



NeuroSpex+: Dual-Task Training of Neuro-Guided Speaker Extraction with Speech Envelope and Waveform

Dashanka De Silva¹, Siqi Cai^{*,3}, Saurav Pahuja¹, Tanja Schultz², Haizhou Li^{1,4}

¹Machine Listening Lab (MLL), University of Bremen, Germany

²Cognitive Systems Lab (CSL), University of Bremen, Germany

³Department of ECE, National University of Singapore, Singapore

⁴School of Data Science, The Chinese University of Hong Kong, Shenzhen, China

ddesilva@uni-bremen.de, caisqi@ieee.org

Abstract

Neuro-guided speaker extraction, i.e. NeuroSpex, aims to isolate the speech signal a listener is attending to in a multi-talker environment using reference cues derived from cortical activity, such as electroencephalography (EEG). Despite remarkable progress, there remains untapped potential. In this study, we propose *NeuroSpex+*, a novel neuro-guided speaker extraction model that integrates an additional task of reconstructing the target speech envelope. By simultaneously optimizing the model for both the target speech envelope and speech waveform, *NeuroSpex+* reinforces the mask generation for speaker extraction. Experimental results demonstrate that the proposed model significantly outperforms baselines, improving overall signal quality.

Index Terms: Speaker extraction, EEG, selective auditory attention, speech envelope

1. Introduction

Humans have an innate ability to focus on a single sound source while ignoring multiple competing noises, a phenomenon known as selective auditory attention (SAA), particularly in challenging environments like cocktail party scenarios [1]. However, traditional machine listening systems struggle to realize this function, and individuals with hearing impairments often experience diminished effectiveness with conventional hearing aids [2].

Target speaker extraction (TSE) [3, 4] isolates a speaker's voice using an auxiliary reference cue that uniquely identifies the target or attended speech. Various reference cues have been explored for guiding speaker extraction, including pre-enrolled speech samples of the target speaker [3], visual cues such as hand and body gestures [5], lip movements [6], and multi-modal approaches that combine audio-visual inputs. However, in real-world scenarios, these cues may not always be accessible or reliable. In contrast, neural responses provide a robust and inherently available reference, making EEG-based TSE a viable and effective solution [7, 8].

Neuroscience studies have demonstrated a strong correlation between attended speech and the elicited neural responses [9, 10], enabling auditory attention detection (AAD) from brain activity [11, 12]. TSE mimics human SAA by isolating the target speech using an auxiliary reference cue [3]. This reference cue is crucial for separating the target speech signal and has led to the development of neuro-guided speaker extraction [2], offering promising avenues for brain-inspired hearing aids for those with hearing impairments.

Recent advancements in EEG-based AAD and TSE have enabled the extraction of attended speech using EEG signals as a reference in multi-speaker environments. Among them, pio-

neering work BISS [13] utilizes the reconstructed speech envelope derived from EEG signals as a reference cue to facilitate TSE in complex auditory environments. In contrast, NeuroHeed [7], and NeuroSpex [8], have used neural response-based reference signals, known as EEG embeddings, to guide the extraction process.

Despite these advances, existing methods have not fully explored the potential of exploiting temporal synchronization between the speech stimulus and the elicited EEG response. Research on SAA has demonstrated that the speech envelope can be decoded from the EEG response [14]. Inspired by this, we propose *NeuroSpex+*, an end-to-end neuro-guided speaker extraction model, in which both the target speech waveform and target speech envelope are used as the reference cue during training. We hypothesize that this dual-task approach will improve the synchronization of EEG embeddings and enhance extraction mask generation, resulting in superior performance in speaker extraction. In contrast to NeuroSpex, *NeuroSpex+* is designed to emphasize the critical role of temporal synchronization between neural responses and attended speech. To achieve this, *NeuroSpex+* introduces an innovative dual-task optimization strategy, which simultaneously facilitates speech envelope reconstruction and enhances the EEG encoder's ability to capture and represent neural activity effectively.

The remainder of this paper is organized as follows: Section II presents the proposed *NeuroSpex+*. Section III describes the dataset, experimental setup, evaluation metrics, and baselines. Section IV discusses the results, followed by an analysis of the findings. Finally, Section V concludes the paper.

