IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 June 2025, accepted 9 July 2025, date of publication 18 July 2025, date of current version 24 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3590490

RESEARCH ARTICLE

Comparison of Linear and Nonlinear Methods for Decoding Selective Attention to Speech From Ear-EEG Recordings

MIKE D. THORNTON[1], DANILO P. MANDIC[2], (Fellow, IEEE), AND TOBIAS REICHENBACH[1]
[1]Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91052 Erlangen, Germany
[2]Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K.

Corresponding author: Tobias Reichenbach (tobias.j.reichenbach@fau.de)

This work was supported by the UKRI CDT in AI for Healthcare (http://ai4health.io) under Grant P/S023283/1.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Imperial College Research Ethics Committee under Application No. 19IC5388 No A1, and performed in line with the Declaration of Helsinki.

ABSTRACT Many people with hearing loss struggle to comprehend speech in crowded auditory scenes, even when they are using hearing aids. However, the focus of a listener's selective attention to speech can be decoded from their electroencephalography (EEG) recordings, raising the prospect of smart EEG-steered hearing aids which restore speech comprehension in adverse acoustic environments. Here, we assess the feasibility of using a novel, ultra-wearable, ear-EEG device to classify the selective attention of normal-hearing listeners who participated in a two-talker competing-speakers experiment. State-of-the-art auditory attention decoding algorithms are compared, including stimulus-reconstruction algorithms based on linear regression as well as non-linear deep neural networks, and canonical correlation analysis (CCA). Meaningful markers of selective auditory attention could be extracted from the ear-EEG signals of all participants, even when those markers are derived from relatively short EEG segments of just 5 s in duration. Algorithms which relate the EEG signals to the rising edges of the speech temporal envelope are more successful than those which make use of the temporal envelope itself. The CCA algorithm achieves the highest mean attention decoding accuracy, although differences between the performances of the three algorithms are both small and not statistically significant when EEG segments of short durations are employed. In summary, our ultra-wearable ear-EEG device offers promising prospects for wearable auditory monitoring.

INDEX TERMS Auditory attention decoding, EEG, hearables.

I. INTRODUCTION

People with hearing loss often struggle to comprehend speech in noisy environments, even when they wear their hearing aids [1], [2]. Various interventions have been developed to address this problem; for example, modern hearing aids incorporate advanced noise suppression technologies, which can reduce the impact of interfering sounds such as wind and other non-speech noises [3]. In cocktail-party scenarios, where the interfering sounds are also speech, a hearing aid must first identify the target speaker, and then selectively amplify that voice alone. If the target speaker is known *a priori* (for example, it is the teacher in a classroom environment), then the voice of that speaker can be wirelessly transmitted to hearing aid users via a hearing loop [4]. In more spontaneous settings such as social scenarios, beamforming technologies are usually utilised [5]. These rely on directional microphones to identify the direction of the listener's head, and attempt to selectively amplify sounds emanating from that direction.

Amongst these interventions, only beamforming hearing aids offer users the ability to selectively and spontaneously

The associate editor coordinating the review of this manuscript and approving it for publication was Filbert Juwono.

choose which voice to enhance. In everyday environments, however, beamforming hearing aids often provide only limited benefits to their users [6], [7]. This limitation arises because individuals do not constantly orient their head towards a target speaker during naturalistic listening. Furthermore, interfering speech can easily divert the listener's attention and shift the direction of their gaze away from the intended speaker [8]. Whilst listeners can learn to adapt their behaviours in order to experience the maximum benefit from their beamforming hearing aids, even with practice they may struggle to quickly reorient their gaze when the focus of their attention changes location or is re-directed. This means that beamforming technology can actually increase the difficulty of locating off-axis sources for hearing-aid users [9], [10].

Alternative techniques for determining the focus of a listener's attention may be used in place of, or in conjunction with, beamforming algorithms. Recent research has demonstrated that the focus of a listener's selective attention to speech can be decoded from neuroimaging data [11], [12], [13]. If a discreet and wearable brain monitoring device can be incorporated into a hearing aid, then it may be possible for the user to control the hearing aid through their attentional focus alone [14], [15], [16]. Most studies identify electroencephalography (EEG) as the most suitable neuroimaging modality for real-world auditory attention decoding, due to its relative accessibility, its high temporal resolution, and the variety of wearable EEG devices which are already commercialised.

Usually, auditory attention is decoded from EEG recordings through a technique known as stimulus-reconstruction [13], [17], [18]. This involves training regression models (known as backward models) to reconstruct features of the speech stimuli from time-aligned EEG signals. The feature of choice is typically taken to be the temporal envelope of speech, since this feature is strongly represented in EEG responses to speech [19]. Moreover, the strength of the EEG response to the speech envelope is modulated by selective attention, with responses to the attended speech envelope being dominant [12]. Typically, the strength of the response is assessed by determining the accuracy with which the attended and ignored speech envelopes can be reconstructed from the EEG signals using regression models, by taking the Pearson correlation between the reconstructed envelopes and the ground-truth envelopes [13]. The speech stream which yields the greatest Pearson correlation is then taken as the estimated attended stream.

Studies which make use of the stimulus-reconstruction approach to auditory attention decoding utilise underlying regression models which fall into two broad categories: linear models (typically based on ridge-regularised least-squares regression), and non-linear models implemented as deep neural networks [13], [14], [20], [21], [22], [23]. Stimulus-reconstruction approaches based on deep neural networks typically achieve higher attention decoding accuracies than their linear counterparts, since they can account for the non-linear processing of the human auditory system. Some studies have also demonstrated that deep neural networks can generalise between participants and datasets remarkably well [24], [25]. As an alternative to the stimulus-reconstruction approach, canonical correlation analysis (CCA) can be used to relate EEG recordings to the speech envelope, and subsequently decode auditory attention [26], [27]. This algorithm applies linear transformations to *both* the EEG signals as well as the speech envelope signals in order to produce pairs of maximally-correlated components. By applying a linear classifier to the resulting correlation coefficients, a marker of selective attention to each competing speech stream can be obtained. In their review study, Geirnaert et al. [14] identified CCA as the best auditory attention decoding algorithm, outperforming all linear and non-linear stimulus-reconstruction approaches.

Most auditory attention decoding studies make use of EEG signals which were collected from multiple electrodes placed across the scalp of each participant. The EEG electrodes are typically fitted with a headcap and conductive gel in order to achieve a low and stable impedance across the skin-electrode interface. In contrast, the sensor montage employed by any prospective EEG-steered hearing aid must be much more discreet and easy to set-up. Ideally, it will consist of just a few electrodes, and the device will not rely upon the application of conductive gel, which dries out with time (thereby changing the noise characteristics of the EEG sensors) and can be difficult to apply properly for inexperienced users. Promisingly, Ciccarelli et al. [20] have demonstrated that dry-contact EEG systems for auditory attention decoding can compete with gel-based systems, and Kappel and colleagues [28] demonstrated that dry-contact in-the-ear EEG can be used for unobstrusively measuring auditory evoked potentials.

Currently, ear-centric sensor montages are popular choices for wearable EEG-monitoring devices [29], [30], [31], [32]. It is already established that such ear-centric devices can be used for long-term and discreet auditory monitoring, for example by detecting steady-state or subcortical responses to clicks and tones [33], [34], [35], [36]. In the context of auditory attention decoding, two types of ear-EEG devices have been studied: the concealed EEG (cEEG) device, which employs a C-shaped array of around-the-ear electrodes; and in-ear sensors, which sit entirely within the ear canal [15], [16], [37]. In the study by Fiedler et al. [16], the authors found that auditory attention could not be decoded with significance when only in-ear electrodes were used. This result is to be expected, since EEG recorded from electrodes with a very small spatial separation has a very low signal-to-noise ratio (SNR) [38], [39]. Instead of using in-ear electrodes exclusively, Fiedler et al. proposed to reference the in-ear electrodes to a nearby location on the scalp; the FT7 location was identified as a promising candidate. The decoding accuracies achieved in that study were similar to

those achieved using a binaural cEEG montage in the study by Holtze et al. [37].

In the study of Fiedler et al., the authors followed the usual stimulus-reconstruction approach to decode auditory attention to continuous speech in two-talkers conditions [16]. However, in place of the commonly-used temporal envelope feature of speech, the authors made use of the onset envelope feature. The onset envelope is the defined as the half-wave rectified derivative of the temporal envelope, and therefore captures the rising edges of the temporal envelope which are predominantly driven by word and syllable onsets [40].

