3) was recorded using Natus Quantum hardware using a sampling rate of 1024 Hz. Left and right channels of the audio stimuli played to the participants were also recorded in sync with neural signals to facilitate segmenting of neural data into trials for further offline analysis.

Neural data was pre-processed and analyzed using MATLAB software (MathWorks). All neural data was first resampled to 1000 Hz and then montaged to a common average reference to reduce recording noise.[53] The neural data was then further downsampled to 400 Hz. Line noise at 60 Hz and its harmonics (up to 180 Hz) were removed using a notch filter. The notch filter was designed using MATLAB's *fir2* function and applied using *filtfilt* with an order of 1000. In order to extract the envelope of the high gamma band (70 – 150 Hz), the neural data was first filtered with a bank of eight filters, each with a width of 10 Hz, spaced consecutively between 70 and 150 Hz.[54] The envelopes of the outputs of these filters were obtained by computing the absolute value of their Hilbert transform. The final envelope of the high gamma band was obtained by computing the mean of the individual envelopes yielded by the eight filters and further downsampling to 100 Hz.

Speech responsive electrodes were determined by comparing neural samples (of the high gamma envelope sampled at 100 Hz) recorded in response to speech with those recorded in response to silence. For each trial, 25 samples corresponding to silence were randomly chosen in a [$t = -0.4$ to -0.1 s] window, with $t = 0$ s being the onset of speech. Similarly, 25 samples corresponding to speech were drawn from a window [$t = 0.1$ to 1.6 s]. These samples were accumulated across trials and a t-statistic (between speech and silence samples) was computed for every electrode. Electrodes with a t-statistic above 5 were considered to be speech-responsive and retained for further analysis.

Stimuli Design and Experiment: The experiment consisted of 28 multi-talker trials with a mean trial duration of 44.2 s (SD = 2.0 s). The total experiment lasted 26 min. Every trial consisted of two concurrent and independent conversations (one to-be-attended, one to-be-ignored) that were spatially separated and continuously moving in the presence of diotic background noise. The to-be-ignored conversation started 3 s later than the to-be-attended conversation. The participants were cued to attend to the conversation that started first.

A total of eight native American English voice actors (four male, four female) were recruited to voice these conversations. These conversations were based on general daily life situations (see Table S1, Supporting Information for conversation transcripts). Every trial consisted of four talkers: two for the to-be-attended conversation (say A and B), two for the to-be-unattended conversation (say C and D). The to-be-attended conversation had one turn-taking (talker switch) at the 50% trial time mark whereas the to-be-ignored conversation had two turn-takings: one at the 25% trial time mark and the other at the 75% trial time mark. Thus, the talker in the to-be-attended conversation would transition from A to B and the talker in the to-be-ignored conversation would transition from C to D to back to C.

To check to which conversation a participant might be attending, repeated words were artificially inserted in both the to-be-attended and the to-be-ignored conversations. Participants were asked to press a button upon hearing a repeated word in the conversation that they were following. The conversation transcripts were force aligned with the audio recordings of the voice actors using the Montreal Forced Aligner tool.[55] The repeated words were inserted in the conversations based on the following criteria:

- The number of repeated words to be inserted in a conversation of a trial was determined by dividing the trial duration (in seconds) by 7 and rounding the result.
- For every trial, an equal number of repeated words were inserted in the to-be-attended and the to-be-ignored conversations.
- A word could be repeated only if its duration was at least 300 ms.
- To make repeated words sound smooth and natural, a Hanning window of 30 ms was applied to both sides of the audio segment corresponding to the repeated word.
- The audio segment corresponding to a repeated word was also prefixed and postfixed with 200 ms of silence.
- The time interval between the onsets of two repeated words in a conversation was constrained to lie between 5.5 to 9.5 s.

- There was always one repeated word whose onset was within 1.5 s post talker switch in the to-be-attended conversation. This was done to check if participants tracked the switch in talkers in the to-be-attended conversation.
- The onset of the first repeated word in a trial was constrained to lie between 5 – 8 s from trial start time. This first repeated word could occur either in the to-be-attended conversation or the to-be-ignored conversation.
- The minimum time gap between a repeated word onset in the to-be-attended conversation and a repeated word onset in the to-be-ignored conversation was set to be at least 2.5 s. This was done to prevent simultaneous overlap of repeated words in the two conversations and to allow for determining to which conversation a participant was attending to.