2. Methodology

We introduce a dual-task model, i.e. *NeuroSpex+*, designed to perform both speech envelope reconstruction and speaker extraction, as depicted in Figure 1. Unlike NeuroSpex [8] that only uses target speech waveform as the training target, *NeuroSpex+* features a novel EEG encoder that is trained to reconstruct the target speech envelope, thus is expected to enforce the target reference with temporal information embedded in the envelope. In brief, the *NeuroSpex+* architecture consists of two primary branches: (1) EEG-to-speech envelope reconstruction; and (2) speaker extraction. The model is trained end-to-end using an integrated loss function that jointly optimizes both tasks.

2.1. Speech Envelope as Training Target of EEG Encoder

The EEG-to-speech envelope reconstruction branch is implemented with the novel EEG encoder and speech envelope decoder. The EEG encoder processes the 64-channel EEG signal to extract temporally synchronized information correlated with

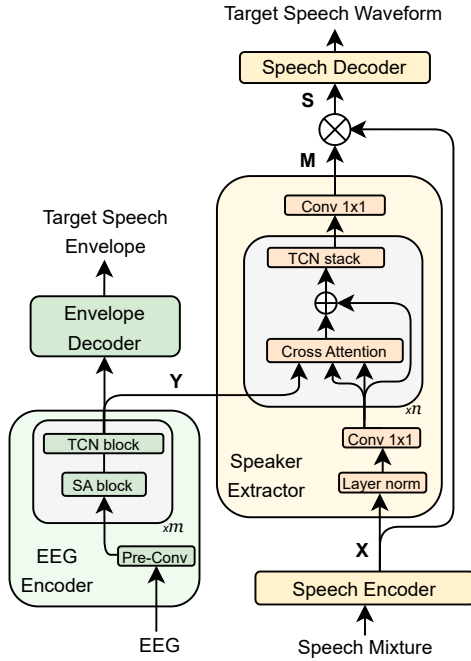


Figure 1: The proposed NeuroSpex+ consists of two main branches: (left) the EEG-to-speech envelope reconstruction branch, which includes an EEG encoder and an envelope decoder; and (right) the speaker extraction branch, which comprises a speech encoder, a speaker extractor, and a speech decoder. NeuroSpex+ takes a speech mixture and its corresponding EEG signals as inputs and produces the target speech waveform as output. \otimes refers to the element-wise multiplication.

the target speech, generating EEG embeddings Y that serve as reference signals for the speaker extractor. The speech envelope decoder then employs these embeddings Y to reconstruct the target speech envelope. Specifically, the EEG encoder begins with a pre-convolution layer, followed by a stack of Self-Attention (SA) [15] and Temporal Convolution Network (TCN) [16] block pairs. The SA mechanism is employed for its ability to capture temporal dependencies in sequential EEG data [17]. The TCN blocks further enhance the model by expanding the receptive field through hierarchical convolutions, enabling it to capture broader temporal contexts and aggregate information across multiple points in the EEG time series data [18]. The EEG encoder exploits temporal dependencies and contextual information inherent in the EEG data, which are crucial for accurately reconstructing the target speech envelope.

In the EEG encoder, as shown in Figure 1, the pre-convolution layer with a kernel size of 1×3 performs initial feature extraction, followed by m pairs of SA and TCN blocks, where $m = 4$, to effectively capture the temporal dynamics of attended speech from EEG features. Each SA block in the encoder includes a Self-Attention layer, residual connection, and layer normalization. As depicted in Figure 2 (a), the TCN block comprises the following layers: a convolution layer with 1×1 kernel to match the input EEG feature dimension, a dilated depth-wise convolution layer with a kernel size of 1×8 and a dilation rate of 2, and a point-wise convolution layer to reduce the number of output channels. Each convolution layer in the TCN block, except for the final point-wise layer, employs

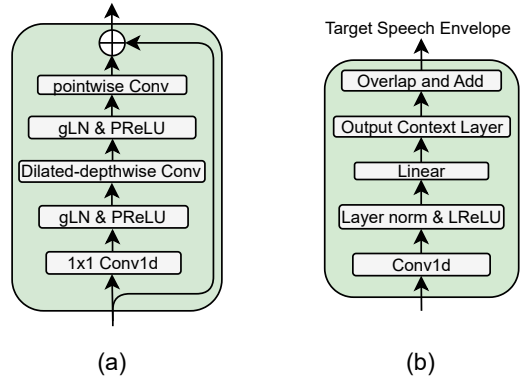


Figure 2: (a) The TCN block of the EEG encoder. Note that the speaker extractor also uses the same TCN block in its TCN stacks. gLN , $PReLU$, and $Conv$ refer to global Layer Normalization, Parametric Rectified Linear units, and convolution operations, respectively. (b) The envelope decoder is designed to reconstruct the speech envelope from EEG embeddings. The $LReLU$ denotes Leaky Rectified Linear units.

