HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2021 April 20.

Published in final edited form as:

Neuroimage. 2020 December ; 223: 117282. doi:10.1016/j.neuroimage.2020.117282.

Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception

Enea Ceolini[a,*], **Jens Hjortkjær**[b,c], **Daniel D.E. Wong**[d,e], **James O'Sullivan**[f,g], **Vinay S. Raghavan**[f,g], **Jose Herrero**[h], **Ashesh D. Mehta**[h], **Shih-Chii Liu**[a], **Nima Mesgarani**[f,g,*]

[a]University of Zürich and ETH Zürich, Institute of Neuroinformatics, Switzerland [b]Department of Health Technology, Danmarks Tekniske Universitet DTU, Kongens Lyngby, Denmark [c]Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital Hvidovre, Hvidovre, Denmark [d]Laboratoire des Systèmes Perceptifs, CNRS, UMR 8248, Paris, France [e]Département d'Études Cognitives, École Normale Supérieure, PSL Research University, Paris, France [f]Department of Electrical Engineering, Columbia University, New York, NY, USA [g]Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA [h]Department of Neurosurgery, Hofstra-Northwell School of Medicine and Feinstein Institute for Medical Research, Manhasset, New York, NY, USA

Abstract

Hearing-impaired people often struggle to follow the speech stream of an individual talker in noisy environments. Recent studies show that the brain tracks attended speech and that the attended talker can be decoded from neural data on a single-trial level. This raises the possibility of "neuro-steered" hearing devices in which the brain-decoded intention of a hearing-impaired listener is

[1]BISS: brain-informed speech separation.

[*]Corresponding authors. enea.ceolini@ini.uzh.ch (E. Ceolini), nima@ee.columbia.edu (N. Mesgarani).

Authors contribution

EC, DW, JeH and NM participated equally in the development of the idea. EC and NM were responsible for the speech separation. DW, JeH and VR were responsible for the EEG analysis. JO and NM were responsible for the iEEG analysis. JoH and AM were responsible for the iEEG data collection. EC created the figures, performed statistical analyses, wrote parts of the paper, and was responsible for the overall paper. SL provided critical feedback on the paper.

CRediT authorship contribution statement

Enea Ceolini: Conceptualization, Formal analysis, Writing - original draft, Writing - review & editing. **Jens Hjortkjær:** Conceptualization, Formal analysis. **Daniel D.E. Wong:** Conceptualization, Formal analysis. **James O'Sullivan:** Formal analysis. **Vinay S. Raghavan:** Formal analysis. **Jose Herrero:** Data curation. **Ashesh D. Mehta:** Data curation. **Shih-Chii Liu:** Supervision. **Nima Mesgarani:** Conceptualization, Formal analysis.

Ethics statement

For the EEG data collection all subjects provided written informed consent to participate. The experiment was approved by the Science Ethics Committee for the Capital Region of Denmark (protocol H16036391) and was conducted in accordance with the Declaration of Helsinki. For the iEEG data collection, all research protocols were approved and monitored by the institutional review board at the Feinstein Institute for Medical Research and informed written consent to participate in research studies was obtained from each subject before the implantation of electrodes.

Data and code availability statement

The data associated with this manuscript is divided in three parts:

• CODE: The code used to reproduce the results present in the paper can be accessed publicly at https://gitlab.com/Enny1991/brain-informed-source-separation

• EEG Data: the data associated with the EEG experiments has been collected during a previous study and can be accessed publicly on Zenodo at the following link https://zenodo.org/record/3618205#.XorgdC-w2_I

• iEEG Data: The data cannot be made public but can be requested from the author.

used to enhance the voice of the attended speaker from a speech separation front-end. So far, methods that use this paradigm have focused on optimizing the brain decoding and the acoustic speech separation independently. In this work, we propose a novel framework called brain-informed speech separation (BISS)[1] in which the information about the attended speech, as decoded from the subject's brain, is directly used to perform speech separation in the front-end. We present a deep learning model that uses neural data to extract the clean audio signal that a listener is attending to from a multi-talker speech mixture. We show that the framework can be applied successfully to the decoded output from either invasive intracranial electroencephalography (iEEG) or non-invasive electroencephalography (EEG) recordings from hearing-impaired subjects. It also results in improved speech separation, even in scenes with background noise. The generalization capability of the system renders it a perfect candidate for neuro-steered hearing-assistive devices.

