SPEAKER-INDEPENDENT BRAIN ENHANCED SPEECH DENOISING

Maryam Hosseini, Luca Celotti, Éric Plourde

NECOTIS, Department of Electrical and Computer Engineering
Université de Sherbrooke, QC, Canada
seyedeh.maryam.hosseini.telgerdi@usherbrooke.ca

ABSTRACT

The auditory system is extremely efficient in extracting attended auditory information in the presence of competing speakers. Single-channel speech enhancement algorithms, however, greatly lack this efficacy. In this paper, we propose a novel deep learning method referred to as the Brain Enhanced Speech Denoiser (BESD), that takes advantage of the attended auditory information present in the brain activity of the listener to denoise a multi-talker speech. We use this information to modulate the features learned from the sound and the brain activity, in order to perform speech enhancement. We show that our method successfully enhances a speech mixture, without prior information about the attended speaker, using electroencephalography (EEG) signals recorded from the listener. This makes it a great candidate for realistic applications where no prior information about the attended speaker is available, such as hearing aids or cell phones.

Index Terms— speech enhancement, deep learning, EEG signals

1. INTRODUCTION

In a cocktail party problem, where multiple competing speakers are present, the auditory system is extremely efficient in extracting attended auditory information while filtering out irrelevant information. Speech enhancement algorithms [1, 2, 3], which aim at removing background noise from a target noisy speech, greatly lack this efficacy.

One problem that is particularly difficult for speech enhancement algorithms is when the background noise is composed of different simultaneous speakers, such as in a speaker separation task. In the past decade, speaker separation algorithms have taken advantage of deep learning approaches, with significant performance improvements compared to traditional methods. However, most speaker separation algorithms require prior knowledge of the number of speakers as well as of the target speaker, if the goal is to extract one specific speaker [4, 5]. This greatly limits the real world applicability of these algorithms. Moreover, these methods mostly use a spectro-temporal representation of the signals and aim at reconstructing the time domain signal thus requiring an estimation of the signal's phase. Possible errors in this estimation therefore impose an upper bound on the performance of these systems.

Recently, it has been shown that the characteristics of an attended speaker can be extracted from the brain waves of a listener [6]. This has given rise to brain controlled speech processing for hearing aids. In these methods, the electroencephalography (EEG) activity of the listener is used to find an estimation of the attended speaker in a multi-speaker setting. To do so, a source separation is first performed on the speech mixture and this estimation is compared with different sound sources in the environment, to find the most likely speaker. The signal corresponding to this speaker is then added back to the mixture, to make it more prominent [7]. The need for source separation in this approach increases the computational cost. To address this problem, Ceolini et. al [8] modified the approach in [7] to use the EEG to guide a speech extraction algorithm. However, this approach requires two different networks trained separately, one for the estimation of the attended speaker and another one for the speech extraction. Moreover, these methods do not perform speech denoising per se but rather aim at amplifying the attended speaker's signal in the speech mixture.

In this paper we address the problem of denoising in a cocktail party environment without prior knowledge about the target speaker, which we refer to as speaker-independent denoising. We propose a novel deep learning technique for speech enhancement and denoising in a multiple speaker situation, using cues extracted from the EEG signals of the attended speaker. We refer to the proposed approach as a Brain Enhanced Speech Denoiser (BESD). In contrast to existing approaches, we propose an end to end speech denoising approach performed entirely in the time domain, thus avoiding the limitations encountered with a spectro-temporal representation. Moreover, all the modules of the proposed approach are trained in a single neural architecture, which lowers the complexity of the algorithm. Most importantly, the proposed speech denoising approach does not need any prior information about the target speaker when performing the enhancement. The proposed BESD could thus be used in applications where no prior information about the attended speaker is

present, such as hearing aids, cell phones or noise cancelling headphones, which is not the case for current speech denoising algorithms.

2. METHODS

2.1. Data acquisition

The data from all subjects in this experiment has been obtained from the authors of [9] [1]. In this subsection, we summarize the data acquisition procedure, for additional details, please refer to [9]. All procedures were performed in accordance with the Declaration of Helsinki and were approved by the Ethics Committees of the School of Psychology at Trinity College Dublin, and the Health Sciences Faculty at Trinity College Dublin [9]. The subjects ($n = 20$) undertook 30 trials, of 60 seconds each. During the experiment, they were presented with two stories, one to the left ear and the other to the right ear. Each story was read by a different male speaker. Subjects were divided in two groups and each group was instructed to pay attention to either the left (8) or the right ear (12). 128-channel EEG data were recorded at a rate of 512 Hz using a BioSemi ActiveTwo system and further downsampled to 128 Hz.

