Single-microphone deep envelope separation based auditory attention decoding for competing speech and music

Tanveer, M Asjid; Jensen, Jesper; Tan, Zheng-Hua; Østergaard, Jan

Published in:
Journal of Neural Engineering

DOI (link to publication from Publisher):
10.1088/1741-2552/add0e7

Creative Commons License
CC BY 4.0

Publication date:
2025

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Tanveer, M. A., Jensen, J., Tan, Z.-H., & Østergaard, J. (2025). Single-microphone deep envelope separation based auditory attention decoding for competing speech and music. *Journal of Neural Engineering*, *22*(3), Article 036006. https://doi.org/10.1088/1741-2552/add0e7

Journal of Neural
Engineering

PAPER • OPEN ACCESS

Single-microphone deep envelope separation based auditory attention decoding for competing speech and music

To cite this article: M Asjid Tanveer *et al* 2025 *J. Neural Eng.* **22** 036006

Journal of Neural Engineering

OPEN ACCESS

RECEIVED
11 October 2024

REVISED
17 April 2025

ACCEPTED FOR PUBLICATION
25 April 2025

PUBLISHED
7 May 2025

PAPER

Single-microphone deep envelope separation based auditory attention decoding for competing speech and music

M Asjid Tanveer[1,*], Jesper Jensen[1,2], Zheng-Hua Tan[1] and Jan Østergaard[1]

[1] Department of Electronic systems, Aalborg University, Aalborg, Denmark
[2] Oticon A/S, Copenhagen, Denmark
[*] Author to whom any correspondence should be addressed.

E-mail: asjidt@es.aau.dk, jje@es.aau.dk, zt@es.aau.dk and jo@es.aau.dk

Keywords: auditory attention, deep learning, EEG, head related transfer functions, source separation, speech and music

Abstract

Objective. In this study, we introduce an end-to-end single microphone deep learning system for source separation and auditory attention decoding (AAD) in a competing speech and music setup. Deep source separation is applied directly on the envelope of the observed mixed audio signal. The resulting separated envelopes are compared to the envelope obtained from the electroencephalography (EEG) signals via deep stimulus reconstruction, where Pearson correlation is used as a loss function for training and evaluation. *Approach.* Deep learning models for source envelope separation and AAD are trained on target/distractor pairs from speech and music, covering four cases: speech vs. speech, speech vs. music, music vs. speech, and music vs. music. We convolve 10 different HRTFs with our audio signals to simulate the effects of head, torso and outer ear, and evaluate our model's ability to generalize. The models are trained (and evaluated) on 20 s time windows extracted from 60 s EEG trials. *Main results.* We achieve a target Pearson correlation and accuracy of 0.122% and 82.4% on the original dataset and an average target Pearson correlation and accuracy of 0.106% and 75.4% across the 10 HRTF variants. For the distractor, we achieve an average Pearson correlation of 0.004. Additionally, our model gives an accuracy of 82.8%, 85.8%, 79.7% and 81.5% across the four aforementioned cases for speech and music. With perfectly separated envelopes, we can achieve an accuracy of 83.0%, which is comparable to the case of source separated envelopes. *Significance.* We conclude that the deep learning models for source envelope separation and AAD generalize well across the set of speech and music signals and HRTFs tested in this study. We notice that source separation performs worse for a mixed music and speech signal, but the resulting AAD performance is not impacted.

1. Introduction

Humans with normal hearing have the capacity to selectively attend to a talker while filtering out unwanted signals such as background noise, reverberation, and interfering talkers [1, 2]. However, hearing impaired (HI) listeners often lose the ability of selective attention and could struggle to follow a conversation, when there are multiple talkers [3]. Speech separation algorithms could therefore play a key role in hearing aid technology by processing a multi-talker signal through the separation of individual talkers by decomposing a mixed audio signal into individual sound sources [4–6].

Recent advancements in deep learning have noticeably improved the accuracy and efficiency of source waveform separation [6–8]. Additionally, methods such as beamforming [9] can be employed to emphasize the target source while suppressing surrounding sources. However, for these algorithms to be useful in multi-talker scenarios, it is necessary to identify the target source, which can be done by different techniques including auditory attention decoding (AAD) [10, 11]. Some studies propose an end-to-end AAD beamforming approach for speaker identification [12], however for such cases a multi-microphone system is necessary for beamforming. Discerning speech in complex auditory scenarios such as cocktail party problems therefore remains a significant concern for individuals using hearing aids [13].

AAD has seen remarkable advancements in recent years. Researchers have employed various methodologies such as magnetoencephalography and electroencephalography (EEG) [14, 15] to decode auditory attention by tracking neural oscillatory patterns. Machine learning models have been used with EEG [16–18] to identify the attended speaker in multi-speaker scenarios while other studies [19] have investigated neural mechanisms underlying selective auditory attention, shedding light on the neural circuits involved in auditory attention processes. The evolution of deep learning models, especially those anticipated for integration into future EEG-guided hearing aids for real-time AAD, holds significant promise in enhancing speech comprehension for individuals with hearing impairments (HIs). These advancements aim to improve focus and clarity in noisy environments with multiple speakers, making communication more accessible and effective [20, 21]. Recent research has shown encouraging developments in creating such models [22–25].

Traditionally, AAD models have typically been restricted to linear regression models [10, 26]. However, recent research has increasingly shifted the focus to using deep learning models for AAD [27, 28]. Deep learning architectures can decipher complex patterns in auditory stimuli and uncover the underlying neural representations [29]. Recent studies have ventured into applying deep learning techniques, particularly convolutional neural networks (CNNs), to decode auditory attention [30, 31]. Additionally, long-short term memory (LSTM) [32] and CNN-LSTM [33] have shown promising results for AAD. The linear models are more commonly used due to their simplicity (less parameters) [34] and have been used across various datasets. Non-linear models in comparison have typically larger number of parameters and have shown to typically outperform linear models [32, 35]. These investigations have underscored the effectiveness of deep learning architectures in discerning auditory attention within shorter time intervals (5-20 seconds) as compared to linear models, which typically require longer time windows to achieve comparable performance [10].

Stimulus reconstruction has shown to be sensitive to auditory attention [10]: in a scenario where a listener is told to focus on one sound in a multi-speaker scenario, the cortical tracking of the attended target increases compared to the unattended target [16, 26, 36]. Previous studies have commonly used stimulus reconstruction for AAD on sources consisting of only speech [11, 31, 36, 37] while sources consisting of music and its combination with speech are not as prevalent in the literature concerning AAD [38–40].

Source waveform separation provides an elegant solution to separating competing talkers in multi-talker scenarios but quickly becomes computationally expensive when the signals have high sampling rates [41]. Envelope source separation is a good alternative where we perform source separation directly on the envelope rather than the waveform of the mixed source, as it allows downsampling of data from high sampling rates (48 kHz, 16 kHz, etc) to a low sampling rate (64 Hz). Additionally, envelope source separation allows one to have quick access to envelopes, which are features commonly used for AAD.

In this study, we propose a novel single microphone deep learning based system on competing speech and music sources, consisting of envelope based source separation and AAD using stimulus reconstruction with Pearson correlation [42] as a loss function. We showcase that envelope based deep source separation model results in less computation time compared to waveform based source separation. Additionally, our models are trained and tested on a dataset with sources consisting of both speech and music. We outperform the previous study by Simon *et al* [10] on this dataset, achieving better performance metrics across all four target and distractor pairs of speech and music. Unlike their approach, which assumes perfectly separated envelopes (i.e. ideal source separation) and uses separate models for each pair, we train a single model across all pairs for both source separation and AAD. We evaluate our models in simulated conditions using head related transfer functions (HRTFs) to analyze the impact that the head, outer ears, and torso have on the acoustic signals. This allows us to test our models' ability to generalize across the HRTF variants. We also showcase that source separation in the envelope domain suffers no AAD performance degradation compared to the ideal case, where sources are not mixed. Finally, we also perform a subject independent study and observe only a minor degradation in AAD performance compared to the case of using subject dependent models. This work extends the framework proposed in [43], which examined AAD leveraging envelope-based source separation in speech and music contexts. The present study advances this approach by evaluating its robustness and generalizability through cross-HRTF analysis, subject-independent decoding, and evaluation on an additional dataset.