Parametric Rectified Linear units ($PReLU$) activation followed by global Layer Normalization (gLN). The dilation in TCNs facilitates the efficient processing of variable-length sequences, enhancing both the robustness and accuracy of the speech envelope reconstruction. The output of the point-wise convolution layer is then combined with the input to the TCN block through a skip connection to preserve low-level details and prevent information loss.

The envelope decoder, as shown in Figure 2 (b), inspired by the VLAAI architecture [14], transforms EEG embeddings Y into target speech envelopes. The envelope decoder features a convolution layer with a 1×8 kernel, Leaky Rectified Linear units ($LReLU$) activation, and layer normalization, followed by a linear layer and an output context layer, collectively referred to as the VLAAI block. Finally, an Overlap and Add layer converts the decoded envelope embeddings into the final envelope waveform [19].

2.2. Speech Waveform as Training Target of Speaker Extraction

The speaker extraction branch is adopted from NeuroSpex [8], incorporating the speech encoder, speaker extractor, and speech decoder components, as illustrated in Figure 1.

The speech encoder converts the single-channel mixture signal into a sequence of utterance-based temporal embeddings, \mathbf{X} , which are synchronized with EEG embeddings. This process is similar to short-time Fourier transform analysis and uses a time-domain speech encoding approach [19] with a 1×1 convolutional kernel and ReLU activation.

The speaker extractor generates the estimation mask, \mathbf{M} , to separate the target speech from the speech mixture. It takes in the speech mixture embeddings from the speech encoder and the reference signal, i.e., EEG embeddings, from the EEG encoder. These inputs are fused through a cross-attention block with a residual connection, where the reference signal serves as the query and the speech mixture embeddings act as the key and value inputs. Following this, a series of TCN blocks are applied to model long-range temporal dependencies and context [8, 19], referred to as the TCN stack. Each stack comprises 4 TCN blocks, and this configuration of cross-attention

and TCN stack pairs is repeated 4 times (i.e., $n = 4$).

The masked speech embeddings, \mathbf{S} , are derived by element-wise multiplication of the generated mask with the speech mixture embeddings, expressed as $\mathbf{S} = \mathbf{M} \otimes \mathbf{X}$. To reconstruct the time-domain single-channel speech waveform, the speech decoder performs the inverse operation of the speech encoder. This process involves passing the masked speech embeddings \mathbf{S} through a linear transformation, followed by an overlap-and-add operation to regenerate the audio signal.

2.3. Dual-Task Loss Function and Optimization

NeuroSpex+ was trained end-to-end using a dual-task loss function that combines the scale-invariant signal-to-distortion ratio (SI-SDR) [20] for the primary task of speaker extraction and the Pearson correlation coefficient (PCC) for the sub-task of speech envelope reconstruction. The SI-SDR is a widely used metric in audio signal processing that measures the distortion of a reconstructed signal compared to the original, independent of the scale of the reconstructed signal [20]. The PCC measures the linear correlation between two signals, ranging from -1 to 1, where 1 and -1 indicate perfect positive and negative correlations, respectively, and 0 indicates no correlation [14]. SI-SDR and PCC are commonly applied in time-domain speaker extraction [7, 19] and speech envelope reconstruction [14], respectively. The dual-task loss function is defined as,

$$\mathcal{L} = \mathcal{L}_{SI-SDR} + \alpha \times \mathcal{L}_{PCC} \quad (1)$$

where α is a scaling parameter for weighting the PCC loss, empirically set to 0.6. SI-SDR is computed in decibels (dB), while PCC is a scalar value. Higher SI-SDR and PCC indicate better speech and reconstruction quality, respectively. Therefore, both negative SI-SDR and PCC are used in the loss function.

