MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

NeuroHeed+: Improving Neuro-steered Speaker Extraction with Joint Auditory Attention Detection

Pan, Zexu; Wichern, Gordon; Germain, François G; Khurana, Sameer; Le Roux, Jonathan

TR2024-025 March 19, 2024

Abstract

Neuro-steered speaker extraction aims to extract the listener's brain- attended speech signal from a multi-talker speech signal, in which the attention is derived from the cortical activity. This activity is usu- ally recorded using electroencephalography (EEG) devices. Though promising, current methods often have a high speaker confusion error, where the interfering speaker is extracted instead of the attended speaker, degrading the listening experience. In this work, we aim to reduce the speaker confusion error in the neuro-steered speaker extraction model through a jointly fine-tuned auxiliary auditory attention detection model. The latter reinforces the consistency between the extracted target speech signal and the EEG representation, and also improves the EEG representation. Experimental results show that the proposed network significantly outperforms the baseline in terms of speaker confusion and overall signal quality in two-talker scenarios.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2024

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

NEUROHEED+: IMPROVING NEURO-STEERED SPEAKER EXTRACTION WITH JOINT AUDITORY ATTENTION DETECTION

Zexu Pan, Gordon Wichern, François G. Germain, Sameer Khurana, Jonathan Le Roux

Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

ABSTRACT

Neuro-steered speaker extraction aims to extract the listener's brain-attended speech signal from a multi-talker speech signal, in which the attention is derived from the cortical activity. This activity is usually recorded using electroencephalography (EEG) devices. Though promising, current methods often have a high speaker confusion error, where the interfering speaker is extracted instead of the attended speaker, degrading the listening experience. In this work, we aim to reduce the speaker confusion error in the neuro-steered speaker extraction model through a jointly fine-tuned auxiliary auditory attention detection model. The latter reinforces the consistency between the extracted target speech signal and the EEG representation, and also improves the EEG representation. Experimental results show that the proposed network significantly outperforms the baseline in terms of speaker confusion and overall signal quality in two-talker scenarios.

***Index Terms*—** Cocktail party problem, auditory attention, speaker extraction, EEG, multi-modal

1. INTRODUCTION

The human brain has a remarkable ability to focus its auditory attention on a particular stimulus, such as a target speech signal, while filtering out other stimuli, such as interfering speech signals, noise, and reverberation, a phenomenon also known as the "cocktail party effect" [1, 2]. Mimicking such ability in machines, speech separation and speaker extraction algorithms have transformed the development of hearing aids [3] and served as important front-ends for many speech processing tasks including speech recognition [4], speaker localization [5], and speech emotion recognition [6].

Speech separation algorithms separate a multi-talker speech signal into individual clean streams [7–9], reaching exceptional performance when the number of speakers is known. However, the separated speech signals have no association with a listener's attention. An additional algorithm is required to detect which of the separated signals is desired, using auxiliary references such as visual signals [10] or brain signals [11–15]. The performance of such cascaded systems may be limited as each algorithm is optimized independently.

Speaker extraction algorithms are better suited to emulating the "cocktail party effect", as they unify the separation and detection steps into a single network to extract only the speech from a target speaker. They typically use an auxiliary reference as prior knowledge to distinguish the target speaker from interfering speakers. The most widely studied auxiliary reference is a pre-recorded speech signal [16–18], however, it can be cumbersome to pre-record many people's voices and select the right one to use for speaker extraction. Other auxiliary references include visual recordings such as face [19–22], gesture [10], direction information [23, 24] or other unique characteristics of the target speaker [25], with the limitation that it is not always feasible for the listener to visually track the target speaker, or to reliably obtain direction information.