The remainder of this paper is arranged as follows. Section 2 discusses the methodology used for envelope separation, AAD and the aforementioned simulations, section 3 discusses the results obtained during this study and section 4 concludes the paper.

2. Materials and methods

2.1. Data acquisition

We utilize the data from [10], hereafter referred to as the AAU dataset. This study included 35 participants (14 females and 21 males, aged 21–33) who had no neurological or HIs. Three participants were native English speakers, while the rest were proficient enough to understand English comfortably. Compensation was provided with their consent. Two participants were excluded due to poor data quality, indicated by a high level of artifacts in the raw data [10].

Each participant completed 32 one-minute trials. During each trial, they were exposed to incoming sound streams positioned at ±30° azimuth in front of them. Participants were instructed to focus on one designated sound (the target) and ignore the other (the distractor), with both sounds being either music or speech (see figure 1). At the end of each trial, participants were asked two questions to assess their attention level and the quality of their listening experience.

The experiment was conducted in a single session for each participant. Continuous EEG data was collected at a sampling rate of 512 Hz using a 64-channel g.HIamp-Research system (g.tec Medical Engineering GmbH, Austria). The audio data used in this study had a sampling rate of 48 kHz. The electrodes were positioned on the scalp according to the 10-20 international system, and the impedance of each electrode was maintained below 5 kOhms.

To emulate the physical situation used when making the EEG recordings we used head-related transfer functions (HRTFs) measured on several subjects. The HRTFs dataset is publicly available [44] and consists of acoustic impulse responses measured in a slightly reverberant listening room between a set of loudspeakers and microphones of hearing aids worn by test subjects. This dataset includes speakers placed radially around the participants at 22.5° intervals and contains full impulse responses used to simulate reflections from head, outer ears, and torso. We utilize HRTFs with speakers positioned at ±22.5° since they are the closest match to those used in the AAU dataset (±30°). These HRTFs are convolved with both the target and distractor audio sources to simulate the effects of reverberation, enhancing the realism of the audio data.

Additionally, we use the DTU dataset [45] which contains EEG recordings of 18 Danish subjects that listened to natural speech in Danish spoken by 1 or 2 speakers in three reverberation settings: anechoic, mild reverberation and high reverberation. For consistency we will be using the 2 speaker scenario with the anechoic setting. Each subject has 20 anechoic trials, with each trial lasting 50 seconds. The dataset will be used to both train and evaluate our AAD model.

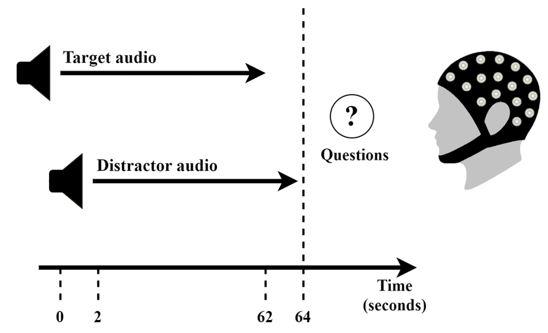

Figure 1. A trial begins with a visual cue that indicates the target source. A 2 s delay is provided to prepare the participant. Participants then listen to two audio streams for 60 s and subsequently answer two follow-up questions.

2.2. Stimuli

The stimuli comprised four categories, categorized into two types (music and speech), each further divided into two genres:

- Piano Music: 8 excerpts of monophonic instrumental pieces played on a piano.
- Electronic Music: 8 excerpts of polyphonic instrumental electronic music.
- Speech Female: 8 excerpts from an English audio book narrated by a woman.
- Speech Male: 8 excerpts from an English audio book narrated by a man.

Participants encountered consecutive blocks of the same target type (e.g. 8 trials of Piano Music followed by 8 trials of Male Speech), with the block order randomized. In each trial, the target and distractor could be music, speech or both. Each excerpt served as a target once, with distractors selected to ensure balanced trials. If both sounds were music, they were of different genres. If both were speech, one had to be male and the other female. The location of the target and distractor (left or right loudspeaker) was randomized for each trial.

2.3. Data pre-processing and splitting

The EEG data underwent pre-processing, initially referencing it to the average of all scalp electrodes. Independent component analysis (ICA) was performed using EEGLAB and an automated detection plugin [46] to remove artifacts related to eye blinking or movement. A low-pass filter was applied to the EEG data at 32 Hz, followed by downsampling to 64 Hz. Previous research indicates that low frequencies (1–8 Hz) [10] works well for envelope reconstruction when using linear models, while for nonlinear models such as deep learning, using a broader frequency range has shown to further improve the performance [31].

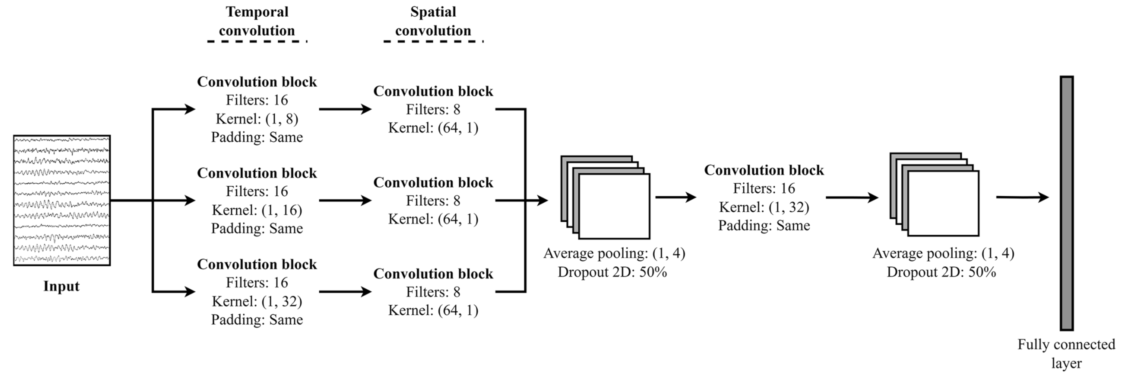

Figure 2. Deep neural network for auditory attention decoding. Each convolution block is followed by batch normalization and tanh activation function.

For each trial, we had access to the normalized target and distractor audio streams. These streams were combined to create the mixed audio source. The envelopes of the mixed audio source, the target audio source, and the distractor audio source were found using the Hilbert transform [47, 48]. The envelopes were low-pass filtered to 32 Hz and then downsampled to 64 Hz to align with the EEG data. Both the EEG and envelopes were normalized to zero mean and unit standard deviation.

After pre-processing, each trial is segmented into time windows of 20 seconds with no overlap between corresponding samples. The data is then split using inter-trial splitting [11, 49, 50] into train, test and validation sets, with trials for both testing and validation drawn from completely unseen trials. To achieve this, out of the 32 trials for each subject, 4 trials are reserved for testing and 4 trials for validation, leaving the remaining 24 trials for training. The 4 trials designated for testing/validation encompass one trial from each of the following four cases:

- Target speech vs. distractor speech.
- Target speech vs. distractor music.
- Target music vs. distractor speech.
- Target music vs. distractor music.

Previous studies have highlighted the importance of proper data splitting [11, 50], as improper splitting can lead to deep learning models training on artifacts specific to each trial, resulting in bloated performance.

2.4. Source envelope separation and auditory attention decoding

For source envelope separation we utilized Conv-Tasnet [6], where typically Conv-Tasnet has only been used for time domain waveform separation, with the exception of a previous study by Tanveer *et al* [43], which showcases how source separation can be done in the envelope domain using Conv-Tasnet. Training was conducted over 100 epochs with a batch size of 32, using the Adam optimizer [51] with default parameters. The best model weights were saved based on performance on the validation data throughout the 100 epochs. The loss function used was the scale-invariant signal-to-noise ratio (SI-SNR) [6]. Permutation invariant training was applied to address the source permutation problem [7].