Keywords

EEG; Neuro-steered; Cognitive control; Speech separation; Deep learning; Hearing aid

1. Introduction

Listeners suffering from hearing loss have difficulty following individual speakers in the presence of ambient and diffuse noise or competing speakers (Conn, 2006; Peelle and Wingfield, 2016). This problem, known as the cocktail party problem (Bregman and Pinker, 1978; Cherry, 1953), is seamlessly solved by normal hearing subjects, but represents a major challenge for the hearing impaired (HI). Automatic speaker separation and speech enhancement algorithms that can be implemented in hearing aid devices have seen tremendous progress in the past decade (Doclo et al., 2015; Gannot et al., 2017). Current speech separation solutions implemented in hearing aid devices are based on array signal processing and beamforming. However, because the microphones are typically placed on the hearing aid itself, the efficacy of the beamforming algorithms is limited by the small number of microphones and insufficient distance between them which is restricted by the size of the subjects head (Doclo et al., 2008). Distributed microphone arrays can solve this problem (Barfuss et al., 2016; Reindl et al., 2014; Schwartz et al., 2017; Zhao et al., 2014) but these solutions are not general and can only be implemented in special circumstances (Ceolini et al., 2020; Kiselev et al., 2017). Because of these limitations, many speech enhancement algorithms for hearing aid applications mainly aim to reduce simple ambient noises and to amplify the sound sources located in front of the user. This procedure, however, is ineffective in multi-talker acoustic environments where the source of noise is other speakers. Listening under such conditions remains the main complaint of hearing aid users.

The recent progress in deep learning (DL) methods has further advanced the state-of-the-art in speech enhancement and speaker separation. These methods have proven considerably more effective, particularly for single-channel speaker separation (Chen et al., 2017; Hershey et al., 2016; Luo and Mesgarani, 2019; Yu et al., 2017). The usability of these methods in hearing aid technologies, however, remains challenging for multiple reasons. First, the majority of these methods assume a fixed number of speakers in the mixed audio.

This inflexibility makes their performance unpredictable in real-world situations where speakers continuously appear and disappear from the acoustic scene. Second, speech separation still remains an unsolved problem when the number of interfering speakers increases and in unseen adverse acoustic environments. In these situations, augmenting the acoustic signal with other informative signals, such as video (Ephrat et al., 2018) or target speaker utterances (Wang et al., 2018; Xiao et al., 2019) has proven very fruitful. The use target speaker utterances is, nevertheless, limited by the necessity of prior knowledge about the identity of the speakers in the scene and this restricts its applicability. Third, separating all the sound sources in the acoustic scene is unnecessary in typical scenarios where the user is only interested in following a target speaker. Separating all speakers can significantly increase the computational cost which is particularly problematic in low-resource embedded applications. Finally, speech enhancement in hearing aid applications is impossible in multi-talker acoustic environments without knowing which speaker is the target and which speakers are interference. One proposed solution measures the listener's brainwaves to determine which speaker the listener wants to focus on (Clark and Swanepoel, 2014). This idea is based on the scientific discovery that the speech of an attended speaker is more strongly encoded in the listener's brain signals compared to background sources. This idea has been implemented using neural signals measured with noninvasive electroencephalography (EEG) (Horton et al., 2014; Kerlin et al., 2010; Power et al., 2012), magnetoencephalography (MEG) (Ding and Simon, 2012), and invasive intracranial EEG (iEEG) (Dijkstra et al., 2015; Golumbic et al., 2013; Mesgarani and Chang, 2012). The framework that uses neural signals to decode and enhance a target speaker in multi-talker speech perception is termed auditory attention decoding (AAD) (Fuglsang et al., 2017; O'Sullivan et al., 2017; Wong et al., 2018a). A typical AAD solution measures the similarity of the neural signals of a listener with each of the individual sound sources. As a result, previous implementations of AAD all started with automatic separation of the sound sources as the initial step. The speaker separation for AAD has been implemented with multichannel approaches, such as beamforming (Aroudi and Doclo, 2019; Van Eyndhoven et al., 2017), or single-channel approaches using neural network models (Han et al., 2019; O'Sullivan et al., 2017). Subsequently, the separated sound sources were compared with the neural signal to determine the target speaker. Performing the speaker separation and source selection steps independently, however, is sub-optimal due to the aforementioned limitations of current speaker separation algorithms.