In this work, we demonstrate how a dry-contact, wirelessly connected ear-EEG device can be used to decode the focus of a listener's auditory attention in two-talker competing-speakers scenarios. The sensor configuration is similar to that proposed by Fiedler et al. in that it consists of two in-ear electrodes, and one adjacent scalp electrode. We adopt three attention decoding algorithms: stimulus reconstruction based on linear models and deep neural networks, and CCA. We further provide a comparison of the attention decoding performance when the temporal envelope feature is used, versus when its onsets is used.

II. MATERIALS AND METHODS

A. SUBJECTS AND STIMULI

Eighteen young, normal-hearing participants (median age: 23 years; 10 males, 8 females) took part in a selective auditory attention experiment. The experimental protocol was approved by the Imperial College Research Ethics Committee (approval number 19IC5388 No A1; approved 20th September 2019).

The experiment consisted of 16 trials: during each trial, participants were presented with two concurrent audiobook chapters – one narrated by a male talker, and the other by a female talker. The audiobooks were presented diotically (i.e. without spatial separation) at a sampling frequency of 44.1 kHz via Seinheiser HD450 headphones, which were placed over the housing modules of the in-ear electrodes. Participants were instructed to attend to one talker (male or female), and to ignore the other. Each trial lasted for approximately 150 seconds.

Across the entire experiment, audiobook material was narrated by the same two talkers. Participants alternated the focus of their attention between the two talkers every four trials. Both speech streams were delivered at equal intensity, with a calibrated sound pressure level of 75 dB SPL (measured using a Brüel & Kjær type-4157 ear simulator).

B. EAR-EEG DEVICE DESCRIPTION

The ultra-wearable, wirelessly enabled ear-EEG device used in this study consisted of two in-ear EEG sensors, one external reference electrode, and a clip-on ground electrode. The reference electrode was affixed to the scalp (see next section), and the ground was clipped to the right earlobe. The in-ear sensors incorporated custom conductive doped-silicon tips, offering a tactile feel similar to that of commercial silicone earphone tips. Signal amplification electronics were housed in bilateral modules located just outside the ear canals. These modules also integrated microphones and 9-axis inertial measurement units (IMUs), which can be used to denoise electrophysiological signals as demonstrated in prior work from our group [41], [42]. An additional lightweight module housing the Bluetooth transmitter rested comfortably on the participants' shoulders.

All sensor signals were sampled at 256 Hz. While the IMU data were not utilized in this study, the microphones captured the mixed speech audio during stimulus presentation. These recordings were used offline to time-align the original audio data with the EEG signals via a cross-correlation-based method. Note that the microphone signals were not used for further analysis; instead, features were extracted directly from the original audio data, as described in the proceeding section. Additional information about the device is available upon request.[1]

C. EAR-EEG DATA ACQUISITION

Dry-contact electrodes were inserted into the ear canals, and the external reference electrode was taped to the lower left forehead (at the FT7 location in the 10–20 system). The ground electrode was clipped to the right earlobe. This configuration yielded one bilateral (cross-head) and one unilateral EEG channel.

The FT7 reference location was chosen based on findings by Fiedler et al. [16], who reported that reference electrodes placed too close to the in-ear sensors achieved very poor attention decoding accuracies. In pilot recordings, we found that placing the reference electrode on the left mastoid led to insignificant attention decoding accuracies, which was also reported by Fiedler et al. (see supplementary information for [16].)

Participants tolerated the ear-EEG setup well, with no major discomfort reported. The most common issue related to the tightness of the Sennheiser headphones, which pressed against the outer ear and occasionally caused mild discomfort near the end of the session. To improve comfort and mechanical stability, the in-ear tips were available in three sizes (small, medium, and large). Once inserted, the earpieces remained remarkably secure and did not require adjustment during the session.

As with other ear-EEG setups, the device was sensitive to movement-related artifacts such as those arising from mastication, swallowing, and electro-ocular activity. To minimize these, participants were encouraged to sit still and to limit unnecessary movements. In our experience, overly strict instructions to avoid movement can cause discomfort and may impair task performance. For this reason, we adopted a more flexible approach, allowing participants to move if needed to remain comfortable and focused throughout the task.

[1] d.mandic@imperial.ac.uk

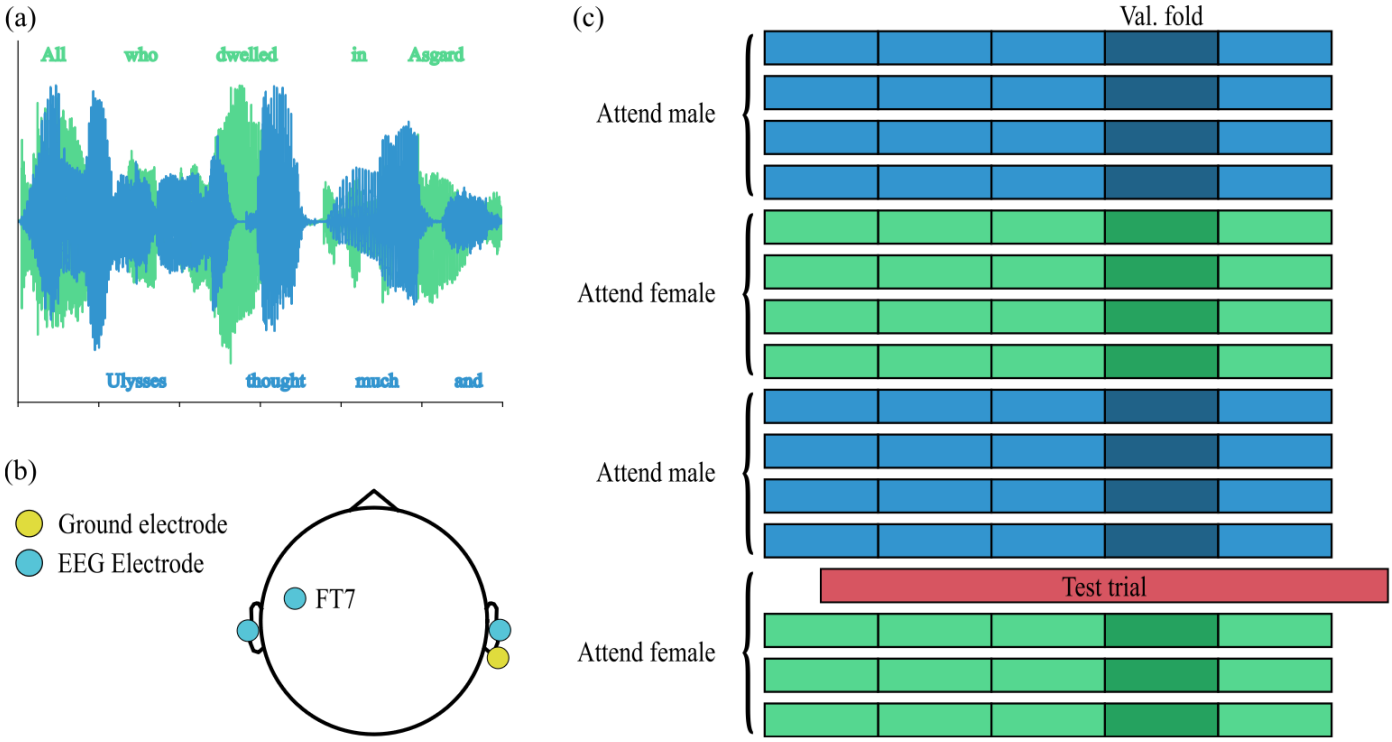

FIGURE 1. Overview of the experiment and analysis protocols. a) Participants listened to a male voice and a female voice which narrated two distinct audiobooks at the same time. They were asked to direct their attention to one of the voices, and to ignore the other. b) The dry-contact EEG sensor montage employed in this work consisted of two in-ear electrodes which were referenced to the FT7 scalp channel, and a clip-on ground electrode which was attached to the right earlobe. c) Participants listened to four short stories, each of which was split into four trials of approximately 2.5 minutes in length. The to-be attended speaker alternated with each story. A nested-cross-validation procedure was used to train and test the attention decoders: data from each trial was held-out in turn; the remaining data were split into five folds. Four of those folds were used to train attention decoders with different hyperparameters, and the remaining fold was used to select the best hyperparameter configurations.

D. SIGNAL PRE-PROCESSING

We related the EEG recordings to the temporal envelopes and the onset envelopes of each speaker separately. There are several types of temporal envelope which have been explored in the literature; we used the auditory-inspired envelope first proposed by Biesmans et al. [43]. This envelope is calculated by filtering the audio signals into 28 sub-bands via a gammatone filterbank; we used the implementation provided in the Brian 2 Hears v0.9.2 python package [44]. The centre frequencies of the filterbank are spaced equidistantly between 50 Hz and 5 kHz on an equivalent rectangular bandwidth scale. The signal in each sub-band is then half-wave rectified, and the resulting signals are averaged together to form a stimulus envelope. As in the study by Fielder et al. [16], the onset envelope feature was formed by half-wave rectifying the first derivative of the temporal envelope feature. In practice, this involved taking the difference between consecutive samples of the temporal envelope feature, and setting negative values to zero. Both features were resampled to 64 Hz after the application of an anti-aliasing boxcar filter.