Google Resonance Audio software development kit (SDK) was used to spatialize the audio streams of the conversations.[56] The trajectories for these conversations were designed based on the following criteria:

- The trajectories were confined to the frontal half of the horizontal plane of the subject in a semi-circular fashion. In other words, the conversations were made to move on a semi-circular path at a fixed distance from the subject spanning −90 degrees (right) to +90 degrees (left).
- The trajectories were initially generated with a resolution of 1 degree and a sampling rate of 0.5 Hz using a first order Markov chain.
- This Markov chain had 181 states (−90 degrees to +90 degrees with a resolution of 1 degree). All states were equally probably of being the initial state.
- The subsequent samples of a trajectory were generated with a probability transition matrix shown in Figure S3 (Supporting Information).
- The resulting trajectories were smoothed with a moving average of five samples and then stretched to span the whole frontal half plane.
- The trajectories were further upsampled using linear interpolation to 10 Hz.
- A pair of trajectories corresponding to a pair of conversations in a trial also followed the following criteria:
  - The spatial separation between the conversations when the second conversation starts was set to be at least 90 degrees.
  - The spatial separation between the conversations during the talker switch in the to-be-attended conversation was ensured to be at least 45 degrees.
  - The correlation of the two trajectories were ensured to be less than 0.5.
- A total of 1000 trajectory sets (each with 28 pairs, one for each of the 28 trials) were generated based on the above criteria.
- To have the trajectories span a uniform joint distribution, the set with the highest joint entropy (computed with a bin size of 20 degrees) was chosen as final.

In addition to the two conversation streams, a single channel background noise was duplicated for both left and right channels introduced in the auditory scene. For every trial, the background noise was either pedestrian noise[26] or speech babble noise.[27] When mixing the three streams the power of the two conversation streams were always kept the same. The power of the background noise stream was suppressed relative to the power of a conversation stream by either 9 or 12 dB. Trial parameters such as background noise type, its power level and voice actor assignments were all counterbalanced across the trials. Stimuli was delivered to the participants with a sampling rate of 44.1 kHz through stereo earphones (Panasonic RP-HJE120).

Speaker-Independent Binaural Speech Separation—Cross-Domain Features: Although the encoder outputs E^L and E^R contain both spectral and spatial information, we added interaural phase difference (IPD) and interaural level difference (ILD) as additional features to increase speaker distinction when speakers are at different locations.[57,58] Specifically, we calculated $\cos(\text{IPD})$, $\sin(\text{IPD})$ and $\text{ILD} \in \mathbb{R}^{F\times H}$

$$\cos(\text{IPD}) = \cos\left(\angle Y^L - \angle Y^R\right) \tag{2}$$

$$\sin(\text{IPD}) = \sin\left(\angle Y^L - \angle Y^R\right) \tag{3}$$

$$\text{ILD} = 10\log_{10}\left(\left|Y^L\right| \oslash \left|Y^R\right|\right) \tag{4}$$

where $Y^L, Y^R \in \mathbb{R}^{F\times H}$ are the STFT output of y^L, y^R, respectively, F is the number of frequency bins, and $\oslash$ is element-wise division operation. The hop size for calculating Y^L and Y^R is the same as that for E^L and E^R to ensure they have the same number of time frames, even though the window length in the encoder is typically much shorter than that in the STFT. Finally, we concatenated these cross-domain features into$[E^L, E^R, \cos(\text{IPD}), \sin(\text{IPD}), \text{ILD}] \in \mathbb{R}^{(2N+3F)\times H}$ as the input to the binaural speech separation module.

Speaker-Independent Binaural Speech Separation—Training and Development Datasets: For the training and development sets, 24000 and 2400 9.6-second binaural audio mixtures were generated, respectively. Each mixture comprised of two moving speakers and one diotic background noise. The moving speech stimuli were created using the methods described in the **Stimuli Design and Experiment** section. Speech was randomly sampled from the Librispeech dataset.[59] For half of the training data, pairs of trajectories that spanned uniform distribution (quantified by joint entropy) were chosen; and for another half of the training data, pairs of trajectories whose average azimuthal difference was smaller than 15 degrees were chosen to enhance the separation model's ability to handle closely spaced moving speakers. Noise was randomly chosen from DEMAND dataset.[60] The SNR, defined as the ratio of the speech mixture in the left channel to the noise, ranged from −2.5 to 15 dB. All sounds were resampled to 16 kHz. The model was speaker-independent as the speakers involved in the testing phase (the voice actors) were not part of the training dataset (Librispeech), ensuring the generalizability and applicability of this system across diverse speakers.

