Speaker-Independent Speech Enhancement with Brain Signals

Maryam Hosseini [1], Luca Celotti [2], and Eric Plourde [2]

[1]Université de sherbrooke
[2]Affiliation not available

November 1, 2023

Abstract

Single-channel speech enhancement algorithms have seen great improvements over the past few years. Despite these improvements, they still lack the efficiency of the auditory system in extracting attended auditory information in the presence of competing speakers. Recently, it has been shown that the attended auditory information can be decoded from the brain activity of the listener. In this paper, we propose two novel deep learning methods referred to as the Brain Enhanced Speech Denoiser (BESD) and the U-shaped Brain Enhanced Speech Denoiser (U-BESD) respectively, that take advantage of this fact to denoise a multi-talker speech mixture. We use a Feature-wise Linear Modulation (FiLM) between the brain activity and the sound mixture, to better extract the features of the attended speaker to perform speech enhancement. We show, using electroencephalography (EEG) signals recorded from the listener, that U-BESD outperforms a current autoencoder approach in enhancing a speech mixture as well as a speech separation approach that uses brain activity. Moreover, we show that both BESD and U-BESD successfully extract the attended speaker without any prior information about this speaker. This makes both algorithms great candidates for realistic applications where no prior information about the attended speaker is available, such as hearing aids, cellphones, or noise cancelling headphones. 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.

Speaker-Independent Speech Enhancement with Brain Signals

Maryam Hosseini, Luca Celotti, and Éric Plourde, *Senior member, IEEE*

***Abstract*—Single-channel speech enhancement algorithms have seen great improvements over the past few years. Despite these improvements, they still lack the efficiency of the auditory system in extracting attended auditory information in the presence of competing speakers. Recently, it has been shown that the attended auditory information can be decoded from the brain activity of the listener. In this paper, we propose two novel deep learning methods referred to as the Brain Enhanced Speech Denoiser (BESD) and the U-shaped Brain Enhanced Speech Denoiser (U-BESD) respectively, that take advantage of this fact to denoise a multi-talker speech mixture. We use a Feature-wise Linear Modulation (FiLM) between the brain activity and the sound mixture, to better extract the features of the attended speaker to perform speech enhancement. We show, using electroencephalography (EEG) signals recorded from the listener, that U-BESD outperforms a current autoencoder approach in enhancing a speech mixture as well as a speech separation approach that uses brain activity. Moreover, we show that both BESD and U-BESD successfully extract the attended speaker without any prior information about this speaker. This makes both algorithms great candidates for realistic applications where no prior information about the attended speaker is available, such as hearing aids, cellphones, or noise cancelling headphones.**

***Index Terms*—Deep learning, EEG signals, Speech enhancement.**

I. Introduction

THE auditory system is extremely efficient in extracting attended auditory information in real-world speech communication, where multiple competing speakers are present. This efficiency is especially important for applications such as speech recognition [1], speech processing for hearing aids and cochlear implants [2] as well as speaker verification [3]. However, current speech enhancement algorithms, aimed at increasing the quality and intelligibility of a degraded speech signal, lack this efficacy. There is thus a need to improve the performance of these algorithms in such conditions [4].

The speech enhancement problem has been extensively studied in the past decades. The first algorithms proposed, such as spectral subtraction [5], Wiener filtering [6] and the Bayesian minimum mean square error (MMSE) [7] used a filter to reduce noise components from the noisy speech signals. Other algorithms like principal component analysis [8], singular value decomposition [9] and the generalized subspace approach [10] used a subspace to separate the clean sound and noise components from the noisy signal and then reconstruct the clean signal from the clean components. Whereas these algorithms perform well under stationary noise conditions, they are less successful at accurately estimating noise components under nonstationary noise conditions leading to bad intelligibility and quality.

In recent years, speech enhancement algorithms, and more specifically speaker separation algorithms, have benefited from deep learning approaches, with a significant improvement in efficiency compared to traditional methods. Methods like Time-domain Audio Separation Net (TasNet) [11], Permutation Invariant Training (PIT) [12], Deep Attractor Network (DAN) [13] and Deep Clustering [14], have been particularly successful in speaker separation tasks, where the background noise is composed of different simultaneous speakers. However, most speaker separation algorithms require some kind of prior knowledge about the auditory scene either during training or inference, such as the number of speakers or more specifically the target speaker if the goal is to extract one specific speaker [11], [12], [13], [14]. In real world situations, the number of speakers can change as the auditory scene changes. Moreover, the need to know the target speaker necessitates prior knowledge about the identity of speakers in the scene. These greatly limit the real-world applicability of these algorithms. Furthermore, these methods are mostly formulated in the time-frequency domain 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.

As mentioned, speech enhancement and speaker separation in the presence of non-stationary noise and for applications such as hearing aids is hard to perform without prior information about the target speaker. Recently, it has been shown that the attended speaker can be decoded from the brain waves of the listener [15], [16], [17]. This has led to the improvement of speech perception in hearing aids. In these applications, the envelope of the attended speaker is first extracted from the brain waves of the listener. Next, a speech separation is performed on the speech mixture to separate the speech sources and the estimated envelope is compared to each of these sources to find the most probable speaker. The speaker separation is typically implemented using either multichannel approaches such as beamforming [17] or single channel approaches using neural networks [16]. The attended speaker is the one that has the highest similarity to the extracted envelope. This estimated speaker is then added back to the mixture to increase the relative intensity of this speaker [16], [18]. One major drawback in these methods is the need for source separation, which increases the computational cost. Moreover, in a real-world application, separating all the

M. Hosseini, L. Celotti and E. Plourde are with the Department of Electrical and Computer Engineering, Université de Sherbrooke, Sherbrooke, Qc, J1K 2R1, Canada e-mail: (Seyedeh.maryam.hosseini.telgerdi, luca.celotti, eric.plourde@usherbrooke.ca.)

sources seems unnecessary as the listener is interested in only one speaker.

To address this problem, Ceolini *et al.* [19] designed a deep learning network, refered to as the Brain-inspired speech separation (BISS) model, that extracts the envelope of the attended speaker from the electroencephalography (EEG) of the listener to jointly perform speech extraction and separation. However, this approach requires training two networks separately, the guiding network and the speaker extraction network, which increases the complexity. This method also uses the time-frequency representation of the speech mixture, which might cause possible errors and inaccuracies.

In this paper we specifically address the problem of speech enhancement in a noisy environment without any prior knowledge about the target speaker, to which we refer as speaker-independent denoising [1]. We first introduce a simple network, the Brain Enhanced Speech Denoiser (BESD) [20], that takes advantage of the attended auditory information present in the brain activity of the listener to denoise a multi-talker speech. We use a Feature-wise Linear Modulation (FiLM) between the brain activity and the sound mixture, to better extract the features of the attended speaker to perform speech enhancement. We build upon this network and propose a deep learning technique for speech enhancement and denoising in a noisy environment that performs much better than BESD. We refer to the proposed approach as a U-shaped Brain Enhanced Speech Denoiser (U-BESD). We also show that our algorithm surpasses the performance of the approach proposed by Ceolini *et al.* [19]. Moreover, compared to [19], these proposed networks are end to end speech enhancement and denoising approaches performed entirely in the time domain, thus avoiding the limitations encountered with a spectro-temporal representation. Furthermore, all the modules of the proposed approaches are trained in a single neural architecture, lowering the complexity of the algorithm. These proposed networks could be used in applications where no prior information about the attended speaker is present, such as hearing aids, cell phones, or noise cancelling headphones.