The EEG signals were preprocessed by first band-passing them from 0.5 Hz to 8 Hz (Type-2 Hamming-sinc FIR filter of order 1691 with a −6 dB stopband attenuation at 0.25 Hz and at 10 Hz). The EEG and envelope signals were finally resampled to a sampling frequency of 64 Hz and standardised across the time dimension by removing the mean of each time series and dividing by its standard deviation.

E. FORWARD MODELS

Forward models, known also as temporal response functions (TRFs), are estimates of the impulse response functions which relate the EEG signals recorded from each channel to a particular feature present in the speech stimulus [45]. The features that we considered in this case were the temporal envelope of speech as well as its onsets.

In practice, TRFs can be estimated by fitting FIR filters to predict the EEG signals from the speech-feature signals, with the filter coefficients themselves representing the TRF estimates. We used 160 filter taps, with each corresponding to a response latency ranging from −1 s (pre-stimulus) to 1.5 s (post-stimulus). Following Biesmans et al. [43], the ridge regularisation parameter of each TRF was selected as the mean eigenvalue of the speech-feature autocovariance matrix. The TRFs were estimated using a cross-validation procedure: for each participant, data from one of the 16 trials was held-out for evaluation. A TRF was then estimated using the data from the remaining 15 trials. Overall, 16 TRFs were fitted per-participant using this procedure, which were averaged in order to produce a single TRF per participant. In this way, we fitted distinct TRFs for each EEG channel,

each stimulus feature (the speech envelope as well as its onsets), and each speaker (the attended speaker and the ignored speaker). In order to assess whether the TRFs were influenced by selective attention, we also formed 'difference TRFs' by subtracting the ignored-speaker TRFs from the attended-speaker TRFs.

It is important to make sure that the TRFs are meaningful before interpreting their coefficients. To this end, a single-sample permutation-based cluster test was used to test whether the instantaneous power of the TRF coefficients was statistically different to random chance. First, a set of null-TRFs was constructed by repeated the cross-validation feature described above, but with temporally misaligned EEG signals and speech-feature signals. We repeated this procedure using 500 different temporal misalignments, yielding 500 null-TRFs per participant, per held-out trial. Temporal clusters in the TRFs were formed by marking samples at which the TRFs' instantaneous power exceeded the 99^{th} percentile of the instantaneous power distribution obtained from the null-TRFs. The size of the largest cluster formed the test statistic of the permutation-based cluster analysis. A null distribution for this test statistic was formed by randomly permuting the signs of the participant's TRFs prior to averaging them; we used 1000 separate permutations to form this distribution. Although the cluster-based permutation tests only test the null hypothesis that each TRF as a whole is different to random chance, clusters with low p-values can still be visualised and interpreted: we retained and inspected all clusters with a p-value less than $p = \frac{0.05}{12}$ (i.e. we Bonferroni corrected for the 12 TRFs shown in Figure 2.)

F. LINEAR BACKWARD MODELS

Like forward models, linear backward models are FIR filters which relate the EEG recordings to features of the auditory stimuli (in this case, the envelopes and onset envelopes of the competing speech streams). However, unlike forward models, backward models predict the speech features from the EEG recordings. The same algorithm (regularised ridge-regression) was used to fit the backward models: in this case, both EEG channels were used to predict speech features, and we considered 64 filter taps corresponding to latencies between 0 s to 1 s. In other words, EEG samples which were delayed up to 1 s relative to the audio stimulus were used to predict the speech features.

We computed participant-specific linear backward models and tuned the ridge regularisation parameter using the nested cross-validation procedure outlined in Figure 1. Data from each trial was held-out in turn for the purpose of evaluating the final models. The remaining data was split into five equally-lengthed folds; each fold was held-out in turn, and the remaining four folds were used to train 19 models with regularisation parameters spaced equally on a log scale between 10^{-9} and 10^{9}. The model which achieved the highest correlation coefficient on the held-out fold was submitted for final evaluation on the held-out trial. As with the forward models, we fitted backward models for each speech feature and each competing speaker. A null distribution for the correlation coefficients was constructed by correlating the reconstructed speech features from a particular testing trial against the actual speech features taken from other trials.

G. CNN-BASED BACKWARD MODELS

The CNN-based backward models were direct analogues of the linear backward models in the sense that they used a temporal window of 2-channel EEG recordings, of 1 s in duration, to reconstruct the speech features which occurred at the onset of the temporal window. However, as universal function approximators, the CNN-based backward models implement non-linear mappings between the EEG recordings and the speech-feature signals, unlike the linear backward models [46]. This means that they may be better placed to account for the fundamentally non-linear nature of the human auditory system [22].

The CNN architecture used in this work consists of several convolutional blocks which are linked together in a feed-forward manner with skip connections. Each convolutional block consists of a convolutional layer, followed by a rectified linear unit (ReLU) activation function, a batch normalisation (BatchNorm) layer, and an average pooling layer. Convolutional layers of artificial neurons act as pattern detectors, since they implement a set of learned multi-channel matched filters; by sequentially stacking convolutional layers, CNNs can detect complex hierarchical patterns [47]. Non-linear activation functions such as the ReLU are essential elements of artificial neural networks which allow them to approximate non-linear functions [46], [48]. BatchNorm performs a learned normalisation of the outputs of the convolutional layer, and has been reported to lead to faster and more stable training of CNNs [49], [50]. Average pooling layers perform downsampling on their inputs through averaging adjacent temporal samples. Since some studies in EEG decoding have reported improved performance when skip connections are used, we utilised them here [25], [51]. After the last convolutional block, a linear combination of the transformed EEG signals is taken to produce an estimate of the envelope (or its onsets) of the attended speech stream.

Participant-specific CNNs were trained using the same nested cross-validation scheme as the linear backward models (see Figure 1). The parameters of the CNNs were tuned using the Adam optimiser with a learning rate of 0.001 [52]. Using the inner cross-validation loop, we tuned two hyperparameters: the kernel size of the convolutional layers (3 or 5) and the number of convolutional blocks (1, 2 or 3).

H. AUDITORY ATTENTION DECODING VIA BACKWARD MODELS

The following procedure was used to perform auditory attention decoding using both types of backward models. Data from the held-out trials were first split into temporal segments of lengths ranging from 0.1 s to 30 s. There was

a 1 s hop length between adjacent segments. For each segment, we obtained reconstructed speech features from the attended-speaker backward models, and these were correlated against the corresponding features derived from both the attended and ignored speech streams. The difference in correlation coefficients, $\Delta\rho = \rho_{\text{attended}} - \rho_{\text{ignored}}$, served as a marker of selective auditory attention. A null distribution for the attention markers was obtained by correlating the reconstructed speech features from each particular segment with the actual speech-features corresponding to a different, randomly selected segment. Single-tailed, unpaired t-tests were used to compare the attention markers against the null markers for each participant.

I. CANONICAL CORRELATION ANALYSIS

Canonical correlation analysis applies linear filters to both the EEG as well as the speech-feature signals, in order to yield new signals, or 'canonical components', which are maximally correlated. Because of this, CCA can be considered as the simultaneous application of forward and backward models. The CCA algorithm yields multiple pairs of canonical components, of which the first pair are the most highly correlated, the second pair are the next most highly correlated, and so on. Importantly, different pairs of canonical components are mutually uncorrelated, i.e. neither of the first pair of components are correlated with either of the second, *et cetera*.

In Geirnaert et al. [14], the authors used CCA to perform auditory attention decoding by performing the following steps: first, they trained the CCA algorithm to maximise the correlation between the EEG signals and the speech envelope of the attended speaker. Then, they passed the EEG signals and corresponding attended and ignored speech envelopes through the algorithm, obtaining two vectors of correlation coefficients ρ_{attended} and ρ_{ignored}; these correspond to the attended and ignored speech streams respectively. Finally, they trained a linear classifier (linear discriminant analysis, LDA) to distinguish between the correlation difference vectors $\rho_{\text{attended}} - \rho_{\text{ignored}}$ and $\rho_{\text{ignored}} - \rho_{\text{attended}}$, thus estimating the identity of the attended speaker. Here, we trained CCA-based attention decoders in same fashion as Geirnaert et al. using the nested cross-validation procedure described above. We trained the LDA classifiers on correlation difference vectors derived from short temporal speech/EEG segments of 5 s in duration. As with the backward models, we evaluated the CCA-based algorithms using various segment durations with a 1 s hop length between segments. The number of canonical components to consider (i.e. the dimensionality of the correlation difference vectors) was tuned via the inner cross-validation loop. Following Geirnaert et al., we did not perform any dimensionality reduction on the original signals prior to applying the CCA algorithm.