Here, we address these issues by proposing a novel approach, brain-informed speech separation (BISS), that combines speaker separation and speaker selection steps of AAD. By jointly performing speech extraction and neural decoding, the neural signal directly guides a robust single channel speech extraction algorithm which is implemented using a neural network model. This method alleviates the need for a prior assumption of the number of speakers in the mixed audio and reduces the source distortion and computational load by extracting the target speaker from the scene. For these reasons, BISS represents a superior candidate for the implementation of a closed-loop, real-time, neuro-steered hearing aid (HA) which naturally adapts to different auditory scenes and number of competing sources.

2. Results

A schematic of BISS is depicted in Fig. 1. In this example, a listener hears two talkers and focuses on one of them (blue). The AAD system, indicated as the *brain decoder*, then decodes the envelope of the attended speech using brain signals (either EEG or iEEG). This decoded envelope (*hint*) is incorporated into a deep-learning-based speech separation algorithm to provide information regarding which of the signals in the acoustic scene has to be extracted. Finally, the enhanced speech of the desired speaker is amplified and delivered to the user thus closing the loop. The details regarding each step of this loop are given in the following paragraphs.

2.1. Brain recordings

2.1.1. EEG—The EEG data used in this work is a subset of the data described in (Fuglsang et al., 2020). EEG recordings from 22 normal hearing (NH) and 22 age-matched HI subjects were collected (NH: mean age 63.0 ± 7.1; HI: mean age 66.4 ± 7.0) after obtaining written consent. HI listeners had sloping high-frequency hearing-loss typical of presbycusis (age-related hearing loss). In 48 trials of ≈ 50 sec each, subjects listened to stories read by either a single talker (S-T) (16 trials), or multi talkers (M-T) (one male, one female, 32 trials). In the M-T trials, the two speech streams were presented at the same loudness level to allow unbiased attention decoding. The two competing speech streams were spatially separated at ± 90° using non-individualized head-related transfer functions (Oreinos and Buchholz, 2013). On each trial, the subjects were cued to attend to either the male or female talker and the attended target was randomized across the experiment. After each trial, the subjects responded to 4 comprehension questions related to the content of the attended speech. Both NH and HI listeners had accurate speech comprehension for both the single-talker (NH: 93.3%, HI: 92.3% correct) and two-talker conditions (NH: 91.9%, HI: 89.8% correct, see (Fuglsang et al., 2020) for details). Despite high accuracy on speech comprehension questions, listening difficulty ratings revealed that the HI listeners rated the two-talker condition as being significantly more difficult than NH listeners did (Fuglsang et al., 2020).

2.1.2. Ieeg—The iEEG data use in this study is the same used in (Han et al., 2019). It has been collected from three subjects undergoing clinical treatment for epilepsy at the North Shore University Hospital, New York. These subjects were implanted with high-density subdural electrode arrays covering their language dominant (left) temporal lobe with coverage over the superior temporal gyrus (STG) (Han et al., 2019). Similar to the EEG experiments, the subjects participated in two experiments, a S-T experiment and a M-T experiment. In both experiments, the subjects listened to stories read by two speakers, one male speaker and one female speaker. In the S-T experiment, the subjects listened to each speaker separately, and in the M-T experiment, the subjects listened to the two speakers talking concurrently with no spatial separation, i.e. the voices were rendered by a single loudspeaker placed in front of the subject. During the M-T experiment, each subject was presented with 11 minutes and 37 seconds of audio, making the S-T experiment twice as long. In the M-T experiment, the audio was separated into 4 blocks. In each block, the subject was asked to focus their attention on only one speaker. At the end of each block, the

subjects were asked to repeat the last sentence of the attended speaker to ensure that they were indeed paying attention to the correct speaker. All the subjects performed the task with high accuracy and were able to report the sentence with an average accuracy of 90.5% (S1, 94%; S2, 87%; and S3, 90%). We used the envelope of the high-gamma power at each site as our measure of neural activation.

2.1.3. Auditory attention decoding (AAD)—We reconstructed the speech envelope of the attended speaker from the raw data collected by EEG or iEEG. The decoder is a spatio-temporal filter that maps the neural recordings to the speech envelope. The mapping is learned using regularized linear regression and is based on the *stimulus reconstruction* method previously presented in (Akbari et al., 2019; Mesgarani et al., 2009; O'Sullivan et al., 2017; Wong et al., 2018a). For both the EEG and iEEG data, a subject-specific linear decoder is trained on S-T data and used to reconstruct speech envelopes on the M-T data. This approach was taken to avoid any potential bias introduced by training and testing on the M-T data. For the iEEG data, only the outputs of a subset of electrodes were used as input to the decoder. The electrode selection was done via a statistical analysis to determine whether a specific electrode is significantly more responsive to speech compared to silence.