2.2. Preprocessing

To lower the amount of memory needed, we downsample the sound stimuli to a sampling rate of 14.7 kHz. The EEG data is first band-pass filtered between 0.1 and 45 Hz. Channels with excessive noise are recalculated by spline interpolation of the surrounding channels. The EEGs are then re-referenced to the average of the mastoid channels. To remove artefacts created by eye blinking and other muscle movements, we perform independent component analysis (ICA). All the analysis are performed in EEGLAB [10].

2.3. Proposed Brain Enhanced Speech Denoiser (BESD)

The proposed Brain Enhanced Speech Denoiser (BESD) (Fig. 1a) has an autoencoder structure with two encoders, one that extracts features of the brain activity and one that extracts features of the speech mixture, as well as a decoder to reconstruct the enhanced speech. We use a general class of fusion methods called conditional normalization (CN) to modulate learned sound features with learned EEG features and vice versa. Our model can be viewed as a development on Conditional Batch Normalization (CBN) [11] and Feature-wise Linear Modulation (FiLM) [12].

The neural activity extracted from the EEG signal is first upsampled by a ratio of 114, to match its dimensions with those of the sound mixture. Each encoder includes three convolutional blocks with a FiLM like modulation between the two sensory pipelines and another single convolutional before the latent space. Each convolutional block (Fig. 1b) is constructed by chaining a 1D convolution of kernel size 25, stride 1 and causal padding, followed by a layer normalization, a leaky ReLU non-linearity and a dropout of 0.3. The only difference between each convolutional block is the filter size. From the input to the latent space, i.e. before the fusion layer, the three different 1D convolutions have filter sizes of respectively 100, 52 and 5 and the last convolution before the Fusion has a filter size of 5 too.

In the FiLM block, we learn four functions of the sound mixture representation s and EEG representation e:

$$\gamma_{s,c} = f_{1,c}(s) \qquad \beta_{s,c} = h_{1,c}(s) \tag{1}$$

$$\gamma_{e,c} = f_{2,c}(e) \qquad \beta_{e,c} = h_{2,c}(e) \tag{2}$$

where c is the feature number. $\gamma_{s,c}$ and $\beta_{s,c}$ modulate the EEG input signal and $\gamma_{e,c}$ and $\beta_{e,c}$ modulate the sound mixture input via a feature-wise transformation as follows:

$$O_{e,c} = \gamma_{s,c} \times e + \beta_{s,c} \tag{3}$$

$$O_{s,c} = \gamma_{e,c} \times s + \beta_{e,c} \tag{4}$$

where $O_{s,c}$ and $O_{e,c}$ are the outputs of the FiLM block. Here, $f_{1,c}$, $h_{1,c}$, $f_{2,c}$ and $h_{2,c}$ are all 1D convolutional layers, with a kernel size of 3 and a number of filters equal to the features dimension of the input to the FiLM layer. As in the original work [12], the modulation is done at several layers along the encoders. The output of both encoders is concatenated in the latent space in what we call Fusion in Fig. 1a.

The decoder is characterized by two convolutional blocks with the same characteristics as the convolutional blocks in the encoder, with filter sizes of 52 and 100. The last layer is a 1D convolution of filter size 1 followed by a hyperbolic tangent.

The network is trained for 60 epochs and a batch size of 16. The stimuli are divided into 2 seconds sections. We use a scale-invariant signal to-distortion ratio (SI-SDR) [13] loss, that has been shown to perform well as a general-purpose loss function for time-domain speech enhancement [14]. We use the Adam optimizer with a learning rate set to 10^{-5}, with a 0.1 drop if we have no change in the loss value for 5 epochs. This proposed approach with the EEG as the input brain activity is referred to as the BESD-EEG approach in the following.