III. RESULTS

A. FORWARD MODELS

Temporal response functions were fitted to relate the EEG recordings to the speech envelopes and their onsets. The TRFs were estimated for both the attended speaker and the ignored speaker. In order to assess whether the TRFs were modulated by selective attention, we formed difference TRFs by subtracting the ignored speech TRFs from the attended speech TRFs.

Single-sample cluster-based permutation tests revealed that all of the TRFs and difference TRFs yielded patterns of instantaneous power which were significantly different to random chance. The temporal response functions for the attended and ignored speakers are depicted in the first two rows of Figure 2. They exhibit morphologies which are typical of TRFs obtained in higher-density EEG studies, with clear and significant components around latencies of 100 ms and 200 ms. The onset envelope TRFs also show pronounced activity at a much later latency of around 500 ms. The bottom row of Figure 2 shows the difference TRFs. For the envelope feature, the amplitude of the first two peaks of the unilateral-channel TRF are significantly modulated by selective attention. As for the onset envelope feature, the first positive peak of both the unilateral and bilateral channel TRFs is significantly modulated by selective attention. In contrast, there is no significant attentional modulation of the later components.

B. LINEAR BACKWARD MODELS

Linear backward models were fitted to predict features of the attended and ignored speech streams from the two-channel EEG signals. Correlation coefficients for each held-out trial were averaged for each participant, and the results are shown in Figure 3a. All eight groups of correlation coefficients have means which are statistically greater than their corresponding null distribution (all $p \ll 0.0001$; single-tailed unpaired t-tests with Bonferroni correction). The models which were trained using features from the attended speech stream produced reconstructed features which were more highly correlated with features of the attended speech stream than those of the ignored speech stream ($p \ll 0.0001$ for both the envelope feature, and its onsets; single-tailed paired t-tests.)

To further investigate how the attended-speaker models could be used for practical auditory attention decoding, we divided each held-out trial into short temporal segments of 5 s in duration with a 1-second hop length. Markers of selective auditory attention were derived from these segments using the procedure outlined in Section II-H. To assess whether these markers were informative at the level of individuals, we computed the attention decoding accuracy for each participant. The results are shown in Figure 3b, which also includes the upper 95% confidence limit of a random binary classifier as a reference. When calculating the upper limit of this classifier, we considered n independent trials of a Bernoulli random variable, where n was taken as the maximum number of non-overlapping 5 s windows available. When using the onset envelope feature, all 18 participants achieved decoding accuracies above this reference threshold. In contrast, three participants did not surpass this threshold when the envelope feature itself was used.

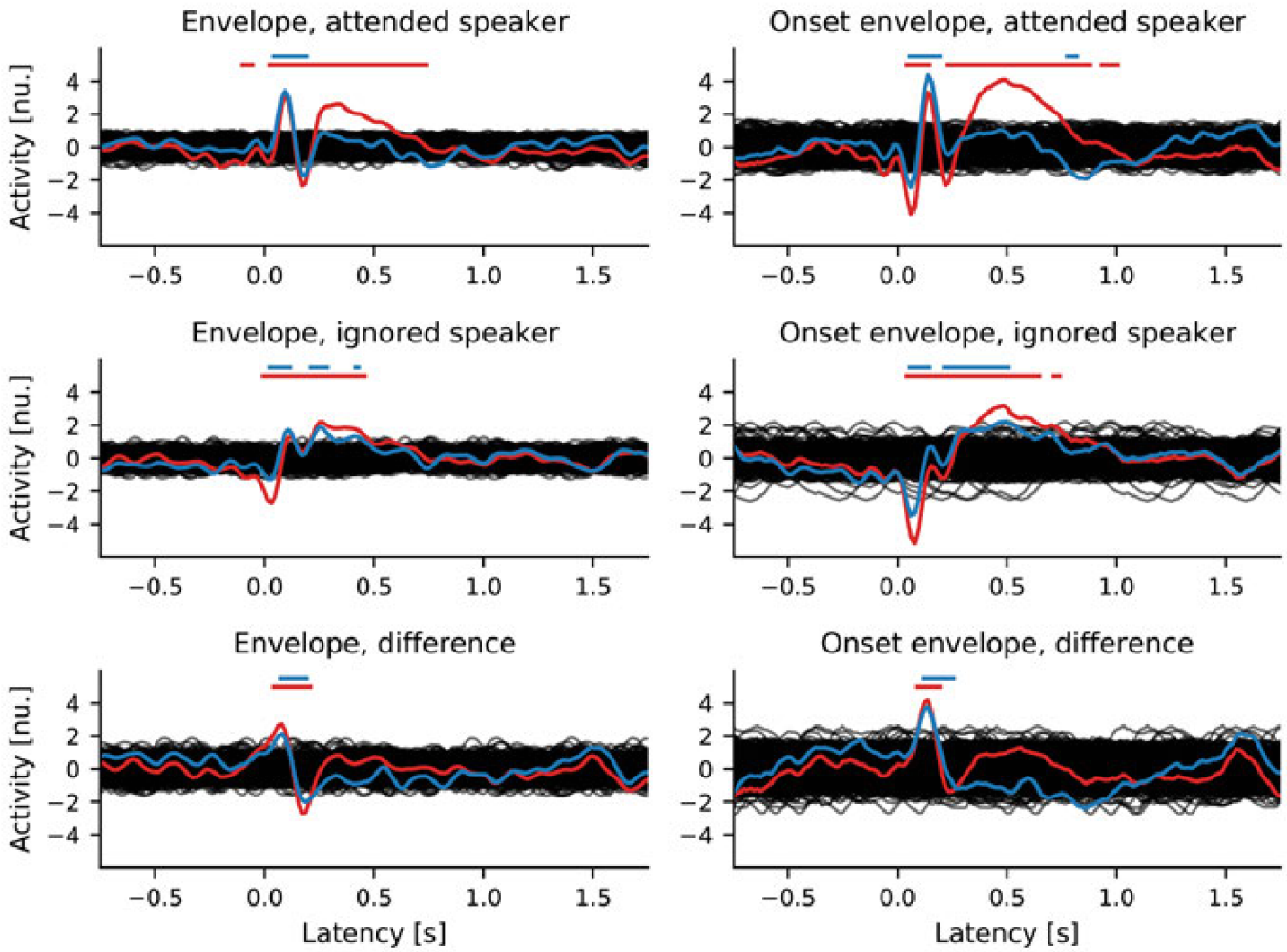

FIGURE 2. Temporal response functions for both the uni-lateral (blue) and bi-lateral (red) EEG channels. The first column shows TRFs computed from the temporal speech envelopes, and the second the onset-envelope TRFs. The first two rows show the grand-average TRFs for attended and ignored speech, respectively, and the bottom row shows the difference between the attended and ignored TRFs. Black lines represent null-TRFs, which were computed after temporally misaligning the EEG and speech-feature signals. Solid red and blue bars indicate temporal clusters with p-values smaller than 0.05/12, as calculated through a cluster-based permutation test.

C. COMPARISON OF DECODING ALGORITHMS

The three attention decoding algorithms were evaluated using segment lengths ranging from 0.1 s to 30 s. For each algorithm and speech feature, the mean attention decoding accuracy is plotted against segment length in Figure 4a. Overall, the mean accuracies lie above the chance level (defined as the 95th percentile of a random binary classifier) across a wide range of window sizes.

The decoding algorithms which utilised the onset envelopes achieved higher mean decoding accuracies than those based on the envelopes themselves. We tested whether these differences were significant using single-tailed, paired t-tests (Bonferroni-corrected for multiple comparisons), for both a short segment length of 5 s as well as a longer segment length of 30 s. For the 5 s window, we only detected a significant difference between the two CCA-based decoders. For the longer window length of 30 s, however, significant differences emerged between the envelope-based decoders and the onset-envelope-based decoders for all three decoder types. These results are shown in Figure 4b.

In general, the CCA decoders appeared to achieve the highest mean decoding accuracies, and the CNN-based decoders achieved the lowest. For each speech feature (onset envelope, and envelope), we performed repeated-measures ANOVA tests to assess whether there were any statistically significant differences between the three decoders. Since we performed these tests for both the 30 s segment length and the 5 s segment length, we performed a total of four ANOVA tests; none of these tests returned a positive result. The results of this analysis are also presented in Figure 4b.