While selecting the right auxiliary reference to use for speaker extraction is hard, an alternative way is to directly model the listener's attention through the neuronal response of their cortical activity, which reflects an interaction of external stimuli with spontaneous patterns produced endogenously [26, 27]. Among various cortical activity recording devices, electroencephalography (EEG) stands out as it is non-intrusive and cost-efficient with high temporal resolution. Auditory attention detection (AAD) studies [28] show that EEG signal could be used to select one of the stimuli that the listener is focusing on in a multi-talker scenario with impressive accuracy. The stimuli studied in AAD are often the clean signals [29, 30] or the separated signals from the speech separation algorithm [14, 15].

The findings in AAD studies enabled neuro-steered speaker extraction, in which the auxiliary reference is the spontaneous neuronal activity, usually recorded using EEG devices. The brain-informed speech separation (BISS) model [31] utilizes the reconstructed speech envelope from the EEG signal as the auxiliary reference. The U-shaped brain-enhanced speech denoiser (UBESD) model [32] and the brain-assisted speech enhancement network (BASEN) [33] go a step further and directly model the EEG signal with a temporal convolutional neural network and fuse the EEG signal with feature-wise linear modulation or convolutional multi-layer cross-attention. The neuro-steered speaker extraction (NeuroHeed) [34] is the current state-of-the-art (SOTA) model. It adopts self-attention for the EEG encoder and proposes an online auto-regressive speaker self-enrollment strategy to reinforce the speaker cue.

Although promising, the correlation of the attended speech with the EEG signal is not as strong as compared to using the corresponding lip recording or a reference speech signal as the auxiliary reference cue. Therefore, NeuroHeed often extracts the interfering speaker instead of the attended one, which is referred to as a speaker confusion error. It creates a negative impact, especially in hearing aids, when the listener is forced to listen to the wrong speaker and cannot switch the attention back. In this work, we aim to improve NeuroHeed by reducing the speaker confusion error.

We draw inspiration from recent AAD studies [29, 30], which are very successful in EEG-speech association. We propose NeuroHeed+, which jointly optimizes the SOTA NeuroHeed model with an auxiliary AAD model. The AAD model maps the EEG signal and the separated speech stimuli in a common feature space, and pushes the EEG and target stimulus representations together while pulling the EEG and interfering stimulus representations away. This reinforces the consistency of the extracted target speech signal to the EEG representation, as well as improves the EEG representation. Experimental results show that the proposed network significantly outperforms NeuroHeed on speaker confusion as well as overall signal quality in two-talker scenarios on the KUL dataset [35].

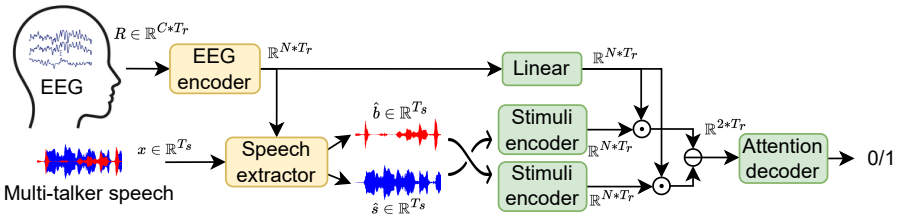


Fig. 1. Our proposed NeuroHeed+ model, which jointly optimizes the speaker extraction model NeuroHeed [34] (in yellow), and an auxiliary auditory attention detection model (in green). The symbol $\ominus$ represents the concatenation of embeddings along the channel dimension, while the symbol $\odot$ refers to the inner product.

2. RELATED WORK: NEUROHEED

Denote a multi-talker discrete-time speech signal as x, that consists of the sum of the target speech signal s and the interfering speech signal b:

$$x = s + b \ \in \mathbb{R}^{T_s} \tag{1}$$

where T_s is the length of the speech signal. The SOTA neuro-steered speaker extraction model NeuroHeed [34] extracts an estimate of s denoted $\hat{s}$, from x, conditioned on the EEG signal $R \in \mathbb{R}^{C*T_r}$ as the reference cue, where C is the number of channels and T_r is the length of the EEG signal. The different lengths T_s and T_r reflect the fact that the sampling rate is different between EEG and audio signals, but they are recorded over the same duration.