Following source separation, we perform stimulus reconstruction using the model illustrated in figure 2. This model is inspired by both Inception-net and EEG-net [52, 53]. The input data consists of 64-channel EEG recordings with lengths corresponding to the time window used. The data first passes through three different temporal convolution blocks, followed by spatial convolution blocks. The outputs of the three spatial convolution blocks are concatenated along the filter dimension, then average pooling is applied along the temporal dimension to downsample the data, with dropout used to prevent overfitting. The final block consists of a convolution block, followed by average pooling, dropout, and a fully connected layer to generate envelopes similar in shape to the target envelopes.

The source envelope separation model [6] provides the target and distractor envelopes from the mixed source, and the AAD model is then trained on the target envelopes. The model is trained for 50 epochs with early stopping to prevent overfitting and a batch size of 32. The Adam optimizer with default parameters was used and the loss function was Pearson correlation:

$$\text{Loss} = -\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}}, \tag{1}$$

where x_i and $\bar{x}$ are the EEG envelope samples and mean EEG envelope while y_i and $\bar{y}$ are the target envelope samples and mean target envelope for each batch.

3. Results

3.1. Source envelope separation

We test how well Conv-Tasnet performs using SI-SNR [6] as a measure of performance, as it is also

Table 1. Scale invariant signal to noise ratio (SI-SNR) scores for source envelope separation across different HRTF variants obtained when training and testing are conducted using the same variant of HRTF. 'Original' refers to the original variant of the data, without any influence of HRTFs.

SI-SNR score (dB) for source envelope separation										
Original	HRTF 1	HRTF 2	HRTF 3	HRTF 4	HRTF 5	HRTF 6	HRTF 7	HRTF 8	HRTF 9	HRTF 10
15.9	13.4	14.1	13.6	13.1	13.9	13.7	13.4	13.6	13.3	13.5

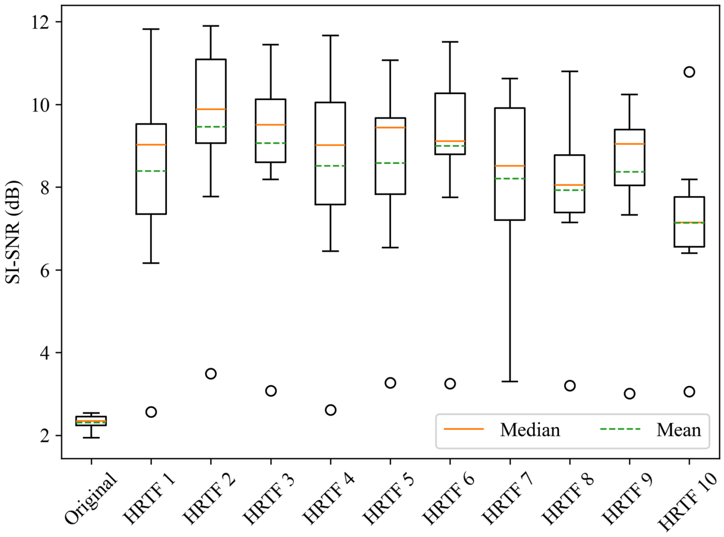

Figure 3. SI-SNR scores obtained by training on each HRTF variant and testing across all HRTFs, showing the models' generalization across different HRTF variants.

used as the loss function for training the model. Additionally, we use 10 randomly selected subjects from the HRTFs [44] to test how well the model performs when exposed to outer ear, head and torso effects.

Table 1 shows the results obtained when the models are trained and tested on the same variant of HRTF (a variant is a particular HRTF from the 10 available). The results indicate that the models generally perform well when trained and tested on the same variant of HRTF. The results for testing across HRTF variants can be found in figure 3, where box plots represent how a model trained on a particular HRTF performs when tested on all other HRTF variants. As we can see, the original model (trained without any HRTF influence) performs very poorly when tested across HRTFs giving an average SI-SNR around 2 dB. Similarly, when checking the other HRTFs' performance, we found that the lowest performing outlier for each HRTF variant is on the original variant (i.e. HRTF variants 1–10, perform worst on the original variant). In general, a model trained on an HRTF from one subject generalizes well across HRTFs from other subjects, showing only slight performance decrements, which is not the case when the model is trained on the AAU dataset (without HRTF influence) and tested on HRTF variants.

It is of interest to understand whether the envelopes obtained by source separation are similar to the envelopes that would be obtained from the original unmixed audio signals. To assess this, we use the strategy shown in figure 4. Here we use the available original target/distractor audio sources, which were combined to make the mixed audio source. We convert this mixed audio source into a mixed envelope, which is then passed through the source envelope separation model to obtain the separated target and distractor envelopes. These separated target and distractor envelopes are then compared with the original target and distractor envelopes, using Pearson correlation. Additionally, we test across the 10 HRTF variants for target and distractor envelopes to see how well the models perform for each case.

Table 2 shows the Pearson correlation score achieved after source envelope separation, between our separated target/distractor envelopes and the original target/distractor envelopes. We pair both targets and distractors to test these conditions: (a) the correlation score should be high for correct pairs (target-target vs. distractor-distractor) and (b) the correlation score should be low for incorrect pairs (target-distractor vs. distractor-target). As we can see, the correlation score shows promising results, giving very high scores when the correct pair is found, which indicates the high accuracy of the given model in source separation. However, we see a comparatively lower score for HRTF 2 and HRTF 4, highlighting that the source envelope separation does not perform that

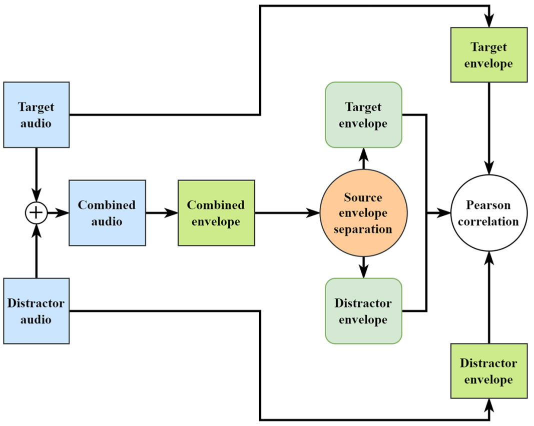

Figure 4. Strategy overview: Relating envelopes from source envelope separation back to their original audio sources, by using Pearson correlation as a measure between separated envelopes and original envelopes.

Table 2. Pearson correlation scores between envelopes from source envelope separation and envelopes from the original sources. We test on 10 different HRTF subjects and the original case without convolving with any HRTF.

Subject	Target\ Target	Target\ Distractor	Distractor\ Distractor	Distractor\ Target
Original	0.96	0.06	0.96	0.06
HRTF 1	0.93	0.09	0.93	0.10
HRTF 2	0.70	0.08	0.69	0.07
HRTF 3	0.93	0.10	0.94	0.09
HRTF 4	0.59	0.07	0.59	0.08
HRTF 5	0.94	0.09	0.94	0.08
HRTF 6	0.93	0.09	0.94	0.09
HRTF 7	0.93	0.10	0.93	0.10
HRTF 8	0.93	0.10	0.93	0.10
HRTF 9	0.92	0.11	0.93	0.11
HRTF 10	0.93	0.09	0.93	0.09

well for these two variants. We visually inspected the time domain signals and the frequency responses of both HRTF 2 and HRTF 4 and compared them to the other HRTF variants. We were not able to observe any noticeable differences. To further investigate this issue, we have also used DPTNet [54] in addition to our Conv-Tasnet model for envelope based source separation. Unfortunately, the DPTNet results indicate weaker overall performance, with notably low correlation scores of 0.33 for certain HRTF subjects (specifically HRTF 3, 4, 5, and 7). The underlying reason for why one model struggles with certain HRTFs while another encounters difficulties with different HRTFs remains unclear.