3. Experiment

3.1. Dataset

We conducted the experiments on the KULeuven dataset [21], which includes EEG recordings from 16 subjects collected with a 64-channel BioSemi ActiveTwo system at a sampling rate of 8,192 Hz. Each subject participated in 8 trials, listening dichotically to Dutch stories narrated by two male speakers, with attention directed toward one speaker. The speech recordings were sampled at 8 kHz, with the attended and unattended speech mixed at 0 dB. EEG was referenced to the average of all electrodes, filtered between 1 and 32 Hz, and down-sampled to 128 Hz. The dataset provides 128 trials across all subjects, totaling 12.8 hours of speech-EEG data, standardized by trial.

3.2. Subject Independent Setup

We employed a subject-independent cross-validation approach to evaluate our proposed model and all baselines. The dataset was divided into three subsets: training, validation, and test. In each cross-validation fold, one subject’s 8 trials were used as the test set, another subject’s 8 trials as the validation set, and the remaining 14 subjects’ trials, totaling 14×8 , as the training set. This process was repeated across 16 folds for 16 subjects, ensuring each subject was used as the test set once. All models were trained on each fold, and the average results were reported. Each trial, with a duration of 360 seconds, was empirically segmented into 4-second windows with a hop length of 1 second for all subsets, resulting in 2,856 segments for both the test and

validation sets and 39,984 for training. Models are trained on the training set, with performance monitored on the validation set after each epoch to adjust hyperparameters.

All implementations are performed using PyTorch on 4 Nvidia RTX A6000 GPUs with random seeds for reproducibility. Models are trained end-to-end with the Adam optimizer (initial learning rate of 0.0001). A learning rate scheduler (0.5 decay) was applied if validation loss stagnated for 5 epochs, and early stopping was employed after 25 epochs without improvement. Training runs for around 100 epochs or until convergence, with a batch size of 16, Xavier initialization for weights, and gradient clipping during training.

3.3. Evaluation Metrics and Baselines

In line with previous studies [8, 7], we use the following evaluation metrics: SDR improvement (SDRi), SI-SDR improvement (SI-SDRi), Perceptual Evaluation of Speech Quality (PESQ), and Short-Term Objective Intelligibility (STOI) for the speaker extraction, and PCC for the envelope reconstruction. SDRi [20] and SI-SDRi measure the improvement in extracted speech quality compared to the mixture signal. PESQ [22] measures the quality and naturalness of extracted speech, STOI [23] assesses intelligibility, and PCC quantifies the similarity between the reconstructed and original speech envelopes. Higher values for all metrics indicate better performance.

We benchmark NeuroSpex+ against three baselines: NeuroSpex [8], the latest model upon which this paper is built; NeuroHeed [7], a previous work that also performs TSE using EEG neural responses as the sole reference cue; BISS [13], a seminal work that uses a reconstructed speech envelope as the reference cue. Compared to BISS, the NeuroSpex+ employs a dual-task strategy where an EEG encoder generates the cue while concurrently reconstructing the speech envelope to capture EEG-attended speech dynamics.

To evaluate the performance of envelope reconstruction, we compared our EEG-to-speech envelope reconstruction branch separately with several baselines: a linear decoder [11], VLAAl [14], and FCNN and CNN by [24]. The VLAAl is a convolution-based architecture that excels in decoding speech envelopes from EEG signals by accurately mapping neural responses. The envelope decoder in our proposed model incorporates components from the VLAAl architecture. FCNN and CNN baselines are employed as feed-forward dense and convolution-based networks, respectively. All models under the ablation study were trained on the same dataset in a subject-independent manner but tested on the validation set.

4. Results

We evaluate the proposed NeuroSpex+ through subject-independent cross-validation, comparisons with baseline models, and an ablation study. The statistical significance of results was assessed using paired t -tests.

Figure 3 depicts the SI-SDRi obtained by our proposed NeuroSpex+ for each subject in the KULeuven dataset. The average SI-SDRi score across all subjects is 17.083 dB, with a standard deviation of 2.597. The violin plots show a uniform distribution of SI-SDRi across subjects, indicating consistent performance with no significant outliers. This suggests that our model exhibits robust performance and generalizability across different subjects.