The NeuroHeed model is presented on the left panel of Fig. 1, and consists of an EEG encoder and a speech extractor. The original NeuroHeed only extracts $\hat{s}$, without estimating the interfering signal $\hat{b}$. NeuroHeed adopts self-attention layers [30] for its EEG encoder and a time-domain dual-path recurrent neural network (DPRNN) [8] for its speech extractor.

Due to the non-intrusive nature of EEG devices in capturing brain signals, the correlation of the attended speech with the EEG signal is usually weak. Generally, the reconstructed speech envelope from the EEG signal and the actual stimuli have a Pearson's correlation of less than 0.3 [31]. Therefore, to reduce speaker confusion errors in the NeuroHeed model, one needs to improve both a) the capability of an EEG encoder to extract a more discriminative EEG representation, and b) the ability of the speech extractor network to correlate the EEG representation to the target speech signal in the mixture speech signal.

3. PROPOSED MODEL: NEUROHEED+

Our proposed NeuroHeed+ model is depicted in Fig. 1, and is an extension of the NeuroHeed model. We propose a multi-task learning framework to jointly perform speaker extraction and auditory attention detection. The speaker extraction network estimates both the target speech signal and the interfering speech signal, while the AAD model recognizes which of the speech signals correlates to the EEG signal. The two networks are separately trained first, and then cascaded and jointly fine-tuned. The AAD network is only used to better train the EEG encoder and the speech extractor, and is discarded at inference time.

3.1. Speaker extraction

The speaker extraction model is shown on the left panel of Fig. 1, and consists of an EEG encoder and a speech extractor. We modify the original NeuroHeed to estimate both the target speech signal $\hat{s}$ and the interfering speech signal $\hat{b}$, such that the cascaded AAD model has access to both.

Different from speech separation models where permutation invariant training [36, 37] is needed, we train the model to always estimate the target signal at a fixed output stream taking advantage of the auxiliary EEG signal. The EEG encoder and the speech extractor are optimized end-to-end using the scale-invariant signal-to-noise ratio (SI-SDR) [38] loss function:

$$\mathcal{L}_{\text{SE}} = \frac{1}{2}(\mathcal{L}_{\text{SI-SDR}}(s, \hat{s}) + \mathcal{L}_{\text{SI-SDR}}(b, \hat{b})), \tag{2}$$

where

$$\mathcal{L}_{\text{SI-SDR}}(s, \hat{s}) = -20 \log_{10} \frac{\left\| \frac{<\hat{s}, s>}{\|s\|^2} s \right\|}{\left\| \hat{s} - \frac{<\hat{s}, s>}{\|s\|^2} s \right\|}. \tag{3}$$

3.2. Auditory attention detection

The AAD model is also shown in Fig. 1. It consists of an EEG encoder that is shared with the NeuroHeed model; a linear layer for EEG representation adaptation; a stimuli encoder that is formed by a time-domain speech encoder [8] with a suitable stride size making the speech representation match the time resolution of the EEG representation, a positional encoding layer, and 5 layers of the self-attention network; and a back-end attention decoder that is formed by two convolutional layers, one adaptive average pooling layer [29, 30], and a sigmoid layer.

The AAD model is first trained to discriminate which of the two clean speech signals s or b is the target speech that corresponds to the EEG signal. This is done by mapping the EEG signal and the two speech signals into a common feature space, and performing a dot-product operation between the EEG feature and the two speech features to obtain the respective similarity scores. The similarity scores are concatenated and passed through the attention decoder to reach a decision.

We minimize the following binary cross-entropy loss for AAD model training:

$$\mathcal{L}_{\text{AAD}} = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}), \tag{4}$$

where $y \in \{0, 1\}$ indicates which of the speech signals corresponds to the EEG signal, while $\hat{y}$ is the predicted probability.