Our data consists of four cases with speech and music acting as both targets and distractors. To assess our source envelope separation we also wish to test how the different models perform for each of the four cases. Figure 5 shows the SI-SNR scores for the envelope separation of each case across all HRTF variants. Each HRTF variant is tested on its specific test set, with 1 trial for each case coming from each subject in the data. The cases of speech vs. speech and music vs. music show comparable results, while notably music vs. speech seems to outperform speech vs. music.

3.2. Auditory attention decoding

After source envelope separation we wish to test how well our model performs for AAD. The model is trained separately on each HRTF variant (including the original data, with no HRTF influence). We test how well each model performs on its specific HRTF test data. Table 3 highlights these results for both accuracy and Pearson correlation (for both target and distractor). We notice that in most cases the performance does not vary noticeably across HRTF variants. This is in line with existing work demonstrating that the AAD performance does not change significantly across reverberant scenarios (anechoic, reverberant, noisy, and reverberant-noisy) [55]. The original variant outperforms all HRTF variants, with HRTF 2 and HRTF 4 performing the worst out of all the variants due to poor source envelope separation (see table 2). In order to further test this theory, we evaluate our AAD model on the idealized envelopes (see table 3) where for HRTF 2 and HRTF 4 variants we achieve an accuracy of 80.03% and 78.03% respectively. This shows that the lower performance seen in table 3 comes as a consequence of poor source envelope separation. Additionally, table 3 shows comparable performance when envelope separation is used compared

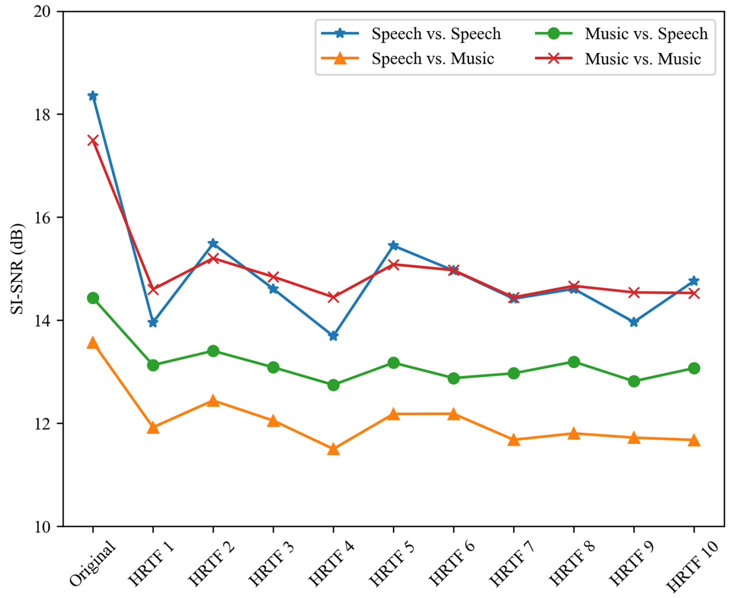

Figure 5. Source envelope separation SI-SNR scores for each of the four cases using both speech and music pairs as both targets/distractors.

to idealized envelopes, highlighting that (excluding HRTF 2 and HRTF 4) de-mixed envelopes perform just as well as original idealized envelopes for AAD.

While the Pearson correlation scores for each HRTF (excluding HRTF 2 and HRTF 4) remain comparable to the original variant, the accuracy fluctuates noticeably. Notably, the Pearson correlation scores under discussion pertain only to the target envelopes.

To further analyze the models we test each variant across all other variants for both accuracy and Pearson correlation. The results can be seen in violin plots shown in figures 6 and 7. The accuracy plot shows that each variant seems to be more evenly spread around the mean which is in contrast to what we see in figure 7, where the Pearson correlation for each variant (excluding the original) seems to be more concentrated around the upper range. We also see that HRTF 2 and HRTF 4 perform poorly in comparison, even across other HRTFs while this is not the case for other HRTFs which generalize well even across HRTF 2 and HRTF 4. This is likely because the envelopes for HRTF 2 and HRTF 4 are poorly separated and do not show strong resemblance to their idealized unmixed counter parts (as shown in table 2), making their respective models less robust and as a result, those models perform poorly across all variants. From the results obtained, it seems that the proposed model for AAD is relatively consistent in its performance across all HRTF variants.

Similarly to SI-SNR, we also wish to test how well our AAD model performs for each of the four pairs consisting of targets/distractors as both speech and music. Figures 8 and 9 show the mean with error bar plot for accuracy and Pearson correlation achieved across all HRTF variants. On average we see HRTF 2 and HRTF 4 perform the worst across all HRTFs, which is similar to the results obtained previously. Additionally, we notice that the music vs. speech scenario has the lowest average accuracy and Pearson correlation compared to the other cases. This is inline with our expectation, as the subjects often claimed that paying attention to music with background speech was the most difficult during data collection. It is worth noting that HRTF 2 and HRTF 4 perform the worst when the target is speech, highlighting the reason that the overall performance drop in these two variants is due to scenarios where the target is speech. This shows that HRTFs in certain cases have a much more notable impact on speech as compared to music. Music vs. music has a higher average accuracy and Pearson correlation score compared to the other three cases, which is interesting when compared to SI-SNR scores in figure 5, where music vs. music shows the lowest performance. This once again indicates that SI-SNR score is not the best indication for what can be achieved by the AAD model.

3.3. Subject independent training and evaluation

One of the key measures for evaluating deep learning models is their ability to generalize to unseen subjects. Our evaluation previously was on unseen trials but subject dependent data. In this section, on the other hand, we will provide subject independent results for two datasets; the AAU dataset [10] and the DTU dataset [45]. The AAD model is trained and then evaluated individually on both datasets. We use

Table 3. Auditory attention decoding results in both accuracy and Pearson correlation for each HRTF variant, including the original variant without any HRTF influence. The table shows Pearson correlation scores for both target/distractor and shows results when using envelope separation with AAD (De-mixed envelopes) and when not using envelope separation, having access to ideal envelopes.

	Accuracy (%) and Pearson correlation for auditory attention decoding										
De-mixed envelopes	Original	HRTF 1	HRTF 2	HRTF 3	HRTF 4	HRTF 5	HRTF 6	HRTF 7	HRTF 8	HRTF 9	HRTF 10
Accuracy	82.4	79.0	64.7	75.5	65.6	77.5	78.3	76.5	77.3	80.1	79.0
(Pearson correlation)$_{target}$	0.122	0.121	0.068	0.112	0.071	0.117	0.116	0.110	0.108	0.116	0.120
(Pearson correlation)$_{distractor}$	0.004	0.001	0.005	0.002	0.001	0.009	0.005	0.007	0.004	0.003	0.008
Idealized envelopes											
Accuracy	83.0	80.1	80.0	82.3	78.1	80.3	78.8	80.5	80.1	81.6	79.5
(Pearson correlation)$_{target}$	0.132	0.128	0.125	0.130	0.124	0.133	0.132	0.129	0.125	0.138	0.126
(Pearson correlation)$_{distractor}$	0.001	0.003	0.001	0.002	0.001	0.002	0.001	0.003	0.004	0.001	0.002

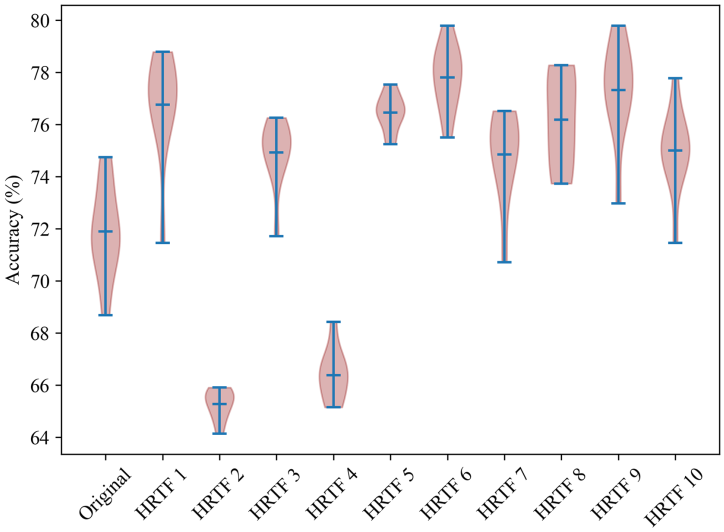

Figure 6. Violin plot, showing accuracy of each HRTF variant's performance across all other variants. The middle line for each violin plot shows the mean while the shape signifies how the values are spread.