2.4. Frequency-band coupling model

Apart from using directly the EEG signals as the input to the proposed BESD approach, we also propose to use instead a frequency-band coupling (FBC) model that estimates the cortical multi-unit neural activity (MUA) from EEGs [15]. This model combines the power of the gamma band (30-45 Hz) and the phase of the delta band (2-4 Hz) of the EEG signals. It has been shown to be a good estimate of the neural activity in the visual and auditory systems [16, 15]. The model is

[1] https://doi.org/10.5061/dryad.070jc

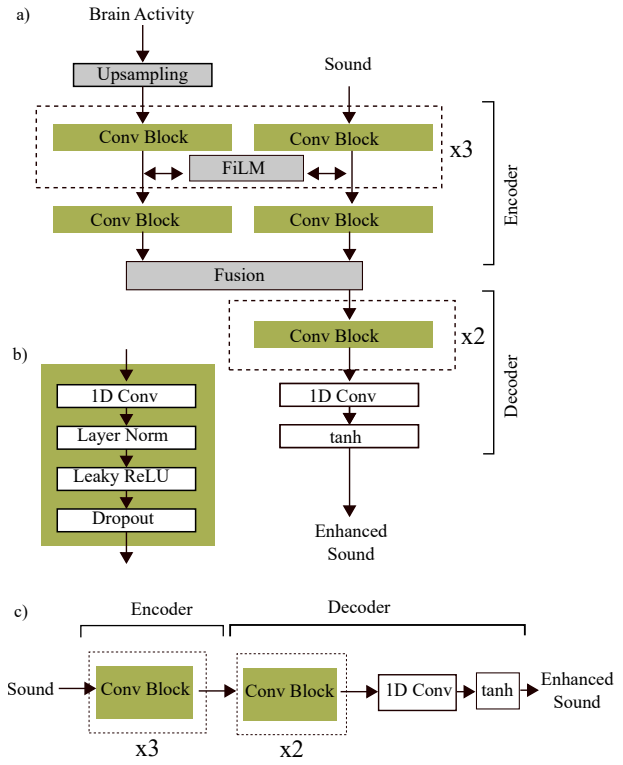

Fig. 1. Illustration of the networks used in the analysis: a) Brain Enhanced Speech Denoiser (BESD) architecture, b) the convolutional block and c) the denoising autoencoder.

presented as a linear combination of these two signals:

$$N(n) = a_\gamma \times P_\gamma(n) + a_\delta \times \angle\delta(n) \quad (5)$$

where $n = 1, ..., M$ is the time index, $N(n)$ is the neural activity at time index n, $P_\gamma(n)$ and $\angle\delta(n)$ are the power of the gamma band and the phase of the delta band respectively and a_γ and a_δ are their respective coefficients.

For the envelope of the gamma band, we used the magnitude of the Hilbert transform of the bandpass filtered EEG signal (between 30-45 Hz) and we extracted the phase of the delta band from the angle of the Hilbert transform of the bandpass filtered EEG signal (between 2-4 Hz). The values of the a_γ and a_δ were both fixed to 0.5 as per [16, 15]. The output of the FBC model is called multi-unit activity (MUA) from here on and the approach that uses the MUA as the input brain activity in Fig. 1a is referred to as BESD-MUA in the following.

3. RESULTS AND DISCUSSION

We evaluated the proposed BESD approach in two different settings. Firstly, we applied the BESD to a speaker specific denoising problem, in which the attended speaker is known. We compared our results to the results of a denoising autoencoder [1] shown in Fig. 1c. To do this, we first trained the BESD and the autoencoder to only extract the second speaker and further tested the trained algorithm, i.e. performed the enhancement per se, also only on the data where the attended speaker was the second speaker. To show the benefits of using the FBC model, we trained the BESD on both the EEG input and the MUA. Secondly, we applied our method to a situation where no prior information is available about the attended speaker when performing the enhancement. The BESD should then be able to extract the attended speaker using the brain signal information and denoise the signal. For this case, we trained the BESD on the data from both of the attended speakers, to enhance the respective attended speaker, with the information available from the recorded EEG of the listener or the MUA. We then performed the enhancement on the data that included both speakers as the attended speaker without any specific information on which speaker needed to be enhanced. For both settings, we used the segmental SNR (segSNR), SI-SDR and short-time objective intelligibility (STOI) [17] as objective performance metrics. We report the results in the form of violin plots, which represents the performance distribution for a given metric.

3.1. Speaker-specific denoising

In this section we present the results for the first setting, where the attended speaker is known. The aim of this experiment is to study if the BESD performs better than a denoising autoencoder with a similar structure, even if the speaker is known to the algorithm. The results are shown in Fig. 2.