2.2. Brain informed speech separation

2.2.1. Input and output features—We trained a speaker-independent speech separation neural network model using the brain signals of the listener to guide the separation. As illustrated in Fig. 1, the two inputs to the speech separation neural network are the noisy audio mixture and the *hint* represented by the attended speech envelope decoded from the listener's neural signals. The audio mixture $y(t)$ consists of the sum of the attended speaker $s_d(t)$ and all undesired sound sources $s_u(t)$ such as other speakers and noise, such that

$$y(t) = s_d(t) + s_u(t) \tag{1}$$

where t represents the time index. The time-frequency representation of this mixture $Y(l, f)$ can be obtained by taking the short-time Fourier transform (STFT) of $y(t)$,

$$Y(l, f) = STFT(y(t)) = S_d(l, f) + S_u(l, f) \tag{2}$$

where l and f are time and frequency bin indices respectively.

The complex mixture spectrogram $\mathbf{Y} \in \mathbb{C}^{F \times L}$ is compressed by a factor of 0.3 to reduce the dynamic range of the spectrogram (Ephrat et al., 2018)

$$\mathbf{Y}^c = (\mathbf{Y})^{0.3} \tag{3}$$

where $\mathbf{Y}^c \in \mathbb{C}^{F \times L}$.

The *hint* input comes from the temporal envelope of the clean speech of the attended speaker:

$$h(t) = \mid s_d(t) \mid^{0.3} \tag{4}$$

where we calculate the absolute value of the waveform, $s_d(t)$, and compress it by a factor of 0.3. During the training of the neural network model, the envelope is calculated from the clean audio signal. During testing, the reconstructed envelope of the attended speaker is used.

In order to extract the speech of the desired speaker from the mixture, the speech separation neural network model is trained to estimate a complex valued mask $\mathbf{M} \in \mathbb{C}^{F \times L}$. The estimated mask $\mathbf{M}$ is applied pointwise to the input STFT $\mathbf{Y}^c$

$$\widehat{\mathbf{S}}_d^c = \mathbf{M} \odot \mathbf{Y}^c \tag{5}$$

The resulting estimated spectrogram is decompressed and inverted to the time domain to obtain an enhanced version of the desired speech $\hat{\mathbf{s}}_d$.

$$\widehat{\mathbf{S}}_d = (\widehat{\mathbf{S}}_d^c)^3 \tag{6}$$

$$\hat{\mathbf{s}}_d = iSTFT(\widehat{\mathbf{S}}_d) \tag{7}$$

We refer the interested reader to Section 4.1.2 for a more detailed description of the model implementation.

2.2.2. Model architecture—Given the recent success of fully convolutional networks for speaker separation (Luo and Mesgarani, 2019), we propose an architecture that only uses 2D convolutions in contrast to the long-short term memory (LSTM) network used in (Ephrat et al., 2018). This architecture is inspired from (Luo and Mesgarani, 2019) but extended to a 2D convolution since the processing is performed in the time-frequency domain as shown in (Liu and Wang, 2019). The use of convolutional layers allows us to decrease the number of parameters in the model and to control the temporal length of the receptive fields.

The general architecture consists of a computational block that fuses the *hint* (see Section 4.1.2), with the mixture audio, followed by a stack of convolutional layers (van den Oord et al., 2016), each identical in its architecture and number of parameters thereby making the architecture modular. A final block applies the estimated complex mask $\mathbf{M}$ to the compressed input mixture spectrogram $\mathbf{Y}^c$ and inverts the estimated output spectrogram to the time domain. We investigate both the causal and non-causal settings of the model. The investigation of the causal setting is crucial if we want to be able to deploy the model to operate in real-time in practical applications.

2.2.3. Noise training scheme—The speech separation model is trained using a clean speech envelope calculated directly from the audio ground truth. However, the envelope estimated from either EEG or iEEG is not a perfect reconstruction of the original envelope.