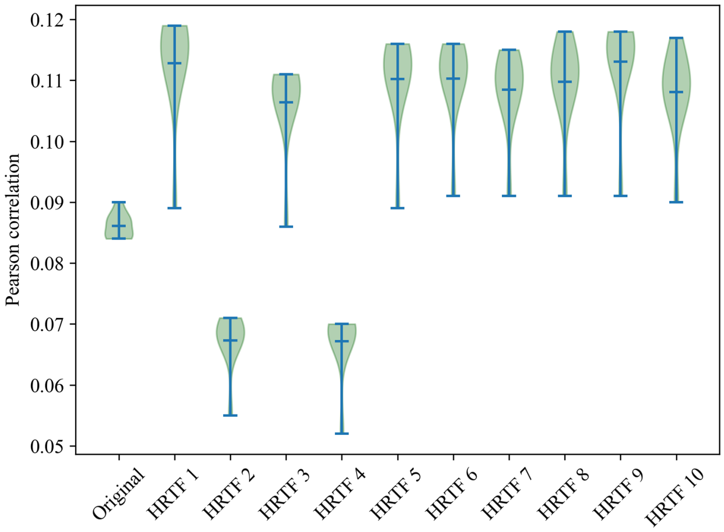

Figure 7. Violin plot, showing Pearson correlation of each HRTF variant and how well the model for each variant generalize across the other HRTF variants.

hold-out method to take out 10 subjects from the AAU dataset and 5 subjects from the DTU dataset, and train the AAD model on the remaining data. The models are then tested on the hold-out subjects. We choose 5 subjects from the DTU dataset for evaluation due to the smaller dataset size and using more hold-out subjects makes it impossible for the deep learning model to learn.

Figure 10 shows the accuracy on the held-out subjects for both AAD models on their respective test datasets. From the figure we can see that the AAD model performs very well on AAU dataset with an average accuracy of 79.5%. However, the model does not seem to perform as well on the DTU dataset with an average accuracy of 66.0%, which is likely due to the much smaller dataset size of the DTU dataset, especially due to only using anechoic trials, making it much harder for the model to learn effectively. Figure 11 shows the Pearson correlation scores for both datasets for both target and distractor. The target Pearson correlation for the AAU dataset once again noticeably outperforms the DTU dataset, with the distractor Pearson correlation scores being much lower for both datasets.

3.4. Source envelope vs. waveform separation

Source envelope separation shows potential due to its lower memory requirement and potential faster processing. In order to compare our models for both source envelope separation and source waveform

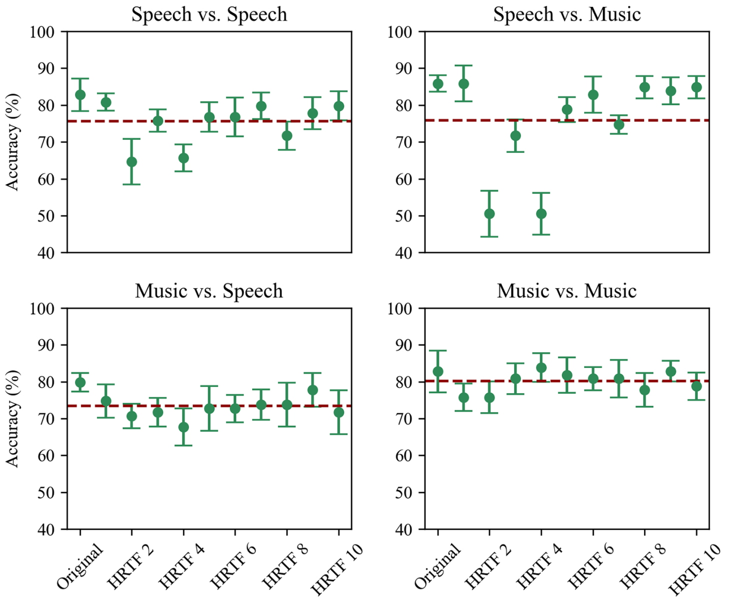

Figure 8. Accuracy error bar plot for each case of music/speech as target/distractor across all HRTF variants. Horizontal dashed line shows the mean across all HRTF variants.

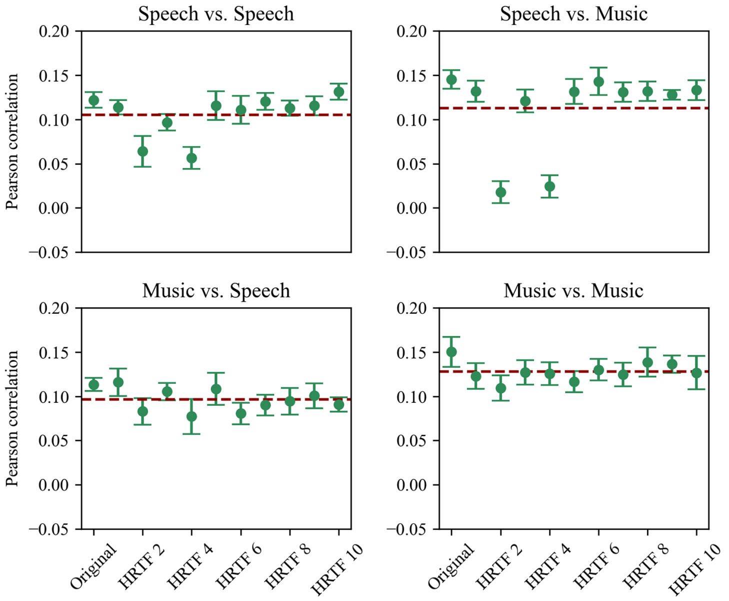

Figure 9. Pearson correlation error bar plot for each case of music/speech as target/distractor across all HRTF variants. Horizontal dashed line shows the mean across all HRTF variants.

separation, followed by AAD, we test our pipeline while only altering the source separation methodology by training Conv-Tasnet on audio data. The audio data is downsampled to 16 kHz from the original 48 kHz. Table 4 shows the AAD performance for both cases. The AAD accuracy and Pearson correlation for source envelope separation outperforms source audio separation. Additionally, we test how much time it takes (on average) for the pipeline to evaluate a single 20 s time window. In order to evaluate this, we test the time taken for evaluation across each sample in the test dataset and average across all times. The source envelope separation scenario evaluates samples approximately 13 times faster than the source audio separation scenario. We also conduct a t-test to determine if the differences in performance for the three given metrics (accuracy, Pearson correlation and time) are statistically significant. The results show that neither accuracy nor Pearson correlation differ significantly. However, the evaluation time is significantly faster for envelope separation compared to waveform separation.

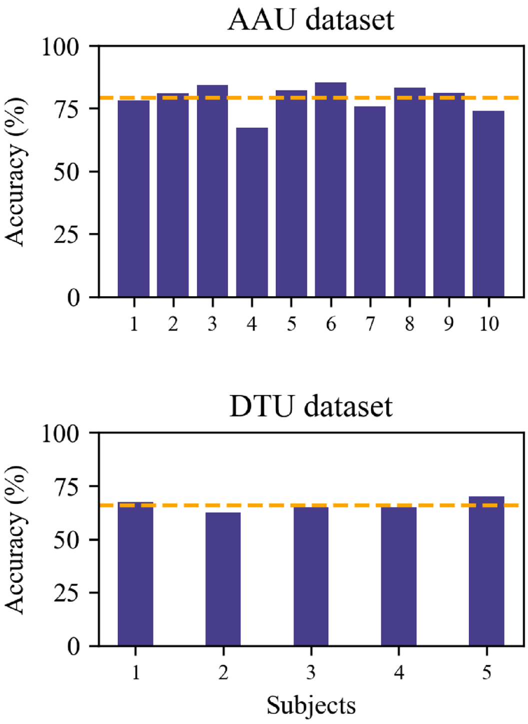

Figure 10. Hold-out subject accuracy for both the AAU dataset and DTU dataset. Both AAD models are trained and evaluated separately for each dataset. The dashed line represents average performance across tested subjects.