It is worth noting that the EEG encoder is initialized from the trained NeuroHeed model and fixed during AAD model training so that it does not become inconsistent with what NeuroHeed expects. The EEG encoder is later unfrozen during the joint fine-tuning of NeuroHeed and the AAD model, so that the fine-tuning can improve the EEG representation for speaker extraction.

3.3. Joint optimization

After the NeuroHeed and AAD models are individually trained, they are cascaded and jointly fine-tuned as shown in Fig. 1. Instead of feeding s and b to the AAD model, we feed the shuffled separated signals $\hat{s}$ and $\hat{b}$ to the AAD model and train it to match $\hat{s}$ with the

EEG signal as opposed to $\hat{b}$. The fine-tuning loss is defined as:

$$\mathcal{L}_{\text{fine-tune}} = \mathcal{L}_{\text{SE}} + \alpha \cdot \mathcal{L}_{\text{AAD}} \quad (5)$$

where α is a scalar weight balancing the two tasks.

The speaker extraction loss $\mathcal{L}_{\text{SE}}$ constrains the speech quality of the extracted signals. The AAD loss $\mathcal{L}_{\text{AAD}}$ backpropagates through the EEG encoder to improve the EEG representation, and also through the speaker extractor to extract discriminative speech signals from the EEG representation, improving the consistency between the separated signal $\hat{s}$ and the EEG representation.

4. EXPERIMENTAL SETUP

4.1. Database

Following [34], we examine our proposed NeuroHeed+ model on the publicly available KULeuven (KUL) dataset [35]. There are 16 normal hearing subjects, with 20 trials per subject, in which the subjects listen to concurrent speech with plugged-in earphones, and one speech signal is played in each ear. The subjects are instructed to listen to the speech in one ear while ignoring the speech in the other ear. We used the first 8 trials for each subject, where they attend to a given signal for the first time. The speech signals are from 4 Dutch stories spoken by 2 male speakers. The speech signals for our speaker extraction models are sampled at 8 kHz. The raw EEG signal is recorded using the BioSemi ActiveTwo system at 8192 Hz with $C = 64$ channels.

As in the experimental evaluation of NeuroHeed [34], we evaluate our models under a known-subject and known-speaker scenario. To do so, we randomly split each trial into training, validation, and test sets with a ratio of 75%, 12.5%, and 12.5%, respectively, without overlap of speech stimulus between sets. For the training set, we use mixture signal augmentation, in which the target stimulus is mixed with the interfering stimulus at a random signal-to-noise (SNR) ratio between -10 dB and 10 dB. There are $400,000$, $3,000$, and $3,000$ utterances for training, validation, and test sets.

4.2. Hyper-parameter settings

The hyper-parameters of the speaker extraction model exactly follow NeuroHeed [34], with N set to 64. For the AAD model stimuli encoder, the time-domain speech encoder has a 1D convolutional layer (Conv1D) with input size 1, output size $N * 2$, kernel size 120, and stride size 60, followed by a rectified linear activation (ReLU), a layer normalization, and a linear layer with input size $N * 2$ and output size N. The self-attention layers in the stimuli encoder have input size N, feedforward size $N*4$, number of heads 1, and dropout 0.1. For the AAD model attention decoder, the first Conv1D has input and output sizes 2, kernel size 15, stride 7, while the second Conv1D has input size 2 and output size 1, kernel size 15, stride 7, and a parametric ReLU is used between the two convolutional layers.

4.3. Training details

We use PyTorch to conduct our experiments. All models are trained on 2 GPUs with 48 GB RAM each. The Adam optimizer is used with a learning rate (lr) warm-up as follows for the first $15,000$ training steps:

$$\text{lr}(n) = 0.1 \cdot N^{-0.5} \cdot n \cdot 15,000^{-1.5} \quad (6)$$

where n is the step number. After the warm-up is done, the lr is halved when the best validation loss (BVL) does not improve within 6 consecutive epochs. The training stops when the BVL does not improve for 10 subsequent epochs. For the joint fine-tuning, the model weights are initialized from the previously trained weights, but the optimizer and lr scheduler are re-initialized. We share a common model between subjects, which has experimentally shown to have better performance than subject-specific models.