II. Proposed approaches

In this section, we present two algorithms to perform speech enhancement using brain activity. First, we present the Brain Enhanced Speech Denoiser (BESD) followed by the U-shaped Brain Enhanced Speech Denoiser (U-BESD) which builds on the BESD.

A. Proposed Brain Enhanced Speech Denoiser (BESD)

The proposed BESD shown in Fig. 1a, has an autoencoder structure with two encoders, one for extracting the features of the brain signal and one for extracting the features of the sound mixture. There is also a decoder that reconstructs the enhanced speech. In the following, we describe in detail each block of Fig. 1a.

[1]Part of this work has been published in [20]. This previous work has been substantially improved here by proposing a new enhancement algorithm (U-BESD) that significantly performs better than that of [20]. Moreover, we also added a more detailed analysis as well as bonified the experimental validation.

As can be seen in Fig. 1a, the encoders contain Conv blocks with a Feature-wise Linear Modulation (FiLM) block, explained further, between the two pipelines.

All convolutional blocks detailed in Fig. 1b, have a similar structure. Each block is constructed by chaining a 1 dimensional (1D) causal convolution with the Glorot weight initialization [21], followed by a Post Conv block (Fig. 1c) consisting of layer normalization, a leaky ReLU nonlinearity and a dropout layer. The only difference between each convolutional block is the filter size, which is decreasing as the network deepens.

To guide the algorithm to better extract the features of the sound mixture and the brain signal, we use a general class of fusion methods called conditional normalization (CN) [22], [23], [24]. These methods use the learned functions of some input to condition the learned features of another input. Here, the fusion algorithm modulates the learned features of the sound mixture using brain signals and vice versa. There are various highly effective methods of CN with different functionalities. Our network can be viewed as a development on FiLM [25]. Here we learn to adaptively affect the output of each layer of the network by applying a normalization function, based on the features extracted from the brain signals and the sound mixture at each layer. As has been shown before, manipulating the intermediate layers of networks can improve their performance [22], [25]. In the FiLM block, the network learns four functions of the inputs, i.e., the sound mixture representation s and the 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. In (1) and (2), $\gamma_{e,c}$ and $\beta_{e,c}$ are the functions that modulate the sound mixture representation and $\gamma_{s,c}$ and $\beta_{s,c}$ are the functions that modulate the EEG signal representation, via a feature-wise linear 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 and the input to the sound and EEG pipeline respectively. Here, $f_{1,c}$, $h_{1,c}$, $f_{2,c}$ and $h_{2,c}$ are all 1D convolutional layers, with a Glorot weight initializer [21], and a number of filters equal to the last dimension of the input to the FiLM block. The FiLM block is applied after every Conv block of the encoder except for the last block. The output of both encoders are concatenated in the latent space in what we call Concatenation in Fig. 1a.

The decoder is composed of 3 1D causal convolutions with a Glorot weight initializer. The last 1D convolution has a filter size of 1 followed by a hyperbolic tangent to reconstruct the enhanced speech of the attended speaker.

B. Proposed U-shaped Brain Enhanced Speech Denoiser (U-BESD)

Apart from BESD, we also propose a second network, U-BESD, that has U-shaped structure (Fig. 1d). Similar to BESD,

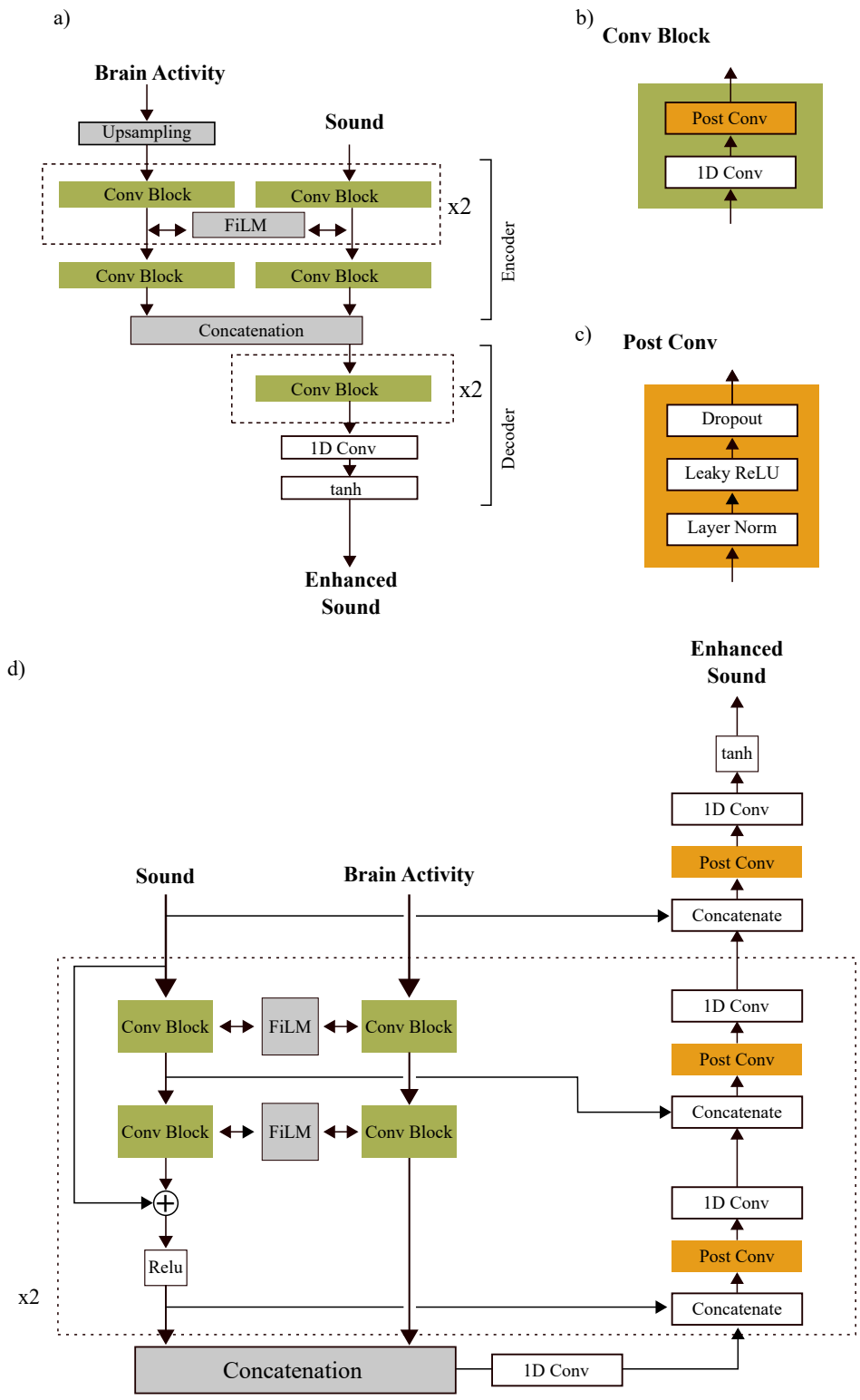

Fig. 1: Illustration of the proposed networks: a) Brain Enhanced Speech Denoiser (BESD) architecture. b) The convolutional block. c) The Post-Conv block. d) U-shaped Brain Enhanced Speech Denoiser (U-BESD) architecture.