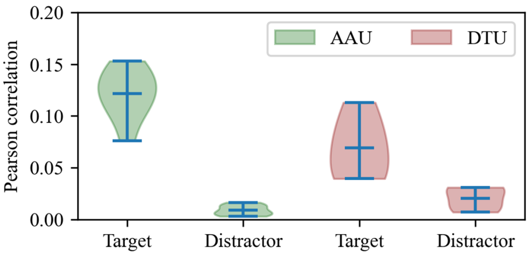

Figure 11. Violin plot with Pearson correlation score for both target and distractor across subjects, for both the AAU dataset and DTU dataset for each AAD model.

Table 4. AAD performance comparison between two source separation methodologies.

	AAD results after source separation		
	Accuracy (%)	Pearson correlation	Time (s)*
Envelope separation	82.4	0.122	0.04
Waveform separation	78.9	0.114	0.53
p-value	0.186	0.337	0.0002

*Average time taken for evaluating 20 s time window segments.

3.5. AAD performance over increasing time windows

In order to evaluate whether higher performance could be achieved at greater time windows, we evaluate our model on the original variant of the dataset at increasing time windows as shown in figure 12. We see that with increasing time windows the final AAD accuracy for the pipeline increases until a saturation level is achieved around 30 s. The results suggest that, for this pipeline and dataset, increasing the time window is unlikely to significantly improve performance which is contrary to typical expectations for correlation-based algorithms. A likely reason is that longer time windows reduce the number of available training samples for the deep learning model.

In order to compare with a baseline we use a linear model in place of the deep learning AAD model, while keeping the source separation model fixed. For the linear model we use a single layer perceptron network, following the same data split for training and

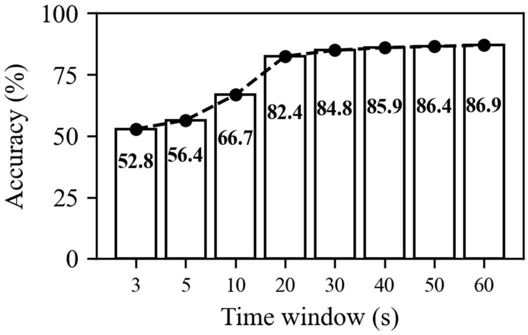

Figure 12. Average accuracy over increasing time window for the given pipeline on the original variant of the dataset.

evaluation. Figure 13 shows the results over increasing time windows for both. From the figure we can see that the deep learning pipeline outperforms the linear model (with source separation) with increasing performance difference over time.

3.6. Comparison to existing linear model

Simon *et al* [10] has worked on the AAU dataset, using a linear model with 60 second time windows to perform stimulus reconstruction and test for AAD. However, their proposed pipeline does not include any source separation but assumes availability of perfectly separated envelopes. Additionally, they train separate models for each of the four target vs. distractor pairs for speech and music as opposed to the proposed study. Figure 14 shows the performance comparison between our single model evaluated on each of the four cases vs. Simon *et al*'s individual model's performance on the four cases, with our pipeline better performing in all four cases. From the figure, we see a similar trend in performance across the four cases, with music vs. speech being the most difficult to classify while speech vs. music is the easiest to classify. Additionally, we note that both our linear model and Simon *et al* [10], do not result in very high performance on the AAU dataset, unlike what is seen commonly in literature where studies show better results when using a linear model for AAD. This suggests that the given dataset, with mixture of speech and music, results in a more challenging environment for linear models to perform.

4. Discussion

In the current study we propose a single microphone system, where we perform source separation directly on the envelopes of mixed audio sources followed by AAD to identify the target source using stimulus reconstruction. The proposed study shows how envelope based source separation is computationally far less expensive than waveform based source separation and provides competing performance when tested with the latter. Furthermore, the study highlights how the pipeline performs when exposed to sources consisting of both speech and music, something lacking in previous literature as generally research on this matter comprises mostly of sources containing only speech.

We evaluate our pipeline across different HRTFs capturing outer ear, head and torso acoustics in a low reverb room to test how well the models generalize across listeners. Based on our results it seems evident that source separation performance seems to vary consistently across the four pairs of speech and music (see figure 5). Additionally, we see that the source envelope separation performance does not noticeably impact AAD performance as seen in figures 6 and 7, with the exception of HRTF 2 and HRTF 4. Furthermore, we see that source envelope separation at times does not represent how well the AAD model may perform as seen from HRTF 2 and HRTF 4 variants' performance, where we see good envelope separation SI-SNR scores (see figure 5) and yet see poor AAD performance (see table 3). This is because although we get good SI-SNR scores, the resulting envelopes do not seem to correlate that well with their idealized counterparts (as seen in table 2).

Table 5 presents a comparison of our proposed deep learning pipeline with previous literature. It is important to recognize that direct comparisons with other literature is challenging due to variations in datasets, experimental design, and individual subject characteristics. While many studies use speech as the target and distractor, our experimental design adds complexity by incorporating sources containing both speech and music. This enhances realism but also increases difficulty. However, in order to make it easier to compare the performance to existing work, we include results where we evaluate our given pipeline for only speech as target/distractor and also for an idealized scenario where we assume access to perfectly separated envelopes without the use of source separation, as done by many previous studies.

Among the literature, one previous study has used envelope based source separation with AAD.

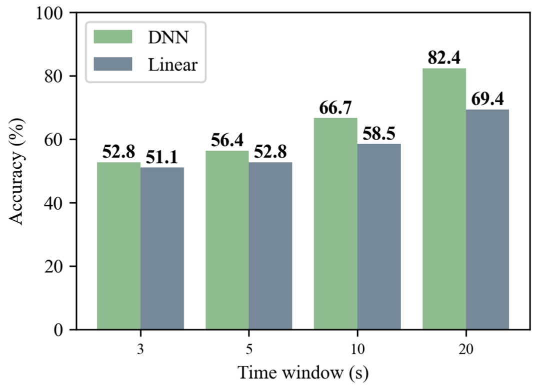

Figure 13. Average accuracy over increasing time window for the given pipeline using deep learning model for AAD vs. linear model for AAD.

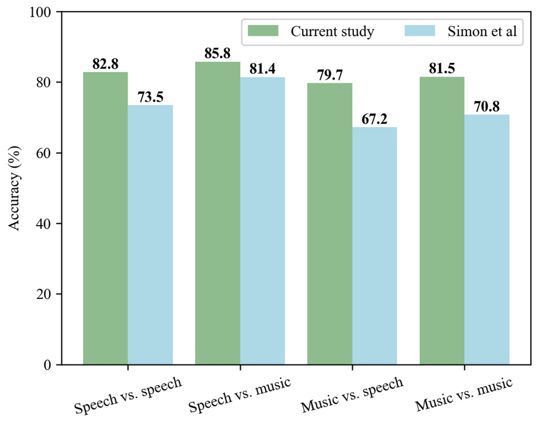

Figure 14. Performance comparison between current study and Simon *et al* [10] for given dataset, across four cases of target/distractor pairs for speech and music.

Table 5. Comparison of envelope based source separation with auditory attention decoding between our study and previous literature, including linear regression (**LR**), deep learning (**DL**), multichannel wiener filter (**MWF**), single-microphone (**SMP**) system and multi-microphone (**MMP**) system. **Idealized**: scenario where no source separation is used, assuming access to perfectly separated envelopes.