IV. DISCUSSION AND CONCLUSION

We have demonstrated that our novel, ultra-wearable ear-EEG device captures auditory responses to the envelope of speech (and its onsets) from normal-hearing human listeners. The recorded responses were shown to be modulated by the

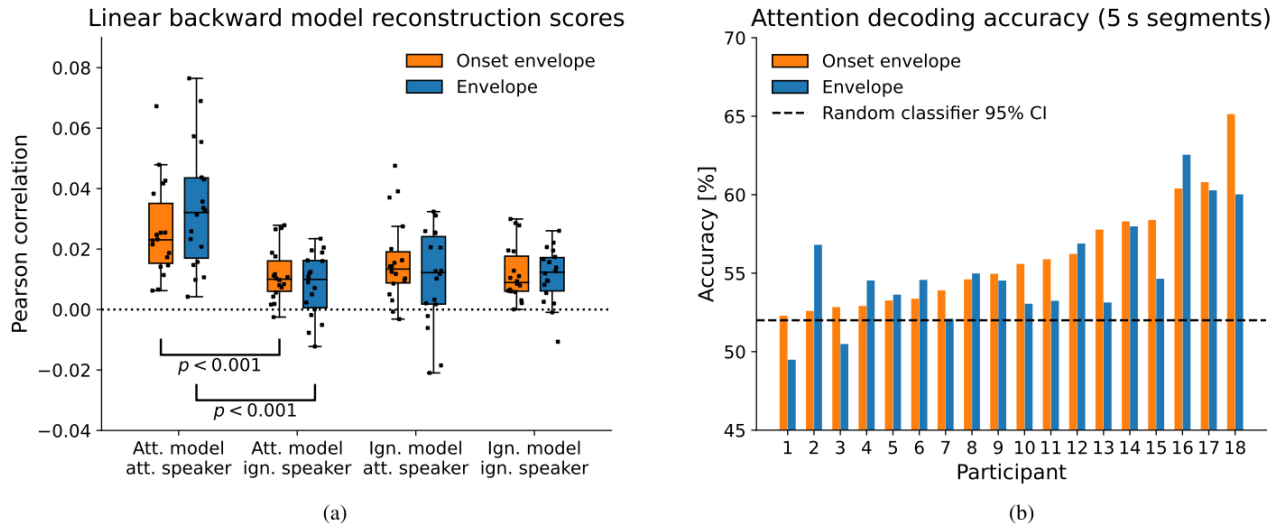

FIGURE 3. Performance of the linear backward models. a) The reconstructed speech features produced by the models were correlated against the actual features of both the attended and ignored speech streams. Correlation coefficients were calculated for each testing trial and participant, and the average correlation coefficients for each participant are displayed as black markers. b) Estimates of the attended talker were inferred from the correlation coefficients obtained in short segments of 5 s in duration. Segments for which the correlation coefficient corresponding to the attended talker exceeded that corresponding to the ignored talker were considered correctly classified. The average success rate over all windows was taken to be the attention decoding accuracy. The upper limit of a random binary classifier's 95% confidence interval is represented by a black dashed line.

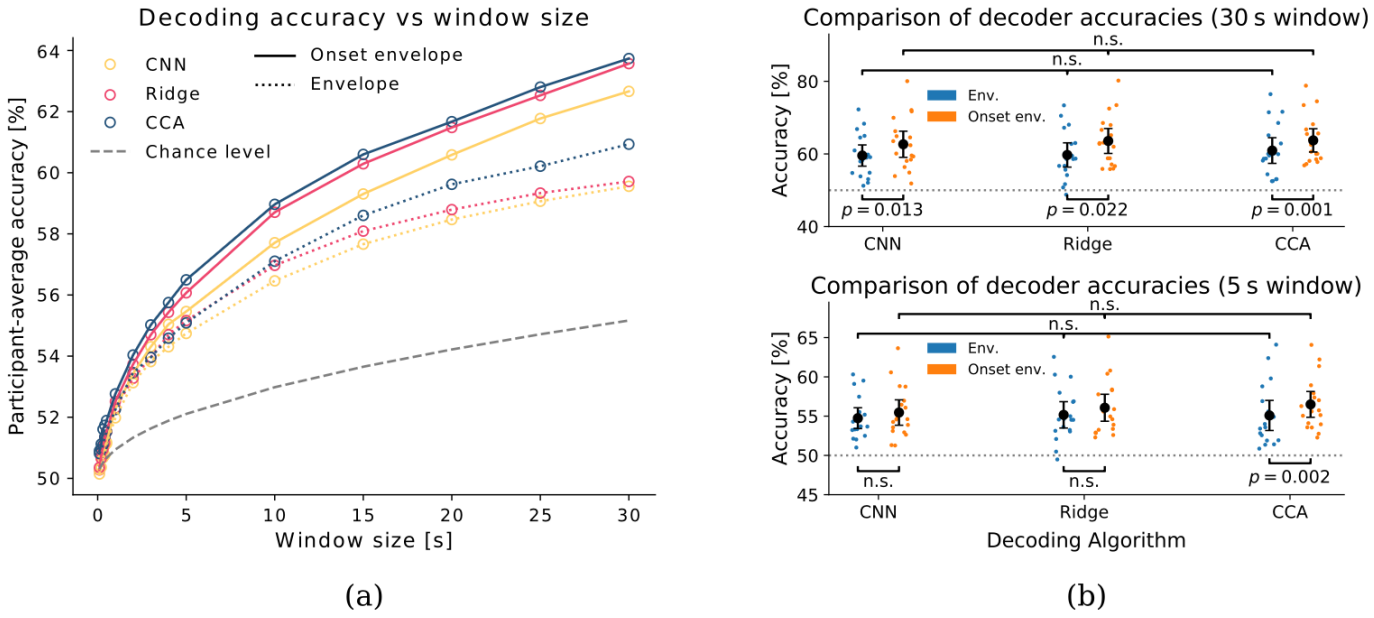

FIGURE 4. Performance of the attention decoding algorithms. a) The decoding accuracies for individual participants were averaged and plotted against segment length. The results for the algorithms which used the speech envelope feature are displayed with broken lines, and solid lines are used for those which utilised the onset envelope. The grey dashed line represents the upper limit of the 95% confidence interval of a random binary classifier. b) The attention decoding accuracies were broken down for individual participants for a longer segment length of 30 s (top) and a shorter window length of 5 s (bottom), and the mean attention decoding accuracies of different algorithms were assessed for statistical differences using single-tailed, paired t-tests.

focus of the listeners' selective attention to speech. We were therefore able to decode which of the two speakers the participants were attending to, using a variety of auditory attention decoding algorithms.

We assessed young, normal-hearing participants in our study because this population generally finds the auditory attention task less cognitively demanding—particularly in the absence of spatial separation between competing talkers—and is better able to tolerate longer experimental sessions. However, our ultimate target population for a prospective EEG-steered hearing aid includes hearing-impaired individuals, who are often older adults. There is broad consensus in the literature that both older adults and individuals with hearing impairment exhibit stronger neural tracking of the

speech envelope compared to young, normal-hearing listeners (for a review, see Gillies et al. [53]), Moreover, in listeners with mild to moderate hearing loss, this enhanced neural tracking appears to be more strongly modulated by attention, leading to improved attention decoding accuracies [54]. These findings suggest that the algorithms developed and validated in young, normal-hearing participants could generalize effectively to older or clinical populations, and may even benefit from enhanced neural envelope tracking in those groups.

Nevertheless, further work is needed to test these algorithms in hearing-impaired individuals and under more ecologically valid conditions, including real-world acoustic environments with background noise. In particular, it will be important to characterise the effect of the relative intensities of target and distracting speech on attention decoding accuracies. Such studies will be essential to assess the robustness and practical utility of EEG-based attention decoding in everyday listening scenarios, and would provide an excellent opportunity to evaluate how older and hearing-impaired users perceive ear-EEG technology in terms of comfort, usability, and acceptability.

Despite our simplified scenario in which listeners attended to one talker whilst ignoring another, it is important to note that real-world acoustic scenarios could involve a multitude of interfering talkers and other sources of background noise. Predictably, the error rate of an attention decoder compounds when the number of candidate speech signals is increased: in a four-talker auditory attention decoding study, Yan et al. [55] reported an accuracy of just 40% when they used a 5-second window length for their linear backward model. Whilst exceeding the chance level of 25% by a considerable margin, their decoder would more often make incorrect classifications than correct classifications. If the error could be significantly reduced in two-talker scenarios, then it would be expected to compound more slowly as additional distractors are added.