Speaker-Independent Binaural Speech Separation—Network Architecture and Training: The binaural separation, post-enhancement, and localizer modules were all designed with a causal configuration of TasNet. For the linear encoder and decoder, we used 96 filters with a 4 ms filter length (equivalent to 64 samples at 16 kHz) and 2 ms hop size. Five repeated stacks witch each having seven 1-D convolutional blocks in the TCN module were used, resulting in an effective receptive field of ≈2.5 s. When calculating cos(IPD), sin(IPD), and ILD, the STFT window size was set to 32 ms and the window shift was set to 2 ms. The binaural separation, post-enhancement, and localizer modules were trained separately. The training batch size was set to 128. Adam[61] was used as the optimizer with an initial learning rate of 1e−3, which was decayed by 0.98 for every two epochs. Each module was trained for 100 epochs.

Canonical Correlation Analysis: Canonical correlation analysis (CCA) was used to determine the attended talker. From the stimuli side, the inputs involved talker spectrograms and trajectories. A 20-bin mel spectrogram representation obtained with a window duration of 30 ms and a hop size of 10 ms was chose. Audio was downsampled to 16 kHz before mel spectrogram extraction. The mel spectrograms of left and right channels were concatenated along the bin dimension. All trajectories were upsampled to 100 Hz from 10 Hz to match the sampling rate of the neural data. Trajectories were pooled across all trials and normalized. Spectrograms were also normalized on a bin-by-bin basis. A receptive field size of 500 ms for neural data and 200 ms for stimuli spectrograms and trajectories were chosen to maximize the correlation between phonemes in continuous speech and their evoked neural responses.[35–37] The starting sample of these receptive fields were aligned in time. Time-lagged matrices were then generated individually for neural data, trajectory and spectrograms.

As done in a previous study,[28] principal component analysis (PCA) was applied individually to time-lagged versions of both spectrogram and trajectory. PCA was also applied to the time-lagged neural data matrix. The top PCA components explaining at least 95% of the variance were retained. This was done to reduce the risk of overfitting in CCA.

CCA filters were trained to project PCA-reduced versions of the attended stimuli and neural data to maximize their correlation. During inference, the trained filters were used to generate correlations for each talker. Attended talker was decided based on voting by the preferences indicated by the first three canonical correlations.

Correction for behavior: For trials in which two or more repeated words were detected in the uncued conversation, the corresponding portions (bounded by button press timings) of the cued to-be-attended and uncued to-be-unattended stimuli were swapped before model training and evaluation. For models trained without correction, no such swapping was done based on behavior.

Enhancement of Attended Conversation: In the System Off setting, the mixture stimuli y has three streams, namely, the two conversations s_1, s_2 and a background noise n.

$$y = s_1 + s_2 + n \tag{5}$$

The binaural speech separation module yields the estimates of the separated speech $\hat{s}_1, \hat{s}_2$ from the mixture y. When editing the scene to enhance the attended conversation, the mixture was attenuated and the separated speech was added as per the following equations.

$$y_{\text{new}} = ky + \alpha_1 \hat{s}_1 + \alpha_2 \hat{s}_2 \tag{6}$$

If $\hat{s}_1 \approx s_1, \hat{s}_2 \approx s_2$, then we have the following.

$$y_{\text{new}} \approx (k + \alpha_1)\, s_1 + (k + \alpha_2)\, s_2 + n \tag{7}$$

In this enhanced mixture, it was ensured that the un-attended talker was attenuated by 9 dB below the attended talker. It was also ensured that the power of the attended talker was the same as that in the original mixture. The attended conversation was determined by canonical correlation analysis (CCA) which uses the iEEG signals and the results of the binaural speech separation. If Conversation 1 was determined to be the attended conversation, the following three equations solved to determine the gain factors k, α_1,α_2 for generating the enhanced mixture.

$$20 \log_{10} \frac{k + \alpha_1}{k + \alpha_2} = 9 \text{ dB} \tag{8}$$

$$k + \alpha_1 = 1 \tag{9}$$

$$\alpha_2 = 0 \tag{10}$$

For generating the enhanced mixtures for the System On (Separated) case, the estimates of the separated speech $\hat{s}_1, \hat{s}_2$ were used. For generating enhanced mixtures for the System On (Clean) case, the original speech streams s_1,s_2 were used.