Table 1. Validation set results for NeuroHeed+ with various configurations. We find our best model according to the reported SI-SDR value in dB. We give a system number (Sys. #) to every different system. α is the scalar weight for the fine-tuning loss. We initialize (Init) and fix different modules during the joint training stage, such as the EEG encoder (EE), speaker extractor (SE), and the green modules in the AAD module (AAD) in Fig. 1.

Sys. #	α	Init SE&EE	Init AAD	Fix AAD	Fix EE	Fix SE	SI-SDR
0 [34]	-	-	-	-	-	-	13.4
1	0.0	-	-	-	-	-	12.3
2	1.0	-	-	-	-	-	13.2
3	1.0	✓	-	-	-	-	14.0
4	1.0	✓	✓	-	-	-	**14.3**
5	1.0	✓	✓	✓	-	-	14.2
6	1.0	✓	✓	✓	✓	-	14.0
7	1.0	✓	✓	✓	-	✓	13.7

4.4. Evaluation metrics

We use the improvement in SI-SDR (SI-SDRi) and the improvement in signal-to-distortion ratio (SDRi) to evaluate the signal quality, while using the improvement in perceptual evaluation of speech quality (PESQi) and the improvement in short term objective intelligibility (STOIi) to evaluate the perceptual quality and intelligibility of the extracted speech with respect to the unprocessed multi-talker speech signals. To evaluate the speaker confusion error, we report the percentage positive rate (PPR) [34], which is defined as the percentage of extracted speech signals that satisfy both i) a positive SI-SDRi value and ii) a higher SI-SDRi value with respect to the target speech signal than to the interfering speech signal. The higher the better for all metrics.

5. RESULTS

We first present in Sec. 5.1 the results on the validation set of our proposed NeuroHeed+ with various training strategies, to select our best model. We then compare our proposed NeuroHeed+ with various baselines on the test set in Sec 5.2, to show the superiority of our model.

5.1. Model tuning

In Table 1, Sys. 0 is the original NeuroHeed model [34], which obtains 13.4 dB. Sys. 1 is the modified NeuroHeed model that estimates both $\hat{s}$ and $\hat{b}$, without joint training. The SI-SDR drops by 1.1 dB from Sys. 0 to Sys. 1, this is probably because the network has limited capacity, thus the performance drops when estimating both signals. Sys. 2 is the NeuroHeed+ model with all modules trained from scratch. The joint learning improves the SI-SDR by 0.9 dB compared with Sys. 1, but is still not better than Sys. 0.

We next explore various initialization and fine-tuning approaches. In Sys. 3, the Speaker extractor and EEG encoder are initialized from Sys. 1 and fine-tuned with Eq. (5), improving the SI-SDR to 14.0 dB. In Sys. 4, the AAD modules are also initialized from pre-training before fine-tuning the whole system, and the SI-SDR further improves to 14.3 dB. In Sys. 5, the AAD modules are fixed, and we fine-tune the EEG encoder and speaker extractor, obtaining a similar SI-SDR to Sys. 4. In Sys. 6, we further fix the EEG encoder and only fine-tune the speaker extractor, leading to

Table 2. Validation set results for NeuroHeed+ with various scalar weight α for the fine-tuning loss. We find our best model according to the reported SI-SDR value in dB.