We compare our NeuroSpex+ with several baseline models, as summarized in Table 1. Note that NeuroSpex+*, which ex-

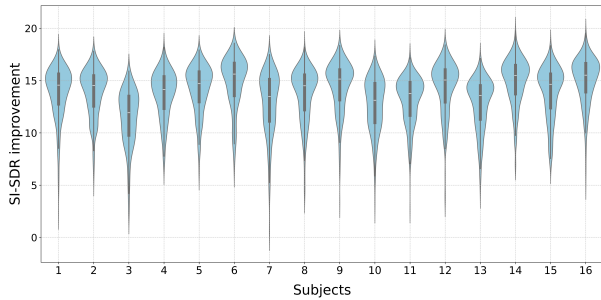


Figure 3: Violin plots of SI-SDR improvement (SI-SDRi) of the extracted speech for each subject using NeuroSpex+.

Table 1: Performance comparison of NeuroSpex+ with its competitive baselines on the test set in terms of SI-SDR (dB) and SI-SDRi (dB) for speech quality, PESQ and STOI for intelligibility. NeuroSpex+* denotes a NeuroSpex implementation using the architecture in Figure 1 without the speech envelope as a training target. ‘*’ and ‘**’ in the SDRi and SI-SDRi columns denote statistically significant differences at $p < 0.01$ and $p < 0.001$, respectively, when compared to NeuroSpex+.

Model	SDRi	SI-SDRi	PESQ	STOI
BISS [13]	2.430**	1.835**	-0.045	0.106
NeuroHeed [7]	12.362*	11.751*	2.282	0.673
NeuroSpex [8]	13.818*	12.275*	2.319	0.795
NeuroSpex+*	14.187*	13.677*	2.408	0.826
NeuroSpex+	15.352	14.745	2.514	0.859

cludes the envelope decoder compared to NeuroSpex+, is used to assess the impact of envelope reconstruction. The distinguishing feature between NeuroSpex+* and NeuroSpex is the EEG encoder, where the EEG encoder in NeuroSpex+* incorporates TCN blocks to improve the representation of temporal dynamics.

It can be observed that both NeuroSpex+ and NeuroSpex+* exhibit statistically significant improvements over previous models across all metrics: SI-SDR and SI-SDRi for speech quality ($p < 0.01$), and PESQ and STOI for intelligibility. Furthermore, NeuroSpex+ shows superior performance relative to NeuroSpex+*, achieving a significant improvement of 1.068 dB in SI-SDRi. This finding supports our hypothesis that the envelope decoder can further improve EEG embedding synchronization and mask extraction, enhancing speaker extraction performance. NeuroSpex+* still outperforms NeuroHeed which can be attributed to the EEG encoder with SA and TCN pairs. Moreover, the BISS baseline shows significantly inferior performance across all models, as it employs a subject-specific linear decoder for envelope reconstruction, which is trained on single-talker data.

An ablation study is conducted to investigate the contributions of the main components of the NeuroSpex+. Table 2 shows the performance of different model configurations, where a single SA and TCN pair is used for the EEG encoder and the envelope decoder has been enhanced. The results show that incorporating envelope reconstruction leads to improved performance. Furthermore, the adoption of the VLA AI block results in a significant enhancement in the reconstruction evaluation

Table 2: Ablation study to evaluate the impact of different model components on overall performance, assessed on the validation set.

EEG Encoder	Envelope Decoder	SI-SDRi (dB)	PCC
Linear	-	3.652	-
SA+TCN	-	10.283	-
SA+TCN	Linear	12.719	0.003
SA+TCN	Conv+Linear	13.067	0.009
SA+TCN	VLA AI block	14.155	0.017

metric, PCC.

To assess the performance of speech envelope reconstruction, we isolated the EEG-to-speech envelope reconstruction branch from NeuroSpex+. We compared it against various envelope reconstruction baselines using PCC, as detailed in Table 3. Although the VLA AI baseline achieved the highest reconstruction score, this result is attributed to its relatively high parameter count. By integrating key components from the VLA AI model into our EEG-to-speech envelope reconstruction branch, which uses fewer parameters, our model also demonstrates competitive performance.

The PCC values reported in Table 3 are relatively low compared to what can normally be expected in typical envelope reconstruction [14, 24]. This is because the attended speech envelopes are reconstructed from elicited EEG responses while having competing speakers in the present, where such complex auditory environments influence the neural representation of attended speech.