However, auditory attention decoders do not need to be optimised for an arbitrary number of talkers. In real-world acoustic scenarios it is appropriate to consider up to several salient voices superposed onto multi-talker babble noise and other environmental sounds. Considering exactly this scenario, Das et al. [56] showed that at moderate SNRs, background babble noise can actually improve the accuracy of an attention decoder in a two-talker scenario. This could be due an increased attentional effect or stronger envelope tracking due to the increased difficulty of the attention task. However, the attention decoding accuracy quickly deteriorated at lower SNRs. This falls broadly in line with studies which report a degradation in neural envelope tracking when the intelligibility of the speech stimulus is decreased [57], [58], [59]. An effective EEG-steered hearing aid should be able to mitigate this by maintaining high speech intelligibility even in adverse listening conditions; this could be facilitated in part by the sophisticated noise-reduction capabilities of conventional hearing aids [57], [60].

Temporal response functions obtained from the ear-EEG device exhibit a clear morphology which is consistent with the existing scalp-EEG-based literature, with strong components at latencies of around 100-200 ms. The onset-envelope TRFs also exhibit broad regions of activity at a much later latency of around 500 ms which was not reported in the ear-EEG study of Fiedler et al. [16]. This is likely because the high-pass filter which we pre-processed the EEG signals with had a much lower cutoff frequency than that which was used in that study (0.5 Hz vs 2 Hz). This means that our TRFs are influenced by contributions from very low-frequency cortical tracking.

Linear backward models were trained to relate the two-channel EEG recordings to speech features of the attended speaker and the ignored speaker. For both speech features, the attended-speaker models better predicted the features of the attended speech stream than those of the ignored speech stream. By contrast, the ignored-speaker models achieved similar correlation scores for both speakers. We therefore used the attended-speaker models as auditory attention decoders. Markers of selective auditory attention were formed by subtracting the correlation coefficients obtained for the ignored speech stream from those obtained for the attended speech stream, and an attention decoding accuracy for each participant was derived from their attention markers. We found that the onset-envelope-based attention decoders produced statistically meaningful attention markers for all 18 participants when a short, practical segment length of 5 s was used, with the corresponding attention decoding accuracies exceeding the upper 95% confidence limit of a random binary classifier. For the envelope-based decoders, decoding accuracies of three participants did not exceed this threshold, with two participants showing very low attention decoding scores of around 50%. Overall, we observed that the onset-envelope-based decoders were more successful than the envelope-based decoders, and there were no clear outlying scores in the data.

Three types of algorithms were used to perform auditory attention decoding: the linear backward models; non-linear CNN-based backward models; and CCA-based decoders. Overall, the attention decoding accuracies were not that high, lying in the region of 55% for short segment lengths of 5 s in duration. These results are comparable to those reported by Fiedler et al. [16], as well as the scalp-EEG study by Narayananet al. [61], who simulated miniature EEG montages by selecting subsets of electrodes from a 255-channel system. These comparisons suggest that our dry-contact ear-EEG system records signals with a signal-to-noise ratio comparable to that of gel-based systems. Moreover, ear-EEG-based decoders achieve performance levels similar to those of low-density scalp-based decoders, with the added advantages of increased wearability and discretion.

Clearly, devices such as smart hearing aids which rely on EEG-based auditory attention decoding are unlikely to provide any benefit to the user unless the attention decoding

accuracy could be significantly enhanced. One possibility would be to employ state-space models to obtain less noisy estimates of the attended speaker label from the original attention markers: Akram et al. [62], [63] developed an algorithm which employs a Kalman filter to denoise the correlation-based attention markers, leading to a moderate improvement in attention decoding accuracy; Hjortkjaer et al. [27] proposed an alternative state-space model which weights the importance of incoming attention markers based on the recent performance of the underlying attention decoder; many more possibilities exist.

An orthogonal approach to improving the attention decoding accuracy of a hearing aid system, for example, could be by leveraging the information available in other sensing modalities via sensor fusion. When a listener is able to see the target talker, the direction of their gaze provides a salient marker of auditory attention. This can be monitored through modern eye-tracking systems, which can be built into a glasses-like form factor, or even more unobtrusively, measured via electro-oculography (EOG) signals. Skoglund et al. [64] recently demonstrated an in-ear EOG system which could decode auditory attention via detecting saccades to the left or right; their system achieved an accuracy of 68%, which is comparable to our EEG-based system. A composite system could potentially achieve much higher attention decoding accuracies.

In naturalistic listening, however, people do not always look in the direction of a target speaker, and more natural cues for selective auditory attention may be desired. Recent studies have suggested that more subtle patterns of eye activity, such as blink rate and microsaccades, may correlate with auditory attention [65], [66], [67]. While these studies support the potential of eye-based attention markers, they generally report small effect sizes and are likely to have very low attention decoding accuracies, especially if recorded from ear-based sensors. However, future work should investigate whether integrating these markers of selective auditory attention with ear-EEG-based markers in a sensor fusion algorithm could lead to improved decoding accuracies.

In addition to decoding accuracy, latency and computational overhead are critical factors for real-time applications of auditory EEG decoders. All attention decoding models considered in this study rely on a fixed window length (e.g., 5 seconds), which imposes an inherent delay—data must first be accumulated over this window before an attention estimate can be produced. Reducing the window length typically degrades decoding accuracy, unless the underlying decoder is also paired with some kind of state-space filter [27], [63], [68]. During preprocessing, we filtered EEG signals in the 0.5—8 Hz band, a range commonly associated with speech envelope tracking. While digital FIR filters can introduce substantial latency due to their high filter order, real-time implementation remains feasible by using IIR filters, which offer comparable attenuation with significantly lower latency. In hardware-constrained scenarios, analog IIR filters may also be implemented directly within hearing devices. From a computational standpoint, the linear models used in this study (e.g., stimulus reconstruction via linear regression and canonical correlation analysis) are highly efficient in terms of both processing time and memory usage. These models have been shown in prior work to operate effectively in near-real-time [27], [69], [70]. In contrast, neural network models—such as the lightweight CNN employed here—require greater computational resources. Efficient real-time deployment of such models may depend on the use of dedicated hardware accelerators (e.g., on-chip AI engines) or offloading computations to a companion device, such as a smartphone.

We did not detect any significant difference in mean accuracy between the three types of decoding algorithm for short segment lengths of 5 s in duration, nor for longer segment lengths of 30 s. It is possible that with only 18 participants, we did not have enough statistical power to detect such a difference. Based on the effect sizes that we observed, we estimate that around 50 participants would have been required to detect statisticaly significant differences—in case these were indeed significant—between the attention decoding accuracies of the decoders for a 30 s window size, and around 90 would have been required for the 5 s window size. Nevertheless, even if the observed differences were in fact statistically significant, they are, at a few percentage points, presumably be too small to be practically relevant.

The deep-learning-based decoders achieved lower mean decoding accuracies than their linear counterparts for all segment lengths, which contrasts with the recent scalp-EEG literatures [22] and [23]. This could be due to the that fact that the ear-EEG signals offered a poorer SNR, since the dry-contact electrodes were placed in locations which are prone to mastication artifacts; the CNNs may therefore have overfitted to artifactual signals during training.

The choice of speech feature (temporal envelope versus its onsets) did impact the final attention decoding accuracy. These speech features were based on biologically-inspired heuristics, and it is possible that by refining them (and possibly tailoring them to individuals, since Figure 3 shows that sometimes one feature is better than another for some participants), further improvements in auditory attention decoding may be made. Such an approach could employ techniques of deep learning to approximate the representations of speech which are present in ear-EEG recordings, following similar approaches to modelling invasively-measured electrophysiological responses to speech [71].

REFERENCES

[1] N. A. Lesica, "Why do hearing aids fail to restore normal auditory perception?" *Trends Neurosci.*, vol. 41, no. 4, pp. 174–185, Apr. 2018.

[2] A. McCormack and H. Fortnum, "Why do people fitted with hearing aids not wear them?" *Int. J. Audiol.*, vol. 52, no. 5, pp. 360–368, Mar. 2013.

[3] R. Bentler and L.-K. Chiou, "Digital noise reduction: An overview," *Trends Amplification*, vol. 10, no. 2, pp. 67–82, Jun. 2006.

[4] J. L. Yanz and D. Preves, "Telecoils: Principles, pitfalls, fixes, and the future," *Seminars Hearing*, vol. 24, no. 1, pp. 029–042, 2003.

[5] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*. Hoboken, NJ, USA: Wiley-Blackwell, 2010, pp. 269–302.