As can be observed, the BESD significantly improves the quality and intelligibility of the enhanced speech compared to the denoising autoencoder, for either the EEG or MUA inputs ($p \ll 0.001$, Mann-Whitney U test). This is somewhat surprising, since the information about the attended speaker is already present when training the model. It has been shown that by manipulating a neural network's intermediate features using FiLM layers, it can carry out diverse tasks [11, 12]. We think that using BESD modulates the intermediate features of the sound and the EEG signal to build a more meaningful representation of the attended speaker, leading to a better performance. Moreover, we saw a significant improvement in the performance of BESD-MUA when evaluating the segSNR and STOI ($p = 1.9 \times 10^{-15}$ and 0.0016, respectively, Mann-Whitney U test) but no significant improvement in terms of SI-SDR.

3.2. Speaker-independent denoising

Next, we investigated the situation where no prior information about the target speaker is available during the enhancement stage. We trained the BESD to automatically extract the attended speaker, using the information present in the EEG signal or MUA, and to further perform the denoising.

Fig. 3 shows the performance of the approach, using the segSNR, SI-SDR and STOI metrics. Since no other speech

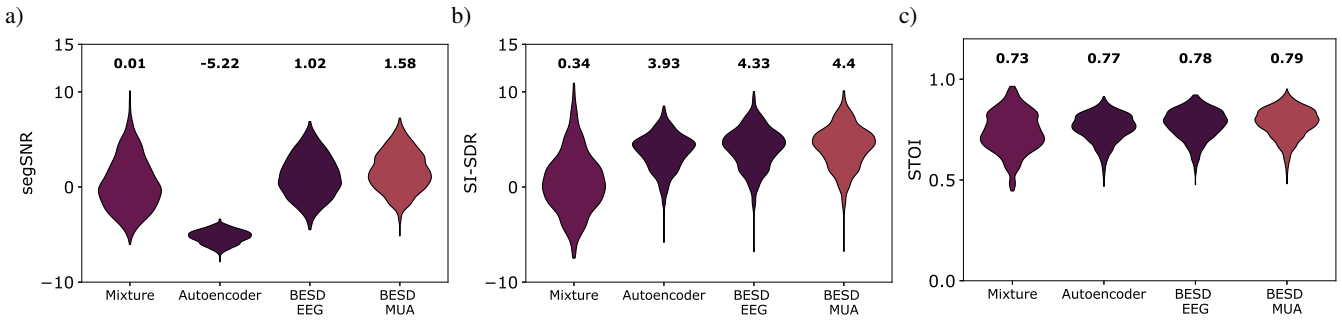

Fig. 2. Speech enhancement performance distributions for speaker-specific denoising. We show the performance for the noisy mixture, autoencoder and BESD with EEG signal and MUA. Medians are shown on top.

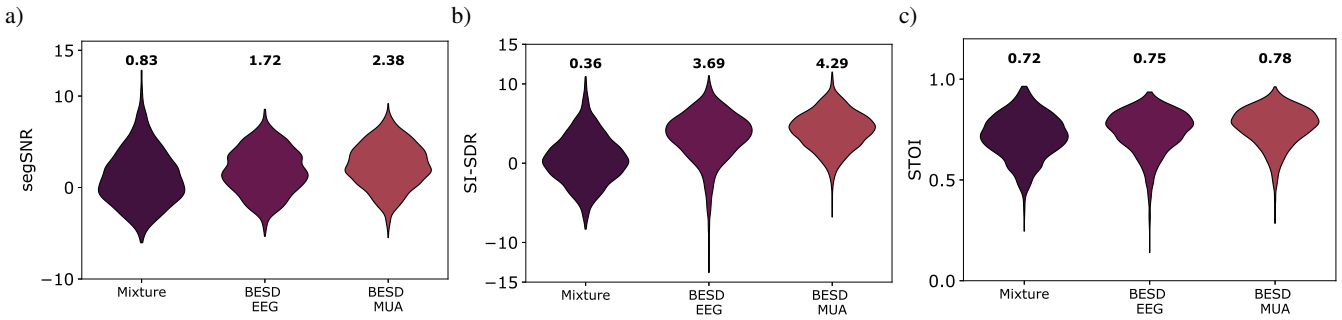

Fig. 3. Speech enhancement performance distributions for speaker independent denoising. We show the performance for the noisy mixture and BESD with EEG signal and MUA. Medians are shown on top.