U-BESD has two encoders, one for extracting the features of the brain signal and one for extracting the features of the sound mixture. It also has a decoder that reconstructs the enhanced speech. In the following, we describe in detail each block of Fig. 1d and the differences between BESD and U-BESD.

One of the differences between BESD and U-BESD is the use of skip connections in the sound encoder. In fact, as can be seen in Fig. 1d, each skip connection passes over one Conv block and is then added to the output of the second Conv block. These skip connections improve the flow of information

to deeper layers by allowing the network to learn small differences with respect to the previous layer and alleviate the problem of vanishing gradients, i.e., when the gradient tends to get very close to zero in the back propagation from the output to the input. The addition of skip connections has been proven effective for image classification [26] and we include them in the network assuming they should improve the performance. The encoder for EEG signals has no skip connections and includes four Conv blocks.

All convolutional blocks are similar to those used in BESD. However, U-BESD uses dilated convolutions instead of normal convolutions as in BESD. The only difference between each Conv block is the dilation rate which increases by a factor of 2 for each convolution as the network deepens. The use of dilated convolutions is the second difference between BESD and U-BESD. In a dilated convolution, the kernel is stretched over a larger area by inserting holes in between its elements, hence the name convolution à trous or convolution with holes. The dilation rate indicates the factor by which the kernel is stretched. Using a dilated convolution is similar to pooling or strided convolutions, but the output dimension remains the same. Dilated convolution with dilation 1 is the same as a normal convolution [27], [28].

U-BESD also uses FiLM blocks between the two encoders, similar to BESD. 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 Concatenation in Fig. 1d.

The decoder is built up of 6 1D causal convolutions with a Glorot weight initializer. Except the last convolution, the output of each convolution is concatenated by a skip connection from the corresponding layer of the encoder, which was not the case in BESD. The output of this concatenation is then passed to a Post Conv block shown in Fig. 1c. The last layer is a 1D causal convolution with filter size of 1 followed by a hyperbolic tangent to reconstruct the enhanced speech of the attended speaker. Several variants of this structure have been studied, which are further detailed in Subsection IV.D.

III. Materials and Methods

In this section, we first summarize the data acquisition, the stimuli used in the experiments, and their preprocessing. We then present an algorithm to estimate the multi-unit neural activity from EEG signal that we further use as an input to the proposed networks. We also present the loss function and the optimizer. Finally, we summarize the evaluation metrics used.

A. Data acquisition

The data from all subjects used in this study have been obtained from the authors of [29][2]. 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. All subjects were native English speakers, reported normal hearing and no history of neurological diseases. A total of 34 subjects (28 males) with a mean age of 27.3 ± 3.2 years participated in the experiments. The data from subject no. 6 were excluded due to noisy recordings.

The subjects undertook 30 trials, of 60 seconds each. During each trial, 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 (17) or the right ear (16 + 1 excluded subject). After each trial, subjects were required to answer multiple choice questions on each story to test their attention. The story line was preserved, such that for each trial, the story began where the last trial ended.

In order not to bias the attention towards one stimulus, the stimuli amplitudes were normalized to have the same root mean square (RMS) level and silent gaps were cut short to a maximum of 0.5 s. Stimuli were presented using Sennheiser HD650 headphones and a presentation software from Neurobehavioral Systems at a sampling rate of 44.1 kHz. Subjects were asked to maintain a visual fixation on a cross hair centered on the screen and to minimize eye blinking and other motor activities.

EEG data were recorded using a 128-channel (plus two mastoids) EEG cap, at a rate of 512 Hz using a BioSemi ActiveTwo system and further downsampled to 128 Hz. For more details regarding the experimental procedure, please refer to [29].

To lower the amount of memory needed, we downsampled the sound stimuli to a sampling rate of 14.7 kHz. We divided the data into the following three groups. First, we randomly kept 5 trials from all subjects as the test data and 2 trials as the validation data. The rest of the data were considered as the training data. For the training and validation sets, each trial of 60 seconds were then cut into 2-seconds signals. For the testing set, each trial was cut into 20-seconds long signals.

B. EEG preprocessing

The EEG data were first band-pass filtered between 0.1 and 45 Hz. This was done to only keep the relevant frequency bands and to remove electrical noise (50 or 60 Hz) or very low frequency noise that is a sign of a drift in the recording environment. To identify channels with excessive noise, the standard deviation (SD) of each channel was compared to the SD of the surrounding channels and each channel was visually inspected. Channels with excessive noise were recalculated by spline interpolation of the surrounding channels.

The EEGs were re-referenced to the average of the mastoid channels to avoid introducing noise from the reference site. When the number of electrodes is dense enough, converting data to an average reference is particularly important. The advantage of re-referencing is that the sum of the outward positive and negative currents across the entire head will be zero [30]. To remove artefacts created by eye blinking and other muscle movements, we performed independent component analysis (ICA). For each subject, any trial that contained

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

too much noise was excluded from the study. All analysis were performed in EEGLAB [31]. As the sampling rate of the sound is 14.7 kHz (after downsampling the original signal of 44.5 kHz by a factor of 3) and the sampling rate of the EEG signal is 128 Hz, we finally upsample the EEG signal by a ratio of 114.

C. Frequency-band coupling model

EEG signals are a noisy mixture of several underlying sources. Therefore, instead of using directly the EEG signal as the input to the proposed BESD and U-BESD approaches, it would be best to directly record the underlying neural activity, which contains more relevant information about the stimuli, and use this activity as the input to the network. However, recording the neural activity directly is an invasive procedure, which therefore strongly restricts its use in human experiments. As a result, we propose to use instead a frequency-band coupling (FBC) model that estimates the cortical multiunit neural activity (MUA) from EEG signals and has been shown to be a good estimate of the neural activity in the visual and auditory systems [32], [30]. This model is presented as a linear combination of the amplitude of the gamma band (30-45 Hz) and the phase of the delta band (2-4 Hz) of the EEG signals:

$$N(t) = a_\gamma \times P_\gamma(t) + a_\delta \times \angle\delta(t) \tag{5}$$

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

For the amplitude of the gamma band, we first band-pass filtered EEG signals between 30-45 Hz, and then we used the magnitude of the Hilbert transform of these signals. Next, we band-pass filtered EEG signals between 2-4 Hz and we extracted the phase of the delta band from the angle of the Hilbert transform of these signals. The values of the a_γ and a_δ were both fixed to 0.5 as per [32], [30]. The output of the FBC model is referred to as multiunit activity (MUA) from here on, and the approaches that use the MUA and EEG signals as the input brain activity are referred to as BESD/U-BESD MUA and BESD/U-BESD EEG respectively in the following.