[6] L. Magnusson, A. Claesson, M. Persson, and T. Tengstrand, "Speech recognition in noise using bilateral open-fit hearing aids: The limited benefit of directional microphones and noise reduction," *Int. J. Audiology*, vol. 52, no. 1, pp. 29–36, Aug. 2012.

[7] M. T. Cord, R. K. Surr, B. E. Walden, and O. Dyrlund, "Relationship between laboratory measures of directional advantage and everyday success with directional microphone hearing aids," *J. Amer. Acad. Audiol.*, vol. 15, no. 5, pp. 353–364, May 2004.

[8] T. A. Ricketts and J. Galster, "Head angle and elevation in classroom environments: Implications for amplification," *J. Speech, Lang., Hearing Res.*, vol. 51, no. 2, pp. 516–525, Apr. 2008.

[9] W. O. Brimijoin, W. M. Whitmer, D. McShefferty, and M. A. Akeroyd, "The effect of hearing aid microphone mode on performance in an auditory orienting task," *Ear Hearing*, vol. 35, no. 5, pp. e204–e212, Sep. 2014.

[10] A. W. Archer-Boyd, J. A. Holman, and W. O. Brimijoin, "The minimum monitoring signal-to-noise ratio for off-axis signals and its implications for directional hearing aids," *Hearing Res.*, vol. 357, pp. 64–72, Jan. 2018.

[11] P. Kidmose, M. L. Rank, M. Ungstrup, D. Looney, C. Park, and D. P. Mandic, "A yarbus-style experiment to determine auditory attention," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol.*, Aug. 2010, pp. 4650–4653.

[12] D. Looney, C. Park, Y. Xia, P. Kidmose, M. Ungstrup, and D. P. Mandic, "Towards estimating selective auditory attention from EEG using a novel time-frequency-synchronisation framework," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2010, pp. 1–5.

[13] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, Jan. 2014.

[14] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigne, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices," *IEEE Signal Process. Mag.*, vol. 38, no. 4, pp. 89–102, Jul. 2021.

[15] M. G. Bleichner, B. Mirkovic, and S. Debener, "Identifying auditory attention with ear-EEG: CEEGrid versus high-density cap-EEG comparison," *J. Neural Eng.*, vol. 13, no. 6, Oct. 2016, Art. no. 066004.

[16] L. Fiedler, M. Wöstmann, C. Graversen, A. Brandmeyer, T. Lunner, and J. Obleser, "Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech," *J. Neural Eng.*, vol. 14, no. 3, Apr. 2017, Art. no. 036020.

[17] D. D. E. Wong, S. A. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, and A. de Cheveigné, "A comparison of regularization methods in forward and backward models for auditory attention decoding," *Frontiers Neurosci.*, vol. 12, Aug. 2018, Art. no. 531.

[18] O. Etard, M. Kegler, C. Braiman, A. E. Forte, and T. Reichenbach, "Decoding of selective attention to continuous speech from the human auditory brainstem response," *NeuroImage*, vol. 200, pp. 1–11, Oct. 2019.

[19] N. Ding and J. Z. Simon, "Cortical entrainment to continuous speech: Functional roles and interpretations," *Frontiers Human Neurosci.*, vol. 8, May 2014, Art. no. 311.

[20] G. Ciccarelli, M. Nolan, J. Perricone, P. T. Calamia, S. Haro, J. O'Sullivan, N. Mesgarani, T. F. Quatieri, and C. J. Smalt, "Comparison of two-talker attention decoding from EEG with nonlinear neural networks and linear methods," *Sci. Rep.*, vol. 9, no. 1, Aug. 2019, Art. no. 11538.

[21] C. Puffay, B. Accou, L. Bollens, M. J. Monesi, J. Vanthornhout, H. Van hamme, and T. Francart, "Relating EEG to continuous speech using deep neural networks: A review," *J. Neural Eng.*, vol. 20, no. 4, Aug. 2023, Art. no. 041003.

[22] T. de Taillez, B. Kollmeier, and B. T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *Eur. J. Neurosci.*, vol. 51, no. 5, pp. 1234–1241, Jan. 2018.

[23] M. Thornton, D. Mandic, and T. Reichenbach, "Robust decoding of the speech envelope from EEG recordings through deep neural networks," *J. Neural Eng.*, vol. 19, no. 4, Jul. 2022, Art. no. 046007.

[24] M. D. Thornton, D. P. Mandic, and T. J. Reichenbach, "Decoding envelope and frequency-following EEG responses to continuous speech using deep neural networks," *IEEE Open J. Signal Process.*, vol. 5, pp. 700–716, 2024.

[25] B. Accou, J. Vanthornhout, H. V. Hamme, and T. Francart, "Decoding of the speech envelope from EEG using the VLAAI deep neural network," *Sci. Rep.*, vol. 13, no. 1, Jan. 2023, Art. no. 812.

[26] A. de Cheveigné, D. D. E. Wong, G. M. Di Liberto, J. Hjortkjær, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, May 2018.

[27] J. Hjortkjær, D. D. E. Wong, A. Catania, J. Märcher-Rørsted, E. Ceolini, S. A. Fuglsang, I. Kiselev, G. Di Liberto, S.-C. Liu, T. Dau, M. Slaney, and A. de Cheveigné, "Real-time control of a hearing instrument with EEG-based attention decoding," *J. Neural Eng.*, vol. 22, no. 1, Feb. 2025, Art. no. 016027.

[28] S. L. Kappel, M. L. Rank, H. O. Toft, M. Andersen, and P. Kidmose, "Dry-contact electrode ear-EEG," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 1, pp. 150–158, Jan. 2019.

[29] D. Looney, P. Kidmose, C. Park, M. Ungstrup, M. L. Rank, K. Rosenkranz, and D. P. Mandic, "The in-the-ear recording concept: User-centered and wearable brain monitoring," *IEEE Pulse*, vol. 3, no. 6, pp. 32–42, Nov. 2012.

[30] S. Debener, R. Emkes, M. De Vos, and M. Bleichner, "Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear," *Sci. Rep.*, vol. 5, no. 1, Nov. 2015, Art. no. 16743.

[31] Z. Wang, N. Shi, Y. Zhang, N. Zheng, H. Li, Y. Jiao, J. Cheng, Y. Wang, X. Zhang, Y. Chen, Y. Chen, H. Wang, T. Xie, Y. Wang, Y. Ma, X. Gao, and X. Feng, "Conformal in-ear bioelectronics for visual and auditory brain–computer interfaces," *Nature Commun.*, vol. 14, no. 1, Jul. 2023, Art. no. 4213.

[32] E. Azemi, A. Moin, A. Pragada, J. H.-C. Lu, V. M. Powell, J. Minxha, and S. P. Hotelling, "Biosignal sensing device using dynamic selection of electrodes," U.S. Patent 20 230 225 659 A1, Jul. 20, 2023.

[33] D. Looney, C. Park, P. Kidmose, M. L. Rank, M. Ungstrup, K. Rosenkranz, and D. P. Mandic, "An in-the-ear platform for recording electroencephalogram," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 6882–6885.

[34] D. Looney, P. Kidmose, M. J. Morrell, and D. P. Mandic, "Ear-EEG: Continuous brain monitoring," in *Brain-Computer Interface Research* (SpringerBriefs in Electrical and Computer Engineering). Cham, Switzerland: Springer, 2014, pp. 63–71.

[35] P. Kidmose, D. Mandic, M. Ungstrup, D. Looney, C. Park, and M. Rank, "A hearing aid adapted for detection brain waves and a method for adapting such a hearing aid," U.S. Patent 9 025 800 B2, May 5, 2015.

[36] M. Garrett, S. Debener, and S. Verhulst, "Acquisition of subcortical auditory potentials with around-the-ear cEEGrid technology in normal and hearing impaired listeners," *Frontiers Neurosci.*, vol. 13, Jul. 2019, Art. no. 730.

[37] B. Holtze, M. Rosenkranz, M. Jaeger, S. Debener, and B. Mirkovic, "Ear-EEG measures of auditory attention to continuous speech," *Frontiers Neurosci.*, vol. 16, May 2022, Art. no. 539.

[38] A. Meiser, F. Tadel, S. Debener, and M. G. Bleichner, "The sensitivity of ear-EEG: Evaluating the source-sensor relationship using forward modeling," *Brain Topography*, vol. 33, no. 6, pp. 665–676, Nov. 2020.

[39] M. C. Yarici, M. Thornton, and D. P. Mandic, "Ear-EEG sensitivity modeling for neural sources and ocular artifacts," *Frontiers Neurosci.*, vol. 16, Jan. 2023, Art. no. 997377.