Study	Source separation	Method	Microphone system	Dataset	Time window (s)	Accuracy (%)
Simon *et al* [10]	Idealized	LR	—	Speech and music	60	73.2
Ciccarelli *et al* [30]	Waveform	DL	SMP	Speech	10	62.0
Nguyen *et al* [34]	Idealized	DL	—	Speech	20	78.0
Mirkovic *et al* [56]	Idealized	LR	—	Speech	60	88.0
Fuglsang *et al* [57]	Idealized	LR	—	Speech	40-50	80.0-90.0
Eyndhoven *et al* [58]	Envelope	MWF	MMP	Speech	30	83.0
Han *et al* [59]	Waveform	DL	SMP	Speech	16	95.0
Das *et al* [60]	Waveform	DL with MWF	MMP	Speech	30	75.0-83.0
O'Sullivan *et al* [61]	Waveform	DL	SMP	Speech	20	73.0
Yan *et al* [62]	Idealized	DL	—	Speech	60	77.5
Current study	Envelope	DL	SMP	Speech and music	20	82.4
	Envelope	DL	SMP	Speech	20	82.8
	Idealized	DL	—	Speech and music	20	83.0
	Idealized	DL	—	Speech	20	83.3
	Envelope	LR	SMP	Speech and music	20	69.4
	Idealized	LR	—	Speech and music	20	70.6

Eynhoven *et al* [58], propose the use of multiplicative non-negative ICA for envelope based source separation followed by MWF for AAD. Their pipeline uses 30 s time windows with a multi-microphone system contrary to the proposed methodology which uses a single microphone system using deep learning for both envelope separation and AAD showing comparative performance, despite a 10 s smaller time window. Ciccarelli *et al* [30] and Das *et al* [60] propose audio based source separation in their AAD pipeline, with the latter producing comparable results to the proposed methodology, although at larger time windows. Han *et al* [59] propose an end-to-end, single-microphone system for waveform separation and AAD using deep learning. The system operates with a 16 s time window and achieves around 95% accuracy. Contrary to our study the authors do not use envelope based stimulus reconstruction but rather use frequency spectrum based stimulus reconstruction, training the AAD model to generate the frequency spectrum of the audio sources. Nguyen *et al* [34] using idealized envelopes (without source separation) output 78.0% accuracy for their deep learning based AAD pipeline, our pipeline outperforms them with an accuracy of 83.3% when only considering speech sources. Yan *et al* [62] proposes a deep learning AAD model in a four speaker scenario, using a 1–60 s time windows. Their model results in an accuracy of 77.5% for 60 s time windows, it is worth noting that a four speaker scenario is a more complex scenario than the proposed two speaker scenario. Fuglsang *et al* [57] introduce a linear regression model in their AAD pipeline. Using time windows of 40–50 s, they are able to achieve a certain degree of robustness towards various reverberant scenarios, ranging from mildly to highly reverberant cases. Our results support these findings by demonstrating that robustness can also be achieved under the influence of HRTFs. It is important to emphasize the inherent difficulty in making direct comparisons with existing literature, primarily due to variations in datasets, model architectures, and evaluation protocols. To address this, we have made an effort to at least approximately obtain comparable experimental conditions by focusing on specific scenarios commonly reported in prior work, such as speech-only target stimuli, and by evaluating performance across varying temporal windows.

5. Conclusion

In this paper we propose a deep learning based source envelope separation and AAD using a single microphone system. The source envelope separation model allows for faster computation and lower memory requirements due to lower sampling rate of envelopes as compared to audio signals. The dataset used in the study consists of both speech and music pairs for targets and distractors. The study looks into how well the models perform on sources consisting of both speech and music and how well the models generalize variants of the AAU dataset that include the effect of HRTFs. The results highlight that the models trained on 20 s time windows generalize well across different HRTFs, while showing that scenarios consisting of target as music and distractor as speech result in the lowest performance as such scenarios are most difficult for subjects to pay attention to the target speech. Additionally, we note that for certain HRTFs, the performance when the target is speech is more negatively impacted than when the target is music. A relevant topic for future work is explaining the performance disparity among different HRTF cases by evaluating with different pipelines, containing different source separation and AAD models.

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

Funding

This work was supported by William Demant Foundation, case no. 20-2428.

Ethical statement

The datasets used in this study are sourced from previously conducted research [10, 45]. The research was conducted in accordance with the principles embodied in the Declaration of Helsinki and in accordance with local statutory requirements.

ORCID iDs

M Asjid Tanveer https://orcid.org/0000-0002-7934-5786
Jesper Jensen https://orcid.org/0000-0003-1478-622X

References

[1] Cherry E C 1953 Some experiments on the recognition of speech, with one and with two ears *J. Acoust. Soc. Am.* **25** 975–9

[2] Pan Z, Wichern G, Germain F G, Khurana S and Le Roux J 2024 Neuroheed+: Improving neuro-steered speaker extraction with joint auditory attention detection *ICASSP 2024-2024 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 11456–60

[3] Wächtler M, Kessler J, Walger M and Meister H 2022 Revealing perceptional and cognitive mechanisms in static

and dynamic cocktail party listening by means of error analyses *Trends Hearing* **26** 23312165221111676

[4] Hershey J R, Chen Z, Le Roux J and Watanabe S 2016 Deep clustering: discriminative embeddings for segmentation and separation *2016 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 31–35

[5] Luo Y, Chen Z and Yoshioka T 2020 Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation *ICASSP 2020-2020 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 46–50

[6] Luo Y and Mesgarani N 2019 Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation *IEEE/ACM Trans. Audio Speech Lang. Process.* **27** 1256–66

[7] Kolbaek M, Yu D, Tan Z-H and Jensen J 2017 Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks *IEEE/ACM Trans. Audio Speech Lang. Process.* **25** 1901–13

[8] Wang D and Chen J 2018 Supervised speech separation based on deep learning: an overview *IEEE/ACM Trans. Audio Speech Lang. Process.* **26** 1702–26

[9] Kellermann W 2008 Beamforming for speech and audio signals *Handbook of Signal Processing in Acoustics* (Springer) pp 691–702

[10] Simon A, Loquet G, Østergaard J and Bech S 2023 Cortical auditory attention decoding during music and speech listening *IEEE Trans. Neural Syst. Rehab. Eng.* **31** 2903–11

[11] Tanveer M A, Skoglund M A, Bernhardsson B and Alickovic E 2024 Deep learning-based auditory attention decoding in listeners with hearing impairment *J. Neural Eng.* **21** 036022

[12] Pu W, Zan P, Xiao J, Zhang T and Luo Z-Q 2020 Evaluation of joint auditory attention decoding and adaptive binaural beamforming approach for hearing devices with attention switching *ICASSP 2020-2020 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 8728–32

[13] Ceolini E, Hjortkjær J, Wong D D E, O'Sullivan J, Raghavan V S, Herrero J, Mehta A D, Liu S-C and Mesgarani N 2020 Brain-informed speech separation (biss) for enhancement of target speaker in multitalker speech perception *NeuroImage* **223** 117282

[14] O'sullivan J A, Power A J, Mesgarani N, Rajaram S, Foxe J J, Shinn-Cunningham B G, Slaney M, Shamma S A and Lalor E C 2015 Attentional selection in a cocktail party environment can be decoded from single-trial eeg *Cereb. Cortex.* **25** 1697–706

[15] Kurmanavičiūtė D, Kataja H, Jas M, Välilä A and Parkkonen L 2023 Target of selective auditory attention can be robustly followed with meg *Sci. Rep.* **13** 10959

[16] Crosse M J, Di Liberto G M, Bednar A and Lalor E C 2016 The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli *Front. Hum. Neurosci.* **10** 604

[17] Chen X, Du C, Zhou Q and He H 2023 Auditory attention decoding with task-related multi-view contrastive learning *Proc. 31st ACM Int. Conf. on Multimedia* pp 6025–33

[18] Mahjoory K, Bahmer A and Henry M J 2023 Convolutional neural networks can identify brain interactions involved in decoding spatial auditory attention *bioRxiv* 2023–11

[19] Golumbic E Z, Cogan G B, Schroeder C E and Poeppel D 2013 Visual input enhances selective speech envelope tracking in auditory cortex at a cocktail party *J. Neurosci.* **33** 1417–26