D. Loss function and optimizer

A scale-invariant signal-to-distortion ratio (SI-SDR) [33] loss function is used, which has been shown to perform well as a general-purpose loss function for time-domain speech enhancement [34]. SI-SDR can be calculated as follows:

$$\text{SI-SDR} = 10\log_{10}\frac{||e_{target}||^2}{||e_{res}||^2} \tag{6}$$

where

$$e_{target} = \frac{r^T g}{||g||^2} g \tag{7}$$

$$e_{res} = e_{target} - r \tag{8}$$

in which g and r are the target speaker and reconstruction of the target speaker in the time domain. e_{target} is the scaled target speaker and e_{res} is the residual noise. Scaling the target speaker insures that the SI-SDR is invariant to the scale of the reconstruction. This is specially desirable in applications where the processing does not necessarily preserve the scale of the target speaker. Moreover, both speech quality and intelligibility are invariant to scaling to a large extent [35].

A higher SI-SDR value indicates a better reconstruction. As a result, the objective of the training is to minimize the following loss:

$$Loss = -\,\text{SI-SDR} \tag{9}$$

The Adablief optimizer is used which has been shown to have a fast training, good generalization and training stability [36] with a learning rate set to 10^{-5} and weight decay of 0.1. The learning rate was reduced by a factor of 0.1 if there was no change in the loss value for 10 epochs.

E. Evaluation metrics

We evaluated and compared the performance of the proposed algorithms through three objective metrics. First, we used the SI-SDR [33], which is proposed as an alternative to the SDR measure from the BSS-eval toolbox [37]. Unlike the SDR, the SI-SDR is invariant to the scale of the processed signal. This metric is defined in (6) and can range from $-\infty$ to $+\infty$. We measured the quality of the enhanced speech using the Perceptual Evaluation of Speech Quality (PESQ) [38] and the intelligibility using the Short-Time Objective Intelligibility (STOI) measure [39]. PESQ was developed to estimate the quality perceived by humans in subjective tests such as the Mean Opinion Score (MOS) and it can range from -0.5 to 4.5. STOI predictions have shown a good correspondence with the measured intelligibility of noisy/processed speech in a large range of acoustic scenarios and it has a range of 0 to 1. It should be mentioned that for all these objective metrics, the greater the metric value, the better is the performance of the algorithm.

IV. Results

We evaluated the approaches, i.e. BESD and U-BESD, in two different settings. We first trained the networks to extract the first speaker using only the data where the subjects attended to the first speaker (subjects 1-17), and we then tested the trained network only on this part of the data. We refer to this setting as speaker-specific denoising. Secondly, we studied the performance of BESD and U-BESD in a speaker-independent setting, i.e., where no prior information is available about the attended speaker when performing the enhancement.

In the speaker-specific setting, we compared BESD and U-BESD to a denoising autoencoder [40] with a structure similar to U-BESD. For both settings, we also investigated whether using the FBC model would increase the performance by training BESD and U-BESD with MUA as the input. In the speaker-independent setting, we compared the performance of U-BESD EEG to a similar model by Ceolini *et al.* [19] by training the U-BESD with the dataset used in their study. Finally, we evaluated the network performance as a function of different parameters and structures.

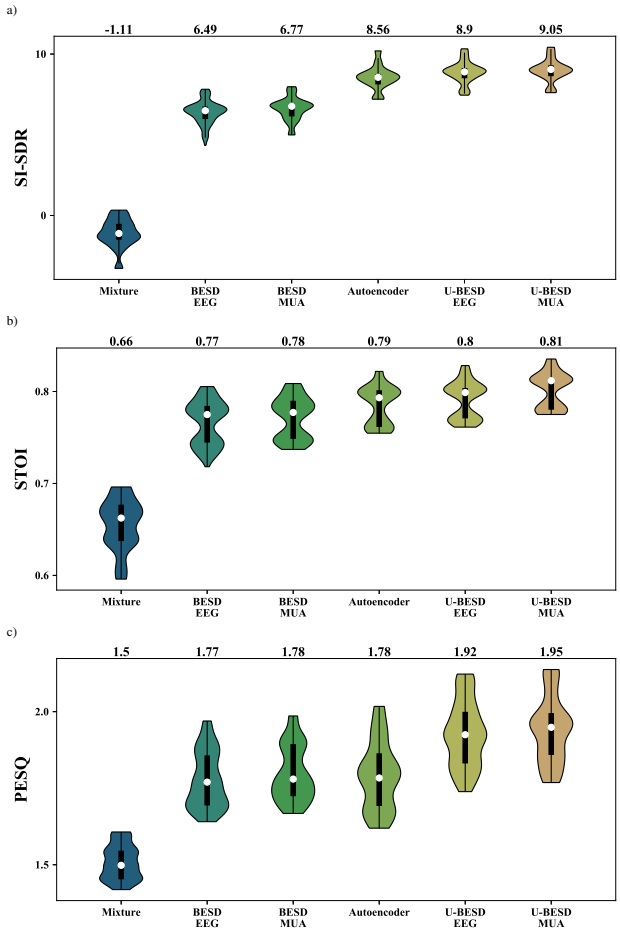

Fig. 2: Speech enhancement performance distributions for speaker-specific denoising using a) SI-SDR, b) STOI and c) PESQ metrics. We show the performance for the noisy mixture, BESD EEG, BESD MUA, denoising autoencoder, U-BESD EEG and U-BESD MUA. Medians are shown on top. The distribution of the values of the metrics are shown in the form of violin plots. The white dot in each plot shows the median. The black bar in the center of the violins shows the interquartile range (IQR). The thin black lines stretched from the bar show *first quartile* $-1.5 \times IQR$ and *third quartile* $+1.5 \times IQR$ respectively. Different colors are used to make distinction between each network easier. With either EEG or MUA, U-BESD significantly increases the performance compared to the BESD network and the noisy mixture ($p \ll 0.001$, Mann - Whitney U test). The denoising autoencoder has a similar structure to U-BESD leading to a better performance compared to BESD.

As mentioned in subsection III-E, we used the SI-SDR, STOI, and PESQ as objective metrics to evaluate the performance of the proposed approaches. We report the results in the form of violin plots, which represent the distribution of the performance over all the test segments for a given metric. In a violin plot, we can find the same information as the box plots. The white dot in the middle shows the median of the distribution. The black bars in the center of the violins show the interquartile range (IQR). The thin black lines stretched from the bar show *first quartile* $-1.5 \times IQR$ and *third quartile* $+1.5 \times IQR$ respectively. The advantage of using violin plots over box plots is that they also show the entire distribution of the data. This is useful especially when the data are multimodal, i.e., they have a distribution with more than one peak.

A. Speaker-specific denoising

In this section we present the results for the speaker-specific setting, where the attended speaker is known. The aim of this experiment is to study if using the U-BESD network increases the performance compared to the BESD and the denoising autoencoder network, even if the speaker is known to the algorithm. The results are shown in Fig. 2. For most subjects, we tested the performance for 15 nonoverlapping segments of 20 seconds leading to 220 estimation segments in total.

As can be seen, U-BESD has significantly less distortion (SI-SDR), better quality (PESQ), and higher intelligibility (STOI) than the speech mixture, BESD and the autoencoder ($p \ll 0.001$, Mann-Whitney U-test). Moreover, comparing the results obtained with U-BESD, with either EEG or MUA as the brain activity input, shows a significant improvement in the performance for the SI-SDR, STOI and PESQ metrics ($p < 0.001$, Mann - Whitney U test). The same can be said about the comparison between BESD EEG and BESD MUA ($p < 0.002$, Mann - Whitney U test).