Sys. #	α	SI-SDR
8	0.001	13.9
9	0.01	14.1
10	0.1	14.1
4	1	**14.3**
11	10	13.9
12	100	13.2

a degradation of the SI-SDR by 0.3 dB compared to Sys. 4. In other words, the 0.8 dB gain from Sys. 2 to Sys. 6 represents the improvements in the speaker extractor's ability brought by the auxiliary AAD task, to correlate the EEG representations with the target speaker during the extraction process. In Sys. 7, we only fine-tune the EEG encoder, and the SI-SDR degrades by 0.6 dB from our best model Sys. 4. In other words, the 0.5 dB gain from Sys. 2 to Sys. 7 represents the improvements in the EEG representation brought by the auxiliary AAD task.

We note that the proposed auxiliary AAD task may be sub-optimal during joint training, as it leaves the opportunity for the speech extractor to learn to output $\hat{s}$ with some EEG information encoded implicitly and for the stimuli encoder to learn to decode that EEG information. However, we still see improvements in the speaker extraction task, which could be because the SI-SDR loss $\mathcal{L}_{\text{SE}}$ constrains $\hat{s}$ and regularizes the speaker extractor training. In addition, the initialization enables the modules to start from a better state instead of quickly converging to a sub-optimal solution.

In Table 2, we present validation set results for the NeuroHeed+ model with various α, using the same initialization and fine-tuning strategy as Sys. 4. We can see that the best α value is 1. Therefore, we select Sys. 4 as our final best NeuroHeed+ model.

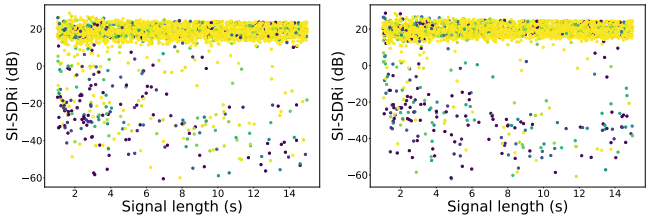
5.2. Comparison with baselines

In Table 3, we compare the test set results of NeuroHeed+ with those of various baselines. Sys. 13 is an oracle system for upper-bound analysis, which performs speech separation using DPRNN [8] first, and then performs ground-truth association by selecting the separated speech signal that has the highest SI-SDR with s. Therefore, it has a PPR of 100%, and a very high SI-SDRi value of 19.4 dB. Sys. 14 is a baseline that performs speech separation using DPRNN first, and then performs the association with a separately trained AAD network. Sys. 15 has the same pipeline as Sys. 14 except that the DPRNN is fine-tuned together with the AAD network. We can see that our proposed NeuroHeed+ model (Sys. 4) outperforms by a large margin in terms of all metrics both Sys. 14 and 15, which perform speech separation without EEG input instead of target speech extraction. The proposed NeuroHeed+ also outperforms NeuroHeed by 1.3 dB in SI-SDRi, 1.2 dB in SDRi, 0.13 in PESQi, 0.02 in STOIi, and 1.5% in PPR (a 16% relative error reduction).

Fig. 2 presents the scatter plot of SI-SDRi of the extracted speech signals for signal lengths ranging from 1 s to 15 s. For both (a) NeuroHeed and (b) NeuroHeed+, the majority of samples have SI-SDRi around 20 dB, meaning that the models extract the correct target speaker with high signal quality. As the signal length increases, both models have fewer samples having negative SI-SDRi values, meaning that the model is able to learn from the longer context when a longer EEG signal is available. Overall, NeuroHeed+ has fewer low SI-SDRi samples compared with NeuroHeed, meaning that NeuroHeed+ makes fewer speaker confusion errors, which explains the average 1.3 dB SI-SDRi gain shown in Table 3.

Table 3. Test set results for EEG-steered speaker extraction models. SI-SDRi and SDRi are reported in dB. The percentage positive rate (PPR) is the percentage of the extracted speech in the test set that has both a positive SI-SDRi value and a higher SI-SDRi value with respect to the attended speech than to the interfering speech. The higher the PPR, the lower the speaker confusion error.