Generally, the decoded envelopes have a Pearson's correlation r of < 0.3 for EEG data and about 0.6 for iEEG data. Because of this, it is important that the speech separation model is robust to a noisy *hint* envelope. We therefore estimate the distribution of the noise in the decoding process and extracted the variance of this noise for both EEG and iEEG data (see Appendix C). The noise has a Gaussian distribution with $\mu = 0$ and $\sigma_{iEEG} = 0.2$ for iEEG and $\sigma_{EEG} = 0.3$ for EEG.

After training the speech separation model with clean speech envelopes, we continued the training using a curriculum training technique (Braun et al., 2017) in which the amount of noise injected into the training data increased continuously for a number of epochs. This training schedule has been shown to be optimal for training a model that is robust to a large range of input signal-to-noise ratio (SNR)s. We use a schedule where the σ of the added noise increases in steps of 0.05 from [0.05, 0.6].

2.3. BISS: Attended speaker separation

2.3.1. Ieeg—We tested the BISS model on the iEEG recordings described in Section 2.1.2. For each subject, we report the violin plots of scale-invariant signal-to-distortion ratio (SI-SDR) (Roux et al., 2019) improvement (signal-to-distortion ratio (SDR) for brevity) from the noisy speech mixture obtained from testing the model with 4 s utterances (Fig. 2). We tested each subject on a set of 69 non-overlapping mixtures of two speakers and computed SDR improvements using the clean reference signal. The results presented in Fig. 2 show a comparable performance across all subjects. Subject 0 was the best with an SDR improvement of 9.5 dB; nevertheless, we did not find any significant difference between the scores of the three subjects. Additionally, the performance of causal and non-causal settings was similar for all subjects. One possible explanation for the similarity of performance across subjects is the noise training procedure in causal and non-causal settings. To test this hypothesis, we tested the performances of the causal and non-causal models using the noisy envelopes, like those used in training, rather than the neurally decoded envelopes as the *hint*. The test showed a decrease in performances gap between the causal and non-causal settings from an initial 1 dB to 0.5 dB. This shows that while there might be a large difference in performance between causal and non-causal settings when using clean envelopes, this difference decreases when using noisy envelopes. This explains the lack of significance between causal and non-causal settings in Fig. 2

Next, we show the effects of the noise curriculum training on the model performance when utilizing neural data. Fig. 3 shows the $r_{diff} = r_{attended} - r_{unattended}$ against SDR for the 69 utterances of Subject 0 attending to the male speaker in the mixture. The top panels show a density plot of the utterances together with their median value, while the bottom panels show every single utterance plotted separately and a linear fit of these points. The leftmost panels show the results for the model without any noise training while the other panels shows the effect of increasing the noise during training.

The top panels show that the median value shifts from below 0 dB, which indicates a failed separation, to above 9 dB, which indicates a very good separation (see Additional Data for audio samples). The bottom panels show that, independent of the noise level used in the training, there is a clear correlation between r_{diff} and the output SDR improvement. This

indicates that the quality of the separation is linearly dependent on the quality of the envelope reconstruction in terms of Pearson's r.

Finally, we show the effect of different Pearson's r values on the estimated mask **M**. In particular, we studied how the masks differ when we compare an utterance with high correlation to an utterance with low correlation. An example, depicted in Fig. 4, shows that the mask for the failed utterance (left) has less sharp edges around the harmonics of the desired speech, while for the successful utterance (right) the mask is sharp around every part of the desired speech and especially sharp around the harmonics. This is true even at smaller time scales where the sharpness of the mask tightly follows the correlation of the reconstructed envelope.

2.3.2. EEG—From the EEG dataset, we focused mainly on the differences between NH and HI groups. For each subject, we tested the performance on 128 non-overlapping segments of 4 seconds.

As in the iEEG case, we look at the differences in performance for the model under causal and non-causal settings. Fig. 5 shows the performance of causal vs non-causal settings for the two subjects groups. As expected, the overall performance is lower for EEG than with iEEG. As with iEEG, we found no significant difference between the causal and non-causal settings (p = 9.3e–01). Moreover, we found no statistical difference between NH and HI for the causal (p = 4.508e–01) and non-causal settings (p = 1.865e–01).

We then looked at the overall performance of each subject in terms of r_{diff} and SDR improvement. Fig. 6 shows the median SDR versus the median r_{diff} for all EEG subjects. Similar to iEEG, both groups show a clear and similar correlation between the r_{diff} and SDR. Overall, the EEG results show a positive correlation with a slope of 14.2 which is very close to the overall positive correlation of iEEG data which is 14.7.