In order to assess the possible performance difference between each subject, we also compared the overall performance of the network for each subject individually, in terms of SI-SDR, STOI, and PESQ. We observed (results not shown) that both BESD and U-BESD approaches had very similar performances for all subjects and all metrics ($p > 0.05$, Mann - Whitney U test). In fact, for the BESD network, the differences in metric medians between the best and worst subjects for SI-SDR, STOI, and PESQ are 0.22 dB, 0.03 and 0.07 respectively. For the U-BESD network, the differences in performance medians between the best and worst subjects for SI-SDR, STOI, and PESQ are 0.2 dB, 0.02 and 0.04 respectively.

B. Speaker-independent denoising

Next, we investigated the situation where no prior information about the target speaker is available during the enhancement. We trained the U-BESD to automatically extract the attended speaker, using the information present in the EEG signal or MUA, and to further perform the denoising. We compare the performance for both BESD and U-BESD and further analyze the performance differences between the speaker-specific and speaker-independent settings in Section V.

The results are shown in Fig. 3. For each subject, we evaluated the performance for 15 nonoverlapping segments of 20 seconds, leading to 441 estimation segments. As can be seen in Fig. 3, for all metrics, U-BESD has significantly less distortions (SI-SDR), better quality (PESQ), and intelligibility (STOI) than both the noisy speech mixture and BESD ($p < 0.03$, Mann - Whitney U test). Moreover, U-BESD MUA shows a significant improvement in performance ($p \ll 0.001$,

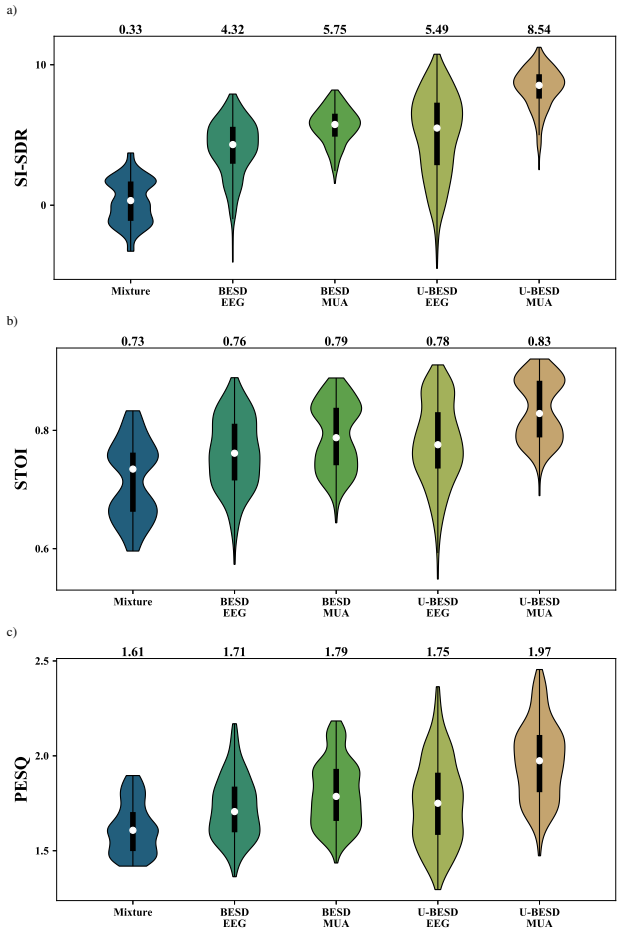

Fig. 3: Speech enhancement performance distributions for speaker-independent denoising using a) SI-SDR, b) STOI and c) PESQ metrics. We show the performance for the noisy mixture, BESD EEG, BESD MUA, U-BESD EEG and U-BESD MUA. Medians are shown on top. With either EEG or MUA, U-BESD significantly increases the performance compared to the BESD network and the noisy mixture ($p < 0.03$, Mann - Whitney U test).

Mann - Whitney U test) compared to U-BESD EEG. The same can be said about the comparison between BESD MUA and BESD EEG ($p \ll 0.001$, Mann - Whitney U test).

We also looked at the overall performance of the network for each subject individually, in terms of the SI-SDR, STOI, and PESQ metrics. The results are shown in Fig. 4 for BESD with MUA and Fig. 5 for U-BESD with MUA. For both BESD and U-BESD networks, it can be seen that generally, subjects 18-33, i.e., subjects who attended the right ear, performed better than subjects 1-17, i.e., subjects who attended the left ear.

Finally, we also compared the performance of the proposed network to that of [19]. To do so, we trained the proposed network using the same data as in [19], i.e., that of [41]. For details regarding the recording and preprocessing of the data please refer to [41]. The results are shown in Fig. 6. Please note that the results shown for [19] are taken directly from the paper. The results are shown for 27 test samples per subject. As can be seen from this figure, U-BESD performs significantly better than the BISS model proposed in [19].

C. Causal vs non-causal configuration

We also studied the performance of U-BESD MUA in the speaker-independent setting under a causal vs a non-causal configuration. Causal convolutions make sure that the model does not violate the order of the data. Meaning, the sample estimated at time step t does not depend on samples at time steps $t+1, t+2, ..., t+T$ [27]. Studying the difference between causal and non-causal convolutions are important because non-causal systems can only be used in applications that do not require real-time processing or low latency. To study the network in a non-causal setting, we used non-causal convolutions. It can be seen in Fig. 7 for all metrics, that using a causal setting significantly decreases the performance of the network ($p < 0.007$, Mann - Whitney U test).

D. Optimizing the network parameters

We further evaluated the effect of different network parameters and structures on the performance of U-BESD in the speaker-independent setting. We first evaluated several possible variations of the FiLM layer and the Concatenation block. We refer here to the structure presented in subsection II-B as Conv-Concat in which we use 1D convolutions in the FiLM block and the concatenation of sound mixture and brain activity input in the Concatenation layer. Firstly, instead of concatenating the sound and brain activity, we pass only sound to the decoder. Secondly, we use fully connected layers in the FiLM block instead of 1D convolutions. The combination of the 1D convolution without concatenation is called Conv-nConcat, fully connected layers with concatenation FC-Concat, and fully connected layers with no concatenation is called FC-nConcat. We also show the performance for networks with no dilated convolutions, Orthogonal weight initialization [42] instead of Glorot weight initialization, and Adam optimizer [43] instead of Adablief as well as a smaller network with 32 filters instead of 64. Furthermore, we also investigated changing the order of layers in the Conv Block from 1D convolution, Layer Norm, Leaky Relu, and dropout (CNRD) to 1D convolution, Leaky Relu, Layer Norm and dropout (CRND).

Table I shows the performance of the network for each of these parameters and structures in terms of the medians of SI-SDR, STOI, and PESQ distributions for the 441 segments of the speaker-independent setting. It also presents the total number of parameters in the network for each case (Model size). The first row of the table corresponds to the U-BESD used for the previous results (U-BESD MUA), which shows the higher performance for the causal settings.