Psychoacoustic Experiment: The online psychoacoustic experiment to evaluate system performance was conducted with 24 normal hearing (self-reported) Amazon MTurk participants. These participants were native speakers of American English based in the US. The experiment lasted for a total of 30 min per participant and each participant was paid $10. All participants were required to wear stereo earphones.

During the experiment, participants listened to trials and answered questions after each trial. The task assigned to the participants during the trial was the same as that of the participants from whom neural data was recorded: to attend to/follow the cued conversation (conversation that starts first) and press spacebar upon hearing a repeated word in the conversation being followed. jsPsych[62] was used to design this web-based experiment.

Every participant listened to a total of 15 trials, 5 trials from each of the following conditions: System On (Clean), System On (Separated) and System Off. Neural data was combined from all the three subjects along the channel dimension to test the system. Since a few subjects could have been paying attention to the uncued (to-be-unattended) stream in any given trial, to prevent combining neural signatures across subjects when the subjects were attending to different streams, trials in which at least one of the subjects had mistakenly attended (tracked at least two or more repeated words) to the uncued (to-be-unattended) stream were discarded. This resulted in 18 trials. For every MTurk participant, the selection of 15 trials from these 18 trials and their corresponding condition assignment (SysOn-Clean, SysOn-Sep, Sys Off) were randomized. The order of presentation of these trials were also randomized with the constraint that the first two trials had to be those with Sys Off condition. Throughout the experiment, the participants were unaware of the conditions assigned to the trials.

After every trial, the participants were prompted with the following four questions:

1) Comprehension: A multiple-choice question based on the content in the to-be-attended conversation with a single correct answer.
2) Difficulty: Participants were asked to rate how difficult or easy it was for them to follow the cued conversation on a scale from 1 to 5 (1 = very difficult, 2 = difficult, 3 = neutral, 4 = easy, 5 = very easy).
3) Sound Localization: The last three seconds of the trial was allowed to be replayed multiple times by the participants. Participants were asked to indicate from one of five equally partitioned sectors of the frontal half plane (left, front left, center, front right, right) where the cued conversation ended.
4) Voice Quality: Participants were also asked to rate the quality of voices in the cued conversation on a scale from 1 to 5 (1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent).

Statistical Analysis: Pre-processing of data is described in their respective sections. In all the figures, error bars represent the standard error of the mean, unless specified otherwise. The details of statistical methods used such as the type of test, p-values, degrees of freedom, etc. are mentioned inline in the main text. All statistical analysis was carried out on MATLAB (MathWorks) R2023a. P-values greater than 0.05 are indicated as not significant (N.S). P-values between 1e-2 and 0.05 are indicated with *. P-values between 1e-3 and 1e-2 are indicated with **. P-values between 1e-4 and 1e-3 are indicated with ***. P-values less than 1e-4 are indicated with ****.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work was funded by grants from the National Institute of Health's National Institute on Deafness and Other Communication Disorders (NIH-NIDCD) and The Marie-Josee and Henry R. Kravis Foundation.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

V.C. and C.H. contributed equally to this work. V.C., C.H., and N.M. designed the experiment. V.C. and N.M. analyzed the neural and behavioral data. C.H. and N.M. developed the binaural speech separation model. A.D.M. and G.M.M. performed the neurosurgeries. V.C., S.B., and C.S. recorded neural data. V.C., C.H., and N.M. wrote the manuscript and made the figures. All others commented and suggested edits to the manuscript.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Keywords

auditory attention decoding, brain-computer interfaces, intracranial electroencephalography, hearing aids, speech separation, deep learning, psychophysics

Received: February 6, 2024
Revised: June 17, 2024
Published online: September 9, 2024