[20] Lunner T, Alickovic E, Graversen C, Ng E H N, Wendt D and Keidser G 2020 Three new outcome measures that tap into cognitive processes required for real-life communication *Ear Hearing* **41** 39

[21] Geirnaert S, Vandecappelle S, Alickovic E, de Cheveigne A, Lalor E, Meyer B T, Miran S, Francart T and Bertrand A 2021 Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices *IEEE Signal Process. Mag.* **38** 89–102

[22] Alickovic E, Lunner T, Wendt D, Fiedler L, Hietkamp R, Ng E H N and Graversen C 2020 Neural representation enhanced for speech and reduced for background noise with a hearing aid noise reduction scheme during a selective attention task *Front. Neurosci.* **14** 846

[23] Alickovic E, Ng E H N, Fiedler L, Santurette S, Innes-Brown H and Graversen C 2021 Effects of hearing aid noise reduction on early and late cortical representations of competing talkers in noise *Front. Neurosci.* **15** 636060

[24] Soroush P Z, Angrick M, Shih J, Schultz T and Krusienski D J 2021 Speech activity detection from stereotactic EEG *2021 IEEE Int. Conf. on Systems, Man and Cybernetics (SMC)* (IEEE) pp 3402–7

[25] Dash D, Ferrari P, Dutta S and Wang J 2020 Neurovad: real-time voice activity detection from non-invasive neuromagnetic signals *Sensors* **20** 2248

[26] Alickovic E, Lunner T, Gustafsson F and Ljung L 2019 A tutorial on auditory attention identification methods *Front. Neurosci.* **13** 153

[27] Lu Y *et al* 2021 Auditory attention decoding from electroencephalography based on long short-term memory networks *Biomed. Signal Process. Control* **70** 102966

[28] Fu Z, Wang B, Wu X and Chen J 2021 Auditory attention decoding from eeg using convolutional recurrent neural network *2021 29th European Signal Processing Conf. (EUSIPCO)* (IEEE) pp 970–4

[29] Li Y, Anumanchipalli G K, Mohamed A, Chen P, Carney L H, Lu J, Wu J and Chang E F 2023 Dissecting neural computations in the human auditory pathway using deep neural networks for speech *Nat. Neurosci.* **26** 2213–25

[30] Ciccarelli G, Nolan M, Perricone J, Calamia P T, Haro S, O'sullivan J, Mesgarani N, Quatieri T F and Smalt C J 2019 Comparison of two-talker attention decoding from eeg with nonlinear neural networks and linear methods *Sci. Rep.* **9** 11538

[31] De Taillez T, Kollmeier B and Meyer B T 2020 Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech *Eur. J. Neurosci.* **51** 1234–41

[32] Xu Z, Bai Y, Zhao R, Zheng Q, Ni G and Ming D 2022 Auditory attention decoding from eeg-based mandarin speech envelope reconstruction *Hear. Res.* **422** 108552

[33] Kuruvila I, Muncke J, Fischer E and Hoppe U 2021 Extracting the auditory attention in a dual-speaker scenario from eeg using a joint cnn-lstm model *Front. Physiol.* **12** 700655

[34] Nguyen N D T, Phan H, Geirnaert S, Mikkelsen K, and Kidmose P 2024 Aadnet: an end-to-end deep learning model for auditory attention decoding (arXiv:2410.13059)

[35] Thornton M, Mandic D and Reichenbach T 2022 Robust decoding of the speech envelope from eeg recordings through deep neural networks *J. Neural Eng.* **19** 046007

[36] Hausfeld L, Riecke L, Valente G and Formisano E 2018 Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes *NeuroImage* **181** 617–26

[37] Zuk N J, Murphy J W, Reilly R B, Lalor E C and Theunissen F E 2021 Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies *PLoS Comput. Biol.* **17** e1009358

[38] Di Liberto G M, Marion G and Shamma S A 2021 Accurate decoding of imagined and heard melodies *Front. Neurosci.* **15** 673401

[39] Di Liberto G M, Pelofi C, Shamma S and de Cheveigné A 2020 Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening *Acoust. Sci. Technol.* **41** 361–4

[40] Ding N and Simon J Z 2012 Emergence of neural encoding of auditory objects while listening to competing speakers *Proc. Natl Acad. Sci.* **109** 11854–59

[41] Pras A and Guastavino C 2010 Sampling Rate Discrimination: 44.1 KHZ VS. 88.2 KHZ *Audio Engineering Society Convention 128* (Audio Engineering Society)

[42] Sedgwick P 2012 Pearson's correlation coefficient *BMJ* **345** e4483

[43] Tanveer M A, Jensen J, Tan Z-H and Østergaard J 2024 Envelope based deep source separation and EEG auditory attention decoding for speech and music *2024 32nd European Signal Processing Conf. (EUSIPCO)* pp 872–6

[44] Moore A H, de Haan J M, Pedersen M S, Naylor P A, Brookes M and Jensen J 2019 Personalized signal-independent beamforming for binaural hearing aids *J. Acoust. Soc. Am.* **145** 2971–81

[45] Fuglsang S A, Wong D and Hjortkjær J 2018 EEG and audio dataset for auditory attention decoding *Zenodo*

[46] Pion-Tonachini L, Kreutz-Delgado K and Makeig S 2019 Iclabel: an automated electroencephalographic independent component classifier, dataset and website *NeuroImage* **198** 181–97

[47] Abdulmunem M E and Badr A A 2017 Hilbert transform and its applications: a survey *Int. J. Sci. Eng. Res* **8** 699–704

[48] Kak S 1970 The discrete hilbert transform *Proc. IEEE* **58** 585–6

[49] Saha S and Baumert M 2020 Intra-and inter-subject variability in eeg-based sensorimotor brain computer interface: a review *Front. Comput. Neurosci.* **13** 87

[50] Puffay C, Accou B, Bollens L, Monesi M J, Vanthornhout J, Van hamme H and Francart T 2023 Relating eeg to continuous speech using deep neural networks: a review *J. Neural Eng.* **20** 041003

[51] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980)

[52] Lawhern V J, Solon A J, Waytowich N R, Gordon S M, Hung C P and Lance B J 2018 EEGnet: a compact convolutional neural network for EEG-based brain–computer interfaces *J. Neural Eng.* **15** 056013

[53] Szegedy C, Ioffe S, Vanhoucke V and Alemi A 2017 Inception-v4, inception-resnet and the impact of residual connections on learning *Proc. AAAI Conf. on Artificial Intelligence* vol 31

[54] Chen J, Mao Q, and Liu D 2020 Dual-path transformer network: direct context-aware modeling for end-to-end monaural speech separation (arXiv:2007.13975)

[55] Aroudi A, Mirkovic B, De Vos M and Doclo S 2019 Impact of different acoustic components on EEG-based auditory attention decoding in noisy and reverberant conditions *IEEE Trans. Neural Syst. Rehabil. Eng.* **27** 652–63

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

[57] Fuglsang S A, Dau T and Hjortkjær J 2017 Noise-robust cortical tracking of attended speech in real-world acoustic scenes *NeuroImage* **156** 435–44

[58] Van Eyndhoven S, Francart T and Bertrand A 2016 EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses *IEEE Trans. Biomed. Eng.* **64** 1045–56

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

[60] Das N, Zegers J, Van hamme H, Francart T and Bertrand A 2020 Linear versus deep learning methods for noisy speech separation for EEG-informed attention decoding *J. Neural Eng.* **17** 046039

[61] O'Sullivan J, Chen Z, Herrero J, McKhann G M, Sheth S A, Mehta A D and Mesgarani N 2017 Neural decoding of attentional selection in multi-speaker environments without access to clean sources *J. Neural Eng.* **14** 056001

[62] Yan Y, Xu X, Zhu H, Tian P, Ge Z, Wu X and Chen J 2024 Auditory attention decoding in four-talker environment with EEG *Proc. Annual Conf. Int. Speech Communication Association (INTERSPEECH)* pp 432–6