Table 3: Performance comparison of different speech envelope reconstruction approaches, evaluated in terms of PCC on the test set.

Envelope Decoder	PCC	#Params
Linear	0.004	10K
CNN [24]	0.012	186K
FCNN [24]	0.009	186K
VLA AI [14]	0.036	1.74M
EEG-to-speech envelope branch	0.027	658K

5. Conclusion

We validated the idea of augmenting NeuroSpex training by using the envelope of target speech as the secondary training target. We have shown that NeuroSpex+ with dual-task training outperforms the original NeuroSpex with single task training on the target speech waveform. As the speech envelope can be derived from the speech waveform itself, the dual-task training in this study is achieved without the requirement of additional resources. Future work should explore integrating additional information from the target speech to further enhance speaker extraction performance.

6. Acknowledgments

This work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen, Germany).

7. References

- [1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, 1953.
- [2] S. Cai, H. Zhu, T. Schultz, and H. Li, "EEG-based auditory attention detection in cocktail party environment," *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 3, 2023.
- [3] C. Xu, W. Rao, E. S. Chng, and H. Li, "Spex: Multi-scale time domain speaker extraction network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1370–1384, 2020.
- [4] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, "Multimodal speakerbeam: Single channel target speech extraction with audio-visual speaker clues," in *INTERSPEECH*, 2019, pp. 2718–2722.
- [5] Z. Pan, X. Qian, and H. Li, "Speaker extraction with co-speech gestures cue," *IEEE Signal Processing Letters*, vol. 29, pp. 1467–1471, 2022.
- [6] 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," *arXiv preprint arXiv:1804.03619*, 2018.
- [7] Z. Pan, M. Borsdorf, S. Cai, T. Schultz, and H. Li, "Neuroheed: Neuro-steered speaker extraction using eeg signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [8] D. De Silva, S. Cai, S. Pahuja, T. Schultz, and H. Li, "Neurospex: Neuro-guided speaker extraction with cross-modal fusion," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 341–348.
- [9] N. Mesgarani and E. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, May 2012.
- [10] J. Kerlin, A. J. Shahin, and L. M. Miller, "Attentional gain control of ongoing cortical speech representations in a "cocktail party";" *Frontiers in Neuroscience*, vol. 8, p. 273, 2014. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fnins.2014.00273/full>
- [11] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 01 2014.
- [12] S. Cai, P. Li, E. Su, and L. Xie, "Auditory attention detection via cross-modal attention," *Frontiers in Neuroscience*, vol. 15, 2021.
- [13] 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, p. 117282, 2020.
- [14] B. Accou, J. Vanthornhout, H. V. hamme, and T. Francart, "Decoding of the speech envelope from EEG using the VLAAI deep neural network," *Scientific Reports*, vol. 13, no. 1, p. 812, 2023.
- [15] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 21–25.
- [16] K. Liu, F. Ke, X. Huang, R. Yu, F. Lin, Y. Wu, and D. W. K. Ng, "DeepBAN: A temporal convolution-based communication framework for dynamic WBANs," *IEEE Transactions on Communications*, vol. 69, no. 10, pp. 6675–6690, 2021.
- [17] M. Borsdorf, S. Cai, S. Pahuja, D. De Silva, H. Li, and T. Schultz, "Attention and sequence modeling for match-mismatch classification of speech stimulus and EEG response," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 799–809, 2024.
- [18] J. Lin, A. J. d. L. van Wijngaarden, K.-C. Wang, and M. C. Smith, "Speech enhancement using multi-stage self-attentive temporal convolutional networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3440–3450, 2021.
- [19] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [20] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.
- [21] W. Biesmans, N. Das, T. Francart, and A. Bertrand, "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 5, pp. 402–412, 2016.
- [22] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processings*, vol. 2. IEEE, 2001, pp. 749–752.
- [23] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech, and Signal Processings*. IEEE, 2010, pp. 4214–4217.
- [24] M. Thornton, D. Mandic, and T. Reichenbach, "Robust decoding of the speech envelope from EEG recordings through deep neural networks," *Journal of Neural Engineering*, vol. 19, no. 4, p. 046007, 2022.