[1] R. Carhart, T. W. Tillman, *Arch. Otolaryngol. – Head Neck Surg.* **1970**, *91*, 273.
[2] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, U. Rass, *EURASIP J Adv. Signal Process* **2005**, *2005*, 152674.
[3] J. Chen, Y. Wang, S. E. Yoho, D. Wang, E. W. Healy, *J. Acoust. Soc. Am.* **2016**, *139*, 2604.
[4] R. Plomp, *Ear Hear* **1994**, *15*, 2.
[5] N. Mesgarani, E. F. Chang, *Nature* **2012**, *485*, 233.
[6] J. O'Sullivan, Z. Chen, J. Herrero, G. M. McKhann, S. A. Sheth, A. D. Mehta, N. Mesgarani, *J. Neural Eng.* **2017**, *14*, 056001.
[7] J. R. Hershey, Z. Chen, J. Le Roux, S. Watanabe, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Shanghai **2016**, 31.
[8] Y. Luo, N. Mesgarani, *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1256.
[9] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, S. Watanabe, TF-GridNet: Integrating Full- and Sub-Band Modeling for Speech Separation **2022**.
[10] C. Han, J. O'Sullivan, Y. Luo, J. Herrero, A. D. Mehta, N. Mesgarani, *Sci. Adv.* **2019**, *5*, eaav6134.
[11] E. Ceolini, J. Hjortkjær, D. D. E. Wong, J. O'Sullivan, V. S. Raghavan, J. Herrero, A. D. Mehta, S.-C. Liu, N. Mesgarani, *NeuroImage* **2020**, *223*, 117282.
[12] B. J. Borgström, M. S. Brandstein, G. A. Ciccarelli, T. F. Quatieri, C. J. Smalt, *Neural Netw* **2021**, *140*, 136.
[13] 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, *Cereb. Cortex* **2015**, *25*, 1697.
[14] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigne, E. Lalor, B. T. Meyer, S. Miran, T. Francart, A. Bertrand, *IEEE Signal Process Mag* **2021**, *38*, 89.
[15] S. Vandecappelle, L. Deckers, N. Das, A. H. Ansari, A. Bertrand, T. Francart, *eLife* **2021**, *10*, e56481.
[16] W. Biesmans, N. Das, T. Francart, A. Bertrand, *IEEE Trans. Neural Syst. Rehabil. Eng.* **2017**, *25*, 402.
[17] G. Ciccarelli, M. Nolan, J. Perricone, P. T. Calamia, S. Haro, J. O'Sullivan, N. Mesgarani, T. F. Quatieri, C. J. Smalt, *Sci. Rep.* **2019**, *9*, 11538.
[18] G. Kidd, T. L. Arbogast, C. R. Mason, F. J. Gallun, *J. Acoust. Soc. Am.* **2005**, *118*, 3804.
[19] C. Han, Y. Luo, N. Mesgarani, *Binaural Speech Separation of Moving Speakers With Preserved Spatial Cues. in Interspeech*, ISCA, xx **2021**, *2021*, 3505.
[20] P. Patel, K. van der Heijden, S. Bickel, J. L. Herrero, A. D. Mehta, N. Mesgarani, *Curr. Biol.* **2022**, *32*, 3971.
[21] P. Patel, L. K. Long, J. L. Herrero, A. D. Mehta, N. Mesgarani, *Cell Rep.* **2018**, *24*, 2051.
[22] A. Bednar, E. C. Lalor, *NeuroImage* **2020**, *205*, 116283.
[23] S. Makov, D. Pinto, P. Har-shai Yahav, L. M. Miller, E. Golumbic, *Cognition* **2022**, *105313.*
[24] G. Marinato, D. Baldauf, *Sci. Rep.* **2019**, *9*, 2854.
[25] W. K. Kirchner, *J Exp Psychol* **1958**, *55*, 352.
[26] J. Barker, R. Marxer, E. Vincent, S. Watanabe, in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, IEEE, Scottsdale, AZ, USA **2015**, 504.
[27] A. J. Spahr, M. F. Dorman, L. M. Litvak, S. Van Wie, R. H. Gifford, P. C. Loizou, L. M. Loiselle, T. Oakes, S. Cook, *Ear Hear* **2012**, *33*, 112.
[28] A. de Cheveigné, D. D. E. Wong, G. M. Di Liberto, J. Hjortkjær, M. Slaney, E. Lalor, *NeuroImage* **2018**, *172*, 206.
[29] I. Kavalerov, et al., in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, New Paltz, NY, USA **2019**, 175.
[30] C. Han, Y. Luo, N. Mesgarani, in *ICASSP 2019 –2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Brighton, USA **2019**, 361.
[31] M. Kolbaek, D. Yu, Z.-H. Tan, J. Jensen, *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1901.
[32] Z.-Q. Wang, P. Wang, D. Wang, *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 2001.
[33] S. Doclo, T. Klasen, T. Van den Bogaert, J. Wouters, M. Moonen, in International Workshop on Acoustic Signal Enhancement **2006**, 1.
[34] J. A. Swets, *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers. xv*, Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, **1996**, 308.
[35] N. Mesgarani, C. Cheung, K. Johnson, E. F. Chang, *Science* **2014**, *343*, 1006.
[36] B. Khalighinejad, G. Cruzatto Da Silva, N. Mesgarani, *J. Neurosci.* **2017**, *37*, 2176.
[37] L. Gwilliams, J.-R. King, A. Marantz, D. Poeppel, *Nat. Commun.* **2022**, *13*, 6606.
[38] A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, IEEE, Salt Lake City, UT, USA **2001**, *2*, 749.
[39] J. Jensen, C. H. Taal, *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 2009.
[40] L. Kong, S. W. Michalka, M. L. Rosen, S. L. Sheremata, J. D. Swisher, B. G. Shinn-Cunningham, D. C. Somers, *Cereb. Cortex* **2014**, *24*, 773.
[41] Y. Deng, I. Choi, B. Shinn-Cunningham, *NeuroImage* **2020**, *207*, 116360.
[42] K. Krumbholz, M. Schönwiesner, D. Y. von Cramon, R. Rübsamen, N. J. Shah, K. Zilles, G. R. Fink, *Cereb. Cortex* **2005**, *15*, 317.
[43] H. M. Kondo, D. Pressnitzer, I. Toshima, M. Kashino, *Proc. Natl. Acad. Sci. U S A* **2012**, *109*, 6775.
[44] D. M. Schneider, R. Mooney, *Annu. Rev. Neurosci.* **2018**, *41*, 553.
[45] S. Haro, H. M. Rao, T. F. Quatieri, C. J. Smalt, *Eur. J. Neurosci.* **2022**, *55*, 1262.
[46] 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, *Neuron* **2013**, *77*, 980.
[47] N. Ding, J. Z. Simon, *Proc. Natl. Acad. Sci. U S A* **2012**, *109*, 11854.
[48] F. R. Willett, E. M. Kunz, C. Fan, D. T. Avansino, G. H. Wilson, E. Y. Choi, F. Kamdar, M. F. Glasser, L. R. Hochberg, S. Druckmann, K. V. Shenoy, J. M. Henderson, *Nature* **2023**, *620*, 1031.
[49] D. A. Moses, S. L. Metzger, J. R. Liu, G. K. Anumanchipalli, J. G. Makin, P. F. Sun, J. Chartier, M. E. Dougherty, P. M. Liu, G. M. Abrams, A. Tu-Chan, K. Ganguly, E. F. Chang, *N. Engl. J. Med.* **2021**, *385*, 217.
[50] S. L. Metzger, J. R. Liu, D. A. Moses, M. E. Dougherty, M. P. Seaton, K. T. Littlejohn, J. Chartier, G. K. Anumanchipalli, A. Tu-Chan, K. Ganguly, E. F. Chang, *Nat. Commun.* **2022**, *13*, 6510.
[51] S. L. Metzger, K. T. Littlejohn, A. B. Silva, D. A. Moses, M. P. Seaton, R. Wang, M. E. Dougherty, J. R. Liu, P. Wu, M. A. Berger, I. Zhuravleva, A. Tu-Chan, K. Ganguly, G. K. Anumanchipalli, E. F. Chang, *Nature* **2023**, *620*, 1037.
[52] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, N. Mesgarani, *Sci. Rep.* **2019**, *9*, 874.