Sys. #	Model	SI-SDRi	SDRi	PESQi	STOIi	PPR
13	Separation-PIT [8]	19.4	19.6	1.22	0.23	100.0
14	Separation-AAD	4.6	9.2	0.64	0.02	74.7
15	+ jointly fine-tuned	12.9	15.2	0.95	0.14	88.5
16	BISS [31]	–0.1	0.5	–0.08	–0.03	59.4
17	UBESD [32]	5.1	5.8	0.09	0.03	80.9
18	BASEN [33]	5.6	6.7	0.22	0.03	75.6
0	NeuroHeed [34]	14.3	15.5	0.95	0.16	90.8
4	**NeuroHeed+**	**15.6**	**16.7**	**1.08**	**0.18**	**92.3**



(a) Sys. 0: NeuroHeed (b) Sys. 4: NeuroHeed+

Fig. 2. SI-SDRi scatter plot of extracted speech signal for various lengths of audio signals in the test set, by (a) the NeuroHeed model [34], and (b) our proposed NeuroHeed+ model. The color represents the AAD probability of making a correct attention detection, with bright yellow meaning correct and dark purple incorrect.

The samples in Fig. 2 are plotted with colors, which represent the AAD output probability score of correctly associating the EEG representation with the clean target signal s instead of the interfering signal b, with bright yellow meaning correct and dark purple incorrect. Because the color is associated with how correlated the EEG representation is to s, it gives an indication as to how easy speaker extraction is. However, the speech extractor uses the EEG representation as conditioning to extract part of a mixture speech signal, so they may not always agree, with one succeeding in selecting the correct speaker while the other fails. Our proposed joint training aims to promote agreement between the speaker extraction model and the AAD model, meaning in particular that for samples where AAD makes accurate classification, the speaker extraction model will make fewer speaker confusion errors. As shown in Fig. 2, NeuroHeed+ indeed has fewer low SI-SDRi samples with a bright yellow color compared with NeuroHeed, meaning that our proposed joint training is able to improve the speaker extraction model in agreeing with the AAD model on those samples.

6. CONCLUSION

In this work, we reduce the speaker confusion error for the SOTA neuro-steered speaker extraction model NeuroHeed. We propose NeuroHeed+, which has a joint learning framework such that the speaker extraction model benefits from the auxiliary AAD task in improving the EEG representation, and improving the EEG-speech association in the speaker extraction processes. Experimental results show that the proposed NeuroHeed+ is effective in extracting the correct speakers and achieving a new SOTA on the KUL dataset.

7. REFERENCES

[1] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acust. united Acust.*, vol. 86, no. 1, pp. 117–128, 2000.

[2] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, 1953.

[3] D. Wang, "Deep learning reinvents the hearing aid," *IEEE Spectr.*, vol. 54, no. 3, pp. 32–37, 2017.

[4] J. Wang, X. Qian, and H. Li, "Predict-and-Update network: Audio-visual speech recognition inspired by human speech perception," *arXiv preprint arXiv:2209.01768*, 2022.

[5] X. Qian, M. Madhavi, Z. Pan, J. Wang, and H. Li, "Multi-target DoA estimation with an audio-visual fusion mechanism," in *Proc. ICASSP*, 2021.

[6] Z. Pan, Z. Luo, J. Yang, and H. Li, "Multi-modal attention for speech emotion recognition," in *Proc. Interspeech*, 2020.

[7] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016.

[8] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. ICASSP*, 2020.

[9] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Making time-frequency domain models great again for monaural speaker separation," in *Proc. ICASSP*, 2023.

[10] Z. Pan, X. Qian, and H. Li, "Speaker extraction with co-speech gestures cue," *IEEE Signal Process. Lett.*, vol. 29, pp. 1467–1471, 2022.

[11] M. Geravanchizadeh and S. Zakeri, "Ear-EEG-based binaural speech enhancement (ee-BSE) using auditory attention detection and audiometric characteristics of hearing-impaired subjects," *J. Neural Eng.*, vol. 18, no. 4, pp. 0460d6, 2021.