We can make the following observations from Table 1:

(i) When a convolution layer is used in the FiLM block (Conv-Concat and Conv-nConcat), the performance of the model is generally better compared to when a fully connected layer (FC-Concat and FC-nConcat) is used. This is due to the fact that the 1D Convolutions have a kernel size of 3, while the fully connected layer is equivalent to a 1D Convolution with kernel size of

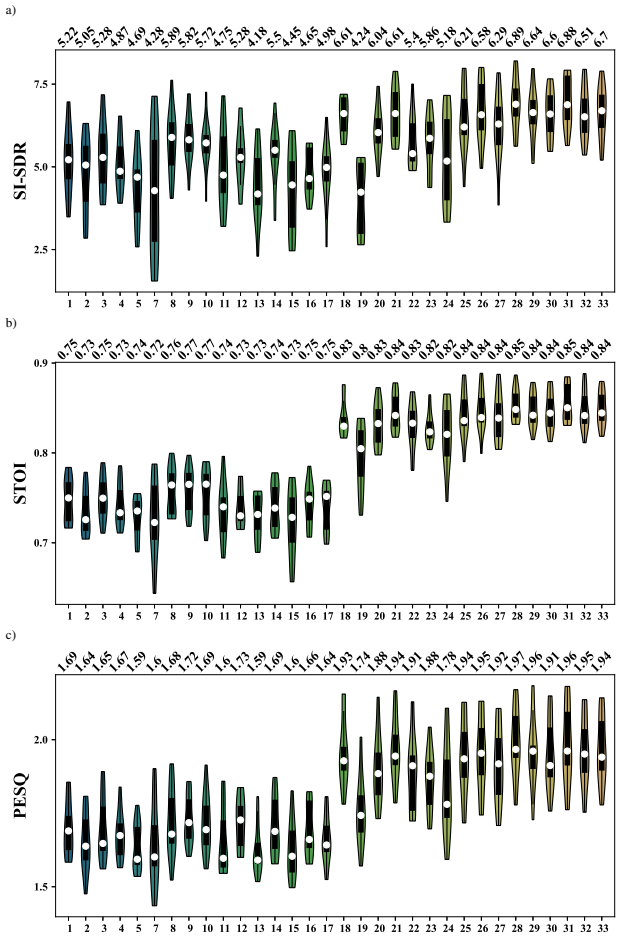

Fig. 4: Speech enhancement performance distributions for speaker-independent denoising for each subject using a) SI-SDR, b) STOI and c) PESQ metrics. We show the performance for BESD MUA. Medians are shown on top for each subject. Each subject was tested on 15 non-overlapping utterances of 20 seconds.

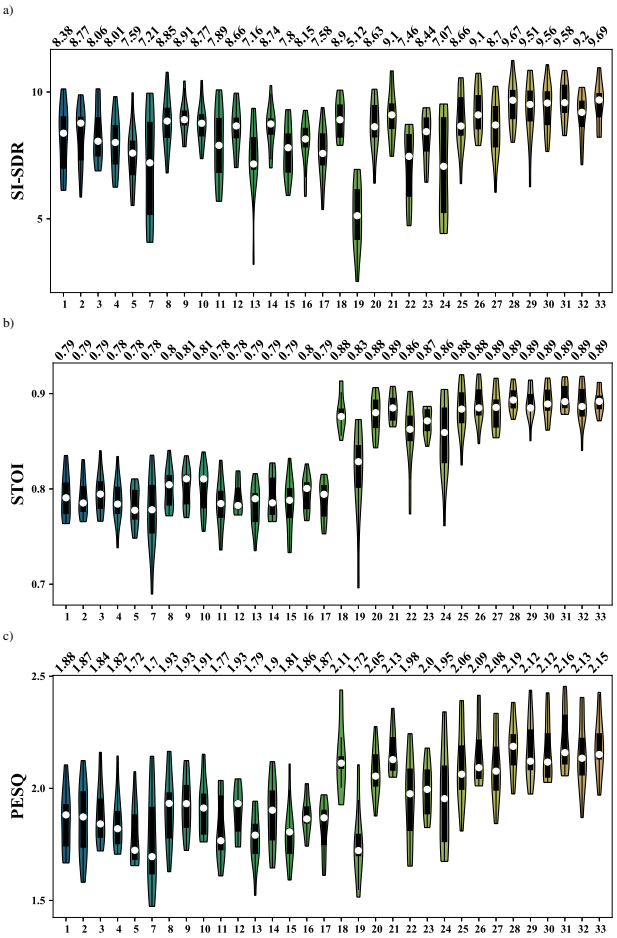

Fig. 5: Speech enhancement performance distributions for speaker-independent denoising for each subject using a) SI-SDR, b) STOI and c) PESQ metrics. We show the performance for U-BESD MUA. Medians are shown on top for each subject. Each subject was tested on 15 non-overlapping utterances of 20 seconds.

TABLE I: The effect of different configurations on the U-BSED performance in the speaker-independent setting.

Fusion type	Weight initialization	Optimizer	Dilation	Order	Nb of filters	Model size	SI-SDR	STOI	PESQ
Conv-Concat	Glorot	Adablief	Y	CNRD	64	1.84 M	**8.53**	**0.83**	**1.97**
Conv-nConcat	Glorot	Adablief	Y	CNRD	64	1.716 M	8.49	**0.83**	1.94
FC-Concat	Glorot	Adablief	Y	CNRD	64	1.712 M	7.62	0.81	1.84
FC-nConcat	Glorot	Adablief	Y	CNRD	64	1.6 M	7.48	0.8	1.84
Conv-Concat	Orthogonal	Adablief	Y	CNRD	64	1.84 M	8.42	0.82	1.92
Conv-Concat	Glorot	Adam	Y	CNRD	64	1.84 M	8.06	0.82	1.92
Conv-Concat	Glorot	Adablief	N	CNRD	64	1.84 M	6.58	0.8	1.82
Conv-Concat	Glorot	Adablief	Y	CRND	64	1.84 M	8.35	0.82	1.92
Conv-Concat	Glorot	Adablief	Y	CNRD	32	0.514 M	7.01	0.79	1.81

1. Therefore, besides having more trainable parameters, Conv-Concat and Conv-nConcat can learn local features of time signals, while the fully connected version constructs only instantaneous features.

(ii) When we concatenate the features learned from the brain activity and the sound mixture in the Concatenation layer (Conv-Concat and FC-Concat) the performance of the network is slightly better than when we pass only the features learned from the sound mixture to the decoder (Conv-nConcat and FC-nConcat). This could be due to the fact that when we pass both of the inputs to the decoder, we provide additional information to the network which increases the ability to extract the attended speaker and denoise the speech mixture. It should be mentioned

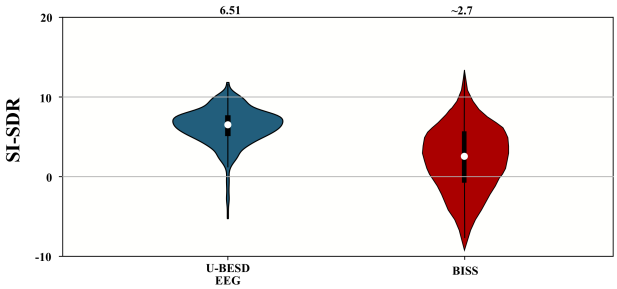

Fig. 6: Comparison between U-BESD EEG and the brain-inspired speech separation (BISS) model [19]. The results for the BISS model have been taken directly from [19].