denoising algorithm, known to the authors, can perform denoising without knowing the target speaker, we compare the performance of the proposed algorithm to the input noisy mixture as well as to the results obtained for the speaker-specific setting. We show that the BESD significantly improves the quality of the enhanced speech ($p \ll 0.001$, Mann-Whitney U test). Furthermore, we observe that the results obtained for the different metrics in the speaker-independent setting (Fig. 3) are on par with those obtained for the speaker-dependent setting (Fig. 2). This shows the great benefit of using brain signals, since even in the absence of any prior information about the attended speaker,the BESD is able to accurately extract information about the attended speaker from the brain signals and perform denoising. Interestingly, using the MUA instead of the EEG signal significantly increased the performance of the BESD for all metrics ($p \ll 0.001$, Mann - Whitney U test). This can be due to the fact that an EEG is a noisy mixture of several underlying cortical sources. By using the FBC model, the most relevant features of the EEG are captured and a better extraction of the attended speaker can be obtained [15].

4. CONCLUSION

It is known that features of the attended speech can be decoded from the brain activity of the listener [6]. In this paper, we use it to design a speaker-independent speech denoising approach, the Brain Enhanced Speech Denoiser (BESD). When the attended speaker is known, our experiments show that the BESD outperforms a denoising autoencoder. Moreover, we show that when no prior information is available about the attended speaker, the BESD performs as well as when the target speaker is known, showing the great advantage of using brain activity signals in a speech denoising algorithm. The proposed BESD is an end to end approach, performed completely in time-domain, where all the modules of the algorithm are trained simultaneously in a single neural architecture, lowering the complexity of the algorithm. Most importantly, given enough training data, the algorithm does not need to know the number of speakers or the target speaker, i.e. it does not need any prior information about the speakers. This makes the BESD a great approach for applications where no prior information about the attended speaker is present, such as hearing aids or cell phones.

5. ACKNOWLEDGEMENT

The authors would like to thank the authors of [9] for kindly providing the data for this study.

6. REFERENCES

[1] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori, "Speech enhancement based on deep denoising autoencoder.," in *Interspeech*, 2013, pp. 436–440.

[2] Ashutosh Pandey and DeLiang Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[3] Lei Sun, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, 2017, pp. 136–140.

[4] Yi Luo and Nima Mesgarani, "TasNet: time-domain audio separation network for real-time, single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 696–700.

[5] Zhuo Chen, Yi Luo, and Nima Mesgarani, "Deep attractor network for single-microphone speaker separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 246–250.

[6] Nima Mesgarani and Edward F Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[7] James A O'Sullivan, Alan J Power, Nima Mesgarani, Siddharth Rajaram, John J Foxe, Barbara G Shinn-Cunningham, Malcolm Slaney, Shihab A Shamma, and Edmund 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, 2015.

[8] Enea Ceolini, Jens Hjortkjær, Daniel DE Wong, James O'Sullivan, Vinay S Raghavan, Jose Herrero, Ashesh D Mehta, Shih-Chii Liu, and Nima Mesgarani, "Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception," *NeuroImage*, p. 117282, 2020.

[9] Michael P Broderick, Andrew J Anderson, Giovanni M Di Liberto, Michael J Crosse, and Edmund C Lalor, "Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech," *Current Biology*, vol. 28, no. 5, pp. 803–809, 2018.

[10] Arnaud Delorme and Scott Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[11] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville, "Modulating early visual processing by language," in *Advances in Neural Information Processing Systems*, 2017, pp. 6594–6604.

[12] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville, "Film: Visual reasoning with a general conditioning layer," *arXiv preprint arXiv:1709.07871*, 2017.

[13] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "SDR–half-baked or well done?," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[14] Morten Kolbæk, Zheng-Hua Tan, Søren Holdt Jensen, and Jesper Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 825–838, 2020.

[15] Marc-Antoine Moinnereau, Jean Rouat, Kevin Whittingstall, and Eric Plourde, "A frequency-band coupling model of EEG signals can capture features from an input audio stimulus," *Hearing Research*, p. 107994, 2020.

[16] Kevin Whittingstall and Nikos K Logothetis, "Frequency-band coupling in surface EEG reflects spiking activity in monkey visual cortex," *Neuron*, vol. 64, no. 2, pp. 281–289, 2009.

[17] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4214–4217.