[12] J. O'Sullivan, Z. Chen, J. Herrero, G. M. McKhann, S. A. Sheth, A. D. Mehta, and N. Mesgarani, "Neural decoding of attentional selection in multi-speaker environments without access to clean sources," *J. Neural Eng.*, vol. 14, no. 5, pp. 056001, 2017.

[13] C. Han, J. O'Sullivan, Y. Luo, J. Herrero, A. D. Mehta, and N. Mesgarani, "Speaker-independent auditory attention decoding without access to clean speech sources," *Sci. Adv.*, vol. 5, no. 5, pp. eaav6134, 2019.

[14] 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," *Cereb. Cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[15] S. Van Eyndhoven, T. Francart, and A. Bertrand, "EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 5, pp. 1045–1056, 2016.

[16] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. Interspeech*, 2019.

[17] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černocký, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 4, pp. 800–814, 2019.

[18] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "SpEx+: A complete time domain speaker extraction network," in *Proc. Interspeech*, 2020.

[19] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–11, 2018.

[20] J. Wu, Y. Xu, S. Zhang, L. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," in *Proc. ASRU*, 2019.

[21] Z. Pan, R. Tao, C. Xu, and H. Li, "Selective listening by synchronizing speech with lips," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1650–1664, 2022.

[22] Z. Pan, M. Ge, and H. Li, "USEV: Universal speaker extraction with visual cue," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 3032–3045, 2022.

[23] M. Elminshawi, S. R. Chetupalli, and E. A. Habets, "Beamformer-guided target speaker extraction," in *Proc. ICASSP*, 2023.

[24] K. Tesch and T. Gerkmann, "Spatially selective deep non-linear filters for speaker extraction," in *Proc. ICASSP*, 2023.

[25] E. Tzinis, G. Wichern, A. Subramanian, P. Smaragdis, and J. Le Roux, "Heterogeneous target speech separation," in *Proc. Interspeech*, Sept. 2022.

[26] D. L. Ringach, "Spontaneous and driven cortical activity: implications for computation," *Curr. Opin. Neurobiol.*, vol. 19, no. 4, pp. 439–444, 2009.

[27] K. D. Harris and A. Thiele, "Cortical state and attention," *Nat. Rev. Neurosci.*, vol. 12, no. 9, pp. 509–523, 2011.

[28] J. Belo, M. Clerc, and D. Schön, "EEG-Based Auditory Attention Detection and Its Possible Future Applications for Passive BCI," *Front. Comput. Sci.*, vol. 3, pp. 661178, 2021.

[29] S. Cai, P. Li, E. Su, and L. Xie, "Auditory attention detection via cross-modal attention," *Front. Neurosci.*, vol. 15, pp. 652058, 2021.

[30] M. Borsdorf, S. Pahuja, G. Ivucic, S. Cai, H. Li, and T. Schultz, "Multi-head attention and GRU for improved match-mismatch classification of speech stimulus and EEG response," in *Proc. ICASSP*, 2023.

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

[32] M. Hosseini, L. Celotti, and E. Plourde, "End-to-end brain-driven speech enhancement in multi-talker conditions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1718–1733, 2022.

[33] J. Zhang, Q.-T. Xu, Q.-S. Zhu, and Z.-H. Ling, "BASEN: Time-domain brain-assisted speech enhancement network with convolutional cross attention in multi-talker conditions," in *Proc. Interspeech*, 2023.

[34] Z. Pan, M. Borsdorf, S. Cai, T. Schultz, and H. Li, "NeuroHeed: Neuro-steered speaker extraction using EEG signals," *arXiv preprint arXiv:2307.14303*, 2023.

[35] N. Das, T. Francart, and A. Bertrand, "Auditory attention detection dataset KULeuven," *Zenodo*, 2019.

[36] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, 2016.

[37] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, 2017.

[38] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. ICASSP*, 2019.