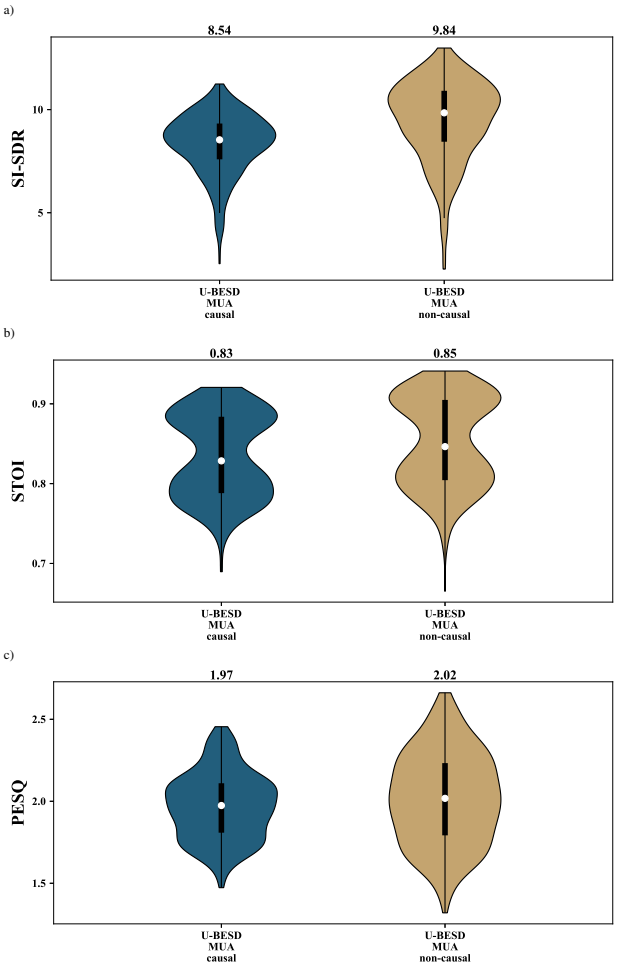

Fig. 7: Speech enhancement performance distributions for speaker-independent denoising for each subject using a) SI-SDR, b) STOI and c) PESQ metrics. We compare the performance for U-BESD MUA for causal vs. non-causal convolutions. Medians are shown on top for each condition.

that to keep other parameters in the network constant, Conv-Concat should be compared to Conv-nConcat and FC-Concat should be compared to FC-nConcat.

(iii) Using Glorot uniform [21] weight initialization improves the performance compared to an Orthogonal weight initialization [42]. Glorot uniform initialization draws samples from a uniform distribution with a variance scaled by the number of each layer's inputs and outputs. This helps the weights to maintain a reasonable range and avoid weight (and gradient) diminishing or exploding. On the other hand, Orthogonal initializer draws samples from a Gaussian distribution and creates an orthogonal weight matrix. The choice of weight initializer is very task dependent and theoretically, there is no reason why using the Glorot uniform initialization produces better results than using the Orthogonal initialization. Whether choosing another type of initializer leads to a better performance needs further investigation.

(iv) Using the Adablief optimizer compared to the Adam optimizer improves the performance of the network. The Adablief optimizer has been shown to have as fast a convergence as Adam, while having both a better generalization and a better training stability compared to Adam [36].

(v) Using dilated convolutions improves the performance of the network. The use of dilated convolutions allows the network's receptive field to grow exponentially without much computational cost. In this way, the network is able to model the long term temporal dependencies in the audio signals [27], [28].

(vi) Using Layer Normalization before Leaky Relu nonlinearity (CNRD) generally leads to a better performance compared to using Leaky Relu before Layer Normalization (CRND). The reason for this improvement however is not clear.

(vii) Increasing the number of filters improves the performance of the network. A larger number of filters lets the network learn more feature maps from the input, which leads to better estimations. However, increasing the number of filters drastically increases the number of parameters (0.541 M compared to 1.84 M) which could lead to overfitting. By using regularization and dropout layers, we aim to decrease the chance of overfitting in the network.

V. Discussion and Conclusion

It has been shown previously that the attended speaker can be decoded from the brain waves of the listener [15], [18]. In this paper, we proposed two networks called Brain Enhanced Speech Denoiser (BESD) and U-shaped Brain Enhanced Speech Denoiser (U-BESD), that use the brain signals of a listener in a noisy environment to perform speech enhancement and denoise a speech mixture. In a speaker-specific setting where the attended speaker is known, we show that U-BESD surpasses BESD as well as a denoising autoencoder with a similar structure. However, the denoising autoencoder performs better than the BESD for SI-SDR and STOI metrics. This is surprising as in this situation, all information about the attended speaker is available to the networks. The fact that the denoising autoencoder and U-BESD perform better than BESD could be due to factors such as:

- Whereas BESD has no skip connections, U-BESD and autoencoder have skip connections, both in the encoder

itself, similar to Resnet [26] and from the encoder to the decoder, similar to Unet [44]. Having both short and long skip connections has been shown to improve performance in deep neural networks for applications such as biomedical image segmentation [45].

- U-BESD and autoencoder both use dilated convolutions. As discussed in Table I, the use of dilated convolutions allows the network's receptive field to grow and model the long-term temporal dependencies in the audio signals [27], [28] and a network with no dilated convolutions, as BESD is, performs significantly worse.

As we just mentioned, the denoising autoencoder used here has a structure similar to the U-BESD, i.e, it uses skip connections and dilated convolutions, and this is why it performs better than BESD. However, if the autoencoder had a structure similar to BESD, then BESD would have produced better results than the denoisng autoencoder, as shown in [20]. The fact that both U-BESD and BESD have a better performance than an autoencoder with a similar structure to theirs indicates that providing the network with additional information, such as the brain activity, is always beneficial and increases the intelligibility and quality and decreases the distortions of the network estimations. By using the brain activity to modulate features learned from sound and vice versa, through FiLM blocks, the network is able to build more meaningful representations of the two inputs and thus performs better at extracting the attended speaker and denoising the speech mixture.

Moreover, the performance of the BESD and U-BESD networks for the speaker-specific task is similar for different subjects. This could be due to the fact that speaker-specific denoising is an easy task since the attended speaker is already known. As a result, the variation of EEG signal over different subjects is not affecting the performance of the network to a significant degree. On the other hand, for the task of speaker-independent denosing, which is admittedly a difficult task, we see significantly different variations over different subjects, as seen in Fig. 4 and Fig. 5. This indicates that for a speaker-independent setting, the performance of the network is affected by the EEG data recorded from each subject, in terms of either the distortions created by the subjects or their attention level. As a result, the network might need subject-specific training and, for a given application to be used by a specific subject, it would probably benefit from further training for that subject.

Although subjects do not perform similarly in the speaker-independent setting, we can see a trend where subjects that listened to the right ear perform better than the ones that listened to the left ear. The speaker for the left ear had a British accent and the speaker for the right had an American accent. Since the experiments were recorded in Ireland, where the accent is somewhat more similar to the British accent than to the American accent, we can speculate that it was easier for the participants to attend to the speaker with the British accent than to the speaker with the American accent and hence the recorded EEG had a better quality leading to a better network performance. Whether people listening to a speaker with different accents affects the EEG signals would however need further investigation.