[40] I. Hertrich, S. Dietrich, J. Trouvain, A. Moos, and H. Ackermann, "Magnetic brain activity phase-locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech signal," *Psychophysiology*, vol. 49, no. 3, pp. 322–334, Mar. 2012.

[41] E. Occhipinti, H. J. Davies, G. Hammour, and D. P. Mandic, "Hearables: Artefact removal in ear-EEG for continuous 24/7 monitoring," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–6.

[42] G. M. Hammour and D. P. Mandic, "Hearables: Making sense from motion artefacts in ear-EEG for real-life human activity classification," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 6889–6893.

[43] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 5, pp. 402–412, May 2017.

[44] (2025). *Brian2hears: A Library for Auditory Modeling With Brian2*. Accessed: Jun. 11, 2025. [Online]. Available: https://github.com/brian-team/brian2hears
[45] E. C. Lalor and J. J. Foxe, "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution," *Eur. J. Neurosci.*, vol. 31, no. 1, pp. 189–193, Jan. 2010.
[46] D. P. Mandic and J. A. Chambers, *Recurrent Neural Networks for Prediction*. Hoboken, NJ, USA: Wiley, Aug. 2001.
[47] L. Stankovic and D. Mandic, "Convolutional neural networks demystified: A matched filtering perspective-based tutorial," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 6, pp. 3614–3628, Jun. 2023.
[48] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.
[49] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, Jan. 2015, pp. 448–456.
[50] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, vol. 31, Dec. 2018, pp. 2488–2498.
[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, Dec. 2014.
[53] M. Gillis, J. Van Canneyt, T. Francart, and J. Vanthornhout, "Neural tracking as a diagnostic tool to assess the auditory pathway," *Hearing Res.*, vol. 426, Dec. 2022, Art. no. 108607.
[54] S. A. Fuglsang, J. Märcher-Rørsted, T. Dau, and J. Hjortkjær, "Effects of sensorineural hearing loss on cortical synchronization to competing speech during selective attention," *J. Neurosci.*, vol. 40, no. 12, pp. 2562–2572, Feb. 2020.
[55] Y. Yan, X. Xu, H. Zhu, P. Tian, Z. Ge, X. Wu, and J. Chen, "Auditory attention decoding in four-talker environment with EEG," in *Proc. Interspeech*, Sep. 2024, pp. 432–436.
[56] N. Das, A. Bertrand, and T. Francart, "EEG-based auditory attention detection: Boundary conditions for background noise and speaker positions," *J. Neural Eng.*, vol. 15, no. 6, Oct. 2018, Art. no. 066017.
[57] T. Van Hirtum, B. Somers, B. Dieudonné, E. Verschueren, J. Wouters, and T. Francart, "Neural envelope tracking predicts speech intelligibility and hearing aid benefit in children with hearing loss," *Hearing Res.*, vol. 439, Nov. 2023, Art. no. 108893.
[58] I. Iotzov and L. C. Parra, "EEG can predict speech intelligibility," *J. Neural Eng.*, vol. 16, no. 3, Mar. 2019, Art. no. 036008.
[59] N. Ding, M. Chatterjee, and J. Z. Simon, "Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure," *NeuroImage*, vol. 88, pp. 41–46, Mar. 2014.
[60] A. Aroudi and S. Doclo, "EEG-based auditory attention decoding: Impact of reverberation, noise and interference reduction," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2017, pp. 3042–3047.
[61] A. Mundanad Narayanan, R. Zink, and A. Bertrand, "EEG miniaturization limits for stimulus decoding with EEG sensor networks," *J. Neural Eng.*, vol. 18, no. 5, Oct. 2021, Art. no. 056042.
[62] S. Akram, J. Z. Simon, S. Shamma, and B. Babadi, "A state-space model for decoding auditory attentional modulation from MEG in a competing-speaker environment," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, Dec. 2014, pp. 460–468.
[63] S. Miran, S. Akram, A. Sheikhattar, J. Z. Simon, T. Zhang, and B. Babadi, "Real-time tracking of selective auditory attention from M/EEG: A Bayesian filtering approach," *Frontiers Neurosci.*, vol. 12, May 2018, Art. no. 262.
[64] M. A. Skoglund, M. Andersen, M. M. Shiell, G. Keidser, M. L. Rank, and S. Rotger-Griful, "Comparing in-ear EOG for eye-movement estimation with eye-tracking: Accuracy, calibration, and speech comprehension," *Frontiers Neurosci.*, vol. 16, Jun. 2022, Art. no. 873201.
[65] B. Holtze, M. Rosenkranz, M. Bleichner, M. Jaeger, and S. Debener, "Eye-blink patterns reflect attention to continuous speech," *Adv. Cognit. Psychol.*, vol. 19, no. 2, pp. 177–200, 2023.
[66] P. Jin, J. Zou, T. Zhou, and N. Ding, "Eye activity tracks task-relevant structures during speech and auditory sequence perception," *Nature Commun.*, vol. 9, no. 1, Dec. 2018, Art. no. 5374.
[67] Q. Gehmacher, J. Schubert, F. Schmidt, T. Hartmann, P. Reisinger, S. Rösch, K. Schwarz, T. Popov, M. Chait, and N. Weisz, "Eye movements track prioritized auditory features in selective attention to natural speech," *Nature Commun.*, vol. 15, no. 1, May 2024, Art. no. 3692.
[68] N. Heintz, S. Geirnaert, I. Van De Ryck, T. Francart, and A. Bertrand, "Probabilistic gain control in a multi-speaker setting using EEG-based auditory attention decoding," in *Proc. 32nd Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2024, pp. 892–896.
[69] A. Aroudi, E. Fischer, M. Serman, H. Puder, and S. Doclo, "Closed-loop cognitive-driven gain control of competing sounds using auditory attention decoding," *Algorithms*, vol. 14, no. 10, Sep. 2021, Art. no. 287.
[70] S. Haro, C. Beauchene, T. F. Quatieri, and C. J. Smalt, "A brain–computer interface for improving auditory attention in multi-talker environments," *BioRxiv*, Mar. 2025, Art. no. 641661.
[71] F. Drakopoulos, S. Sabesan, Y. Xia, A. Fragner, and N. A. Lesica, "Modeling neural coding in the auditory brain with high resolution and accuracy," *BioRxiv*, Jun. 2024, Art. no. 599294.

MIKE D. THORNTON received the M.Phys. degree in physics from the University of Oxford, U.K., in 2020, and the Ph.D. degree in AI and machine learning from Imperial College London, U.K., in 2025.

He is currently a Postdoctoral Researcher with Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany. He has particular research interests in neural signal processing and the neural basis of speech processing.

Dr. Thornton's team developed award-winning submissions to the 2023 and 2024 ICASSP Auditory EEG Decoding Signal Processing Grand Challenges.

DANILO P. MANDIC (Fellow, IEEE) received the Ph.D. degree in nonlinear adaptive signal processing from Imperial College London, U.K., in 1999.

He is currently a Professor of machine intelligence with Imperial College London. His publication record includes two research monographs titled *Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures, and Stability* (Wiley, First Edition), in August 2001, and *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear, and Neural Models* (Wiley, First Edition), in April 2009, an edited book titled *Signal Processing Techniques for Knowledge Extraction and Information Fusion* (Springer), in 2008, and a two-volume research monograph titled *Tensor Networks for Dimensionality Reduction and Large Scale Optimization* (Now Publishers), in 2016 and 2017.

Dr. Mandic received the Denis Gabor Award for Outstanding Achievements in Neural Engineering by the International Neural Networks Society, the 2018 Best Paper Award from *IEEE Signal Processing Magazine* for his article on tensor decompositions, and the President's Award for Excellence in Postgraduate Supervision at Imperial College London. He has been an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS, IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS.

TOBIAS REICHENBACH received the M.Sc. degree in physics from Leipzig University, Germany, and the Ph.D. degree in physics from LMU Munich, Germany.

He is currently a Professor of sensory neuroengineering with the Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany, and a Visiting Professor with the Department of Bioengineering, Imperial College London. He previously held positions as a Postdoctoral Researcher in the group of Prof. A. J. Hudspeth at The Rockefeller University, New York, NY, USA, and subsequently as a Lecturer, a Senior Lecturer, and a Reader with Imperial College London. His multidisciplinary research combines methods from artificial intelligence with computational neuroscience and neuroimaging to advance our understanding of the neural processing of complex natural signals, with applications in medicine and technology.

Dr. Reichenbach's research has been recognized by several international awards and fellowships, such as the Career Award at the Scientific Interface from the Burroughs Wellcome Fund, USA, and a Fellowship from the Engineering and Physical Sciences Research Council, U.K.