Finally, we looked at the distribution of performance for each subject individually across the 128 utterances. We only considered trials in which the decoding of utterances was successful, i.e. with $r_{diff} > 0$. Fig. 7 shows the distribution and the median SDR performance for all individual subjects, ordered by increasing SDR. The difference in performance between the best and worst subjects is 4.6 dB, with the best and worst subjects having median SDRs of 6.8 dB and 2.2 dB, respectively.

3. Discussion and conclusion

We present a brain-controlled speech separation algorithm that uses the single-trial neural responses of a listener attending to a speaker to extract and enhance that speaker from the mixed audio. By utilizing the information provided by the envelope reconstruction algorithm, our method can extract the attended speaker from a mixture of two speakers as well as from speech-shaped background noise in the auditory scene, making it a viable solution for neuro-steered hearing aids (HAs).

Auditory attention decoding, which has been used with both EEG (Horton et al., 2014; Kerlin et al., 2010; Power et al., 2012) and iEEG (Dijkstra et al., 2015; Golumbic et al.,

2013; Mesgarani and Chang, 2012), generally assumes that the clean speech of the speakers in a mixture is available to be compared to the neural signals to determine the target source. This access to clean sources is not realistic in real-world applications. Recent work on AAD has tackled the problem of lack of access to clean sources (Han et al., 2019; Van Eyndhoven et al., 2017). Our work extends these studies by proposing a novel framework that combines the steps of speaker separation and speaker selection by turning speech separation into speech extraction. Not only does this framework readily generalize to competing speakers or background noise, it also requires significantly less computation (see Appendix A) because only the target speaker is extracted.

Moreover, we showed that speech separation quality (SDR) is highly correlated with stimulus reconstruction accuracy. This close correlation between these two quantities reveals two desired aspects of the proposed framework. First, it confirms our initial hypothesis that speech separation quality is higher in a model that takes additional information as input (see results in Appendix A), in this case the target speaker envelope reconstructed from the neural responses of the listener (Ephrat et al., 2018). Moreover, it offers a more general solution with respect to speaker extraction (Wang et al., 2018; Xiao et al., 2019) since the information about the target speaker can be obtained directly from the subject's brain on a trial-to-trial basis and does not have to be known a priori. Second, the speech separation quality of the model in the proposed framework follows the attention level of the subject which directly affects the reconstruction accuracy (r_{diff}) Mesgarani and Chang (2012), and thus reflects the intent of the subject. In closed-loop applications of AAD (Wong et al., 2018b), the separated target speech is typically added to the original mixed signal in order to both amplify the target speaker, but also to maintain the audibility of other sources to enable attention switching (usually 6–12 dB). Since BISS framework creates an output SDR which is correlated with the attention of the subject (r), this alleviates the need to render the mixture speech with a particular SNR since the SNR will naturally reflect the attention of the subject. This attention driven target SNR could help with attention switching in closed-loop applications (Geirnaert et al., 2020; Han et al., 2019).

The results obtained from applying AAD to EEG data are similar to the results obtained with iEEG but with smaller Pearson's r of the reconstructed envelope and lower SDR of separated speech. Even though these results are less accurate, they are in accordance with the predictions made using iEEG for AAD. In particular, the r_{diff} and the output SDR are highly correlated, confirming again that the model follows the subject's attention. Moreover, the AAD results using EEG show no significant difference in target speech enhancement (SDR) between HI and NH subjects. This shows that the proposed BISS can be used by HI subjects, which is a crucial aspect for the applicability of the framework to neuro-steered HAs.

Additionally, it is worth noting that the same speech separation model was used to produce the results presented from both iEEG and EEG. This shows the versatility of the proposed approach. Not only can the framework be applied successfully in the presence of different languages (Danish and English in this case) and noise (see Appendix B), but it is also unaffected by different methods of reconstruction and different types of brain signals used. These findings suggest that the BISS approach is a robust speech separation front-end.

Moreover, the finding that BISS results in no significant difference between causal and non-causal speech separation models increases its usability in real-time systems which require causal, short-latency implementation ($< 20ms$). Finally, we showed how BISS can decouple the optimization of front-end (speech separation) and back-end (AAD) systems even when a small amount of data is available. This joint optimization can also be done when large amounts of data are available.

While our method uses basic neural signal decoding (speech envelope reconstruction), there are many other ways this can be improved, for example, by reconstructing the speech spectrograms (Akbari et al., 2019). Moreover, the neural decoding can be done either with classification (de Cheveigné et al., 2018) or state space models (Miran et al., 2018). These methods can be easily integrated into the BISS framework because it takes as the *hint* any signal that is correlated with the attended speech.