In a speaker-independent setting where no prior information is available about the attended speaker, we show that U-BESD surpasses BESD as well as the noisy mixture. Additionally, it also surpasses a similar model by Ceolini *et al.* [19] when trained on the same dataset. Moreover, by comparing Fig. 2 and Fig. 3, we observe that the performance of the U-BESD MUA in the speaker-independent setting, while being slightly lower, is comparable to the speaker-specific setting. This shows that in the absence of any information about the attended speaker, U-BESD is able to use the brain signals to successfully extract the attended speaker with 100 percent accuracy, and to denoise the speech mixture.

Interestingly, using the MUA as opposed to the EEG as the input to the network increases the performance. This is because the EEG signals are a noisy mixture of several underlying sources. By using the FBC model to estimate MUAs instead of using directly the EEGs, the representation of the brain signal given to the network would be more meaningful and the network would then better extract the attended speaker [30].

In addition, we evaluated the difference in performance of the U-BESD MUA for a causal vs. a non-causal convolution setting. It can be seen from Fig. 7, that a non-causal setting has generally a higher performance. This is because the non-causal setting has information available from future input samples. This could make the non-causal setting unsuitable for real world applications since it introduces additional delays in the network. However, in the current setting and since the filters are only 25 samples long (corresponding to 1 ms at a sampling frequency of 14.7 kHz), the added samples from the future would only imply a 17 ms delay which could be suitable for many applications.

In conclusion, the best of the two proposed approaches, U-BESD, represents a significant step towards speaker-independent speech enhancement and denoising. U-BESD is an end-to-end approach where all modules of the algorithm are trained simultaneously in a single neural architecture, thus lowering the complexity of the algorithm. Moreover, the processing is performed completely in the time domain. Most importantly, given enough training data, no prior information about the attended speaker (i.e., the number of speakers or the target speaker) is needed. The combination of high accuracy, short latency, and small model size makes U-BESD a suitable choice for both offline and real-time, low-latency speech processing applications such as hearing aids and telecommunication devices, where no prior information about the attended speaker is available.

ACKNOWLEDGMENT

The authors would like to thank the authors of [29], [41] for kindly providing the data for this study. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

REFERENCES

[1] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust automatic speech recognition: a bridge to practical applications*. Academic Press, 2015.

[2] Y. Zhao, D. Wang, E. M. Johnson, and E. W. Healy, "A deep learning based segregation algorithm to increase speech intelligibility for hearing-impaired listeners in reverberant-noisy conditions," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1627–1637, 2018.

[3] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5554–5558.

[4] C. Xu, W. Rao, E. S. Chng, and H. Li, "Time-domain speaker extraction network," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 327–334.

[5] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1979, pp. 208–211.

[6] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[8] A. Bouzid, N. Ellouze *et al.*, "Speech enhancement based on wavelet packet of an improved principal component analysis," *Computer Speech & Language*, vol. 35, pp. 58–72, 2016.

[9] B. Lilly and K. Paliwal, "Robust speech recognition using singular value decomposition based speech enhancement," in *IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications (IEEE TENCON)*, 1997, pp. 257–260.

[10] Y. Hu and P. C. Loizou, "A subspace approach for enhancing speech corrupted by colored noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2002, pp. I–573.

[11] Y. Luo and N. 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.

[12] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

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

[14] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.

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

[16] 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," *Journal of Neural Engineering*, vol. 14, no. 5, p. 056001, 2017.

[17] A. Aroudi and S. Doclo, "Cognitive-driven binaural beamforming using EEG-based auditory attention decoding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 862–875, 2020.

[18] 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, 2015.

[19] 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*, p. 117282, 2020.

[20] M. Hosseini, L. Celotti, and É. Plourde, "Speaker-independent brain enhanced speech denoising," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1310–1314.

[21] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[22] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville, "Modulating early visual processing by language," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6594–6604.

[23] T. Kim, I. Song, and Y. Bengio, "Dynamic layer normalization for adaptive neural acoustic modeling in speech recognition," *arXiv preprint arXiv:1707.06065*, 2017.

[24] G. Ghiasi, H. Lee, M. Kudlur, V. Dumoulin, and J. Shlens, "Exploring the structure of a real-time, arbitrary neural artistic stylization network," *arXiv preprint arXiv:1705.06830*, 2017.

[25] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," *arXiv preprint arXiv:1709.07871*, 2017.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[28] 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, p. 1256–1266, Aug 2019.

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

[30] M.-A. Moinnereau, J. Rouat, K. Whittingstall, and E. Plourde, "A frequency-band coupling model of EEG signals can capture features from an input audio stimulus," *Hearing Research*, p. 107994, 2020.

[31] A. Delorme and S. 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.

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

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

[34] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. 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.

[35] B. C. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

[36] J. Zhuang, T. Tang, S. Tatikonda, N. Dvornek, Y. Ding, X. Papademetris, and J. S. Duncan, "Adabelief optimizer: Adapting stepsizes by the belief in observed gradients," *arXiv preprint arXiv:2010.07468*, 2020.

[37] C. Févotte, R. Gribonval, and E. Vincent, "BSS-EVAl toolbox user guide–revision 2.0," 2005. [Online]. Available: https://hal.inria.fr/inria-00564760/document

[38] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)- a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 749–752 vol.2.

[39] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. 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.

[40] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, 2013, pp. 436–440.

[41] S. A. Fuglsang, J. Märcher-Rørsted, T. Dau, and J. Hjortkjær, "Effects of sensorineural hearing loss on cortical synchronization to competing speech during selective attention," *Journal of Neuroscience*, vol. 40, no. 12, pp. 2562–2572, 2020.

[42] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[44] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[45] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep learning and data labeling for medical applications*. Springer, 2016, pp. 179–187.

Maryam Hosseini Graduated with a B.Eng. in Electrical Engineering from Ferdowsi University of Mashhad, Mashhad, Iran and a M.Eng in Biomedical Engineering from Amirkabir University of Technology, Tehran, Iran. She is currently working toward the Ph.D. degree in electrical engineering in the Computational Neuroscience and Intelligent Signal Processing Research Group, Université de Sherbrooke, Sherbrooke,QC. Her research interests include computational neuroscience, deep learning, and machine learning. Particularly, her research focuses on understanding the neural representation of an auditory stimulus in the auditory brain. Her goal is to study background noise features that exist in the signals recorded from the the auditory structures and use these in designing speech enhancement models.

Luca Celotti Received his B.Sc. and M.Sc. in Physics and Biophysics from the Universidad Autonóma de Madrid, Spain, in 2012, with an Erasmus at the Imperial College of London. He is pursuing a PhD at Université de Sherbrooke in Machine Learning and Computational Neuroscience, with emphasis in language generation and spiking networks.

Éric Plourde obtained joint B.Eng. (Electrical Engineering) and M.A.Sc. (Biomedical Engineering) degrees from the École Polytechnique de Montréal in 2002. After spending 2 years in the industry, he completed a Ph.D. in speech processing at McGill University in 2009. From 2009 to 2011, he was a post-doctoral researcher at the Neuroscience Statistics Research Laboratory with joint affiliations at the Massachusetts General Hospital, Harvard Medical School and the M.I.T. He is now a Full Professor in the Department of Electrical and Computer Engineering, Université de Sherbrooke, Quebec, Canada. His research deals mainly with speech enhancement, neural signal processing and auditory neuroscience.