[53] N. E. Crone, D. Boatman, B. Gordon, L. Hao, *Clin. Neurophysiol.* **2001**, *112*, 565.

[54] E. Edwards, M. Soltani, W. Kim, S. S. Dalal, S. S. Nagarajan, M. S. Berger, R. T. Knight, *J Neurophysiol* **2009**, *102*, 377.

[55] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, M. Sonderegger, *Interspeech* **2017**, 498.

[56] M. Gorzel, A. Allen, I. J. Kelly, A. Gungormusler, Efficient Encoding and Decoding of Binaural Sound with Resonance Audio, **2019**.

[57] Z.-Q. Wang, J. Le Roux, J. R. Hershey, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Calgary, AB, Canada **2018**, 1.

[58] R. Gu, J. Wu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, D. Yu, End-to-End Multi-Channel Speech Separation **2019**.

[59] V. Panayotov, G. Chen, D. Povey, S. L. Khudanpur, *An ASR corpus based on public domain audio books. in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, South Brisbane, Queensland, Australia **2015**, 5206.

[60] J. Thiemann, N. Ito, E. Vincent, *Proc. Mtgs. Acoust.*, Montreal, Canada **2013**, 035081.

[61] D. P. Kingma, J. A. Ba, A Method for Stochastic Optimization, **2014**.

[62] J. R. De Leeuw, R. A. Gilbert, B. Luchterhandt, *J. Open Source Softw.* **2023**, *8*, 5351.