By addressing the real-world constraints of AAD and speech separation, BISS represents a promising step towards the real world implementation of neuro-steered HAs.

4. Materials and methods

In this section we first describe the details about the EEG and iEEG data that has been used for the AAD part of BISS. We then describe the neural network model which represents the speech separation front-end. In particular, we describe the network architecture and the training scheme which are unique to BISS.

4.1. Processing of brain signals

EEG recordings were performed in an electrically shielded double-walled sound booth at the Technical University of Denmark (DTU). EEG was recorded using a BioSemi ActiveTwo system with 64 scalp electrodes. The stimuli were presented via ER-3 insert earphones (Etymotic Research). Preprocessing of the EEG data included re-referencing to the average of two posterior electrodes (TP7, TP8), band-pass filtering between 1 and 9 Hz, down-sampling to 64 Hz, and removal of eye-blink artefacts. For full details on the data collection and preprocessing refer to (Fuglsang et al., 2020).

For iEEG, more information about the data collection process and preprocessing is provided in (Han et al., 2019).

4.1.1. Audio signals processing—For the input, we use 4 seconds of audio that is transformed to the frequency domain with a STFT using a window size of 512 and a step size of 125. The choice of the length in time (4 seconds) is completely arbitrary and, differently from (Luo and Mesgarani, 2018b), was not changed during the training process. The choice of 125 was made because the audio sampling rate is 8 kHz and we want an output rate of 64 Hz that matches the envelope sampling rate. Because of the Hermitian property of the Fourier transform on real data, we can keep only the positive frequencies of the transformed signal thus obtaining as input a 3D tensor of size $2 \times 257 \times 257$.

For the output mask we used a complex-valued mask instead of a real-valued magnitude mask. Using a real-valued magnitude mask forces the use of the noisy phase when inverting the estimated separated spectrogram to the time domain, and it was shown that using the compressed complex mask gives better results (Ephrat et al., 2018). Because we use a complex STFT with overlapping windows, there exists an ideal complex mask that perfectly isolates the desired source (Williamson et al., 2016) from the mixture. Unfortunately, the mask values can be arbitrarily high and unbounded, and this poses a problem for the training process. For this reason, we use a hyperbolic tangent compression that limits the output mask values to the range [−1, 1]. Therefore we can only compute an approximation of the ideal mask.

4.1.2. Detailed model architecture—The *hint fusion* (panel d in Fig. 8) consist of two different processing steps that allow us to concatenate the audio waveform of the mixture $\mathbf{Y}^c$ with the desired speech envelope $H(l)$. First, the mixture waveform is transformed in the frequency domain by means of the STFT. The real and imaginary parts are then concatenated along a new axis effectively producing a 3D tensor of size $2 \times F \times L$. A 1×1 2D convolution with C feature maps is then applied to obtain a 3D tensor of shape $C \times F \times L$. Similarly, the desired speech envelope is processed with a 1×1 1D convolution and expanded to become a 3D tensor of shape $1 \times F \times L$. Finally, the two tensors are concatenated along the feature map axis to obtain a 3D tensor of shape $(C + 1) \times F \times L$

The network has S stacks (panel c in Fig. 8). Each stack, indexed with s, is composed of multiple blocks. Each block (panel b in Fig. 8), indexed with i, is a residual block that receives two inputs: the skip connection ($\mathbf{r}$) from the input and the output ($\mathbf{o}$) of the previous block. As shown in Equations (8) and (9), the skip connection is the sum of the input plus the output of each convolutional step, while the output of the block is the output of the convolution summed with the residual connection of the current input.

$$\mathbf{p}_i^s = \mathbf{c}_i^s + \mathbf{s}_{i-1}^s \tag{8}$$

$$\mathbf{o}_i^s = \mathbf{o}_{i-1}^s + \mathbf{c}_i^s \tag{9}$$

The skip input to the first block in a stack is a matrix of zeros, while the output of the last block, and thus of the stack, is the skip path (left part of panel b in Fig. 3).

Each block contains a convolutional step (panel a in Fig. 8) which has the same architecture for all blocks but has a different dilation factor defined by the block index i. In particular, the dilation factor for block i will be 2^i. The convolutional step has three parts, as shown in Equations (10) to (12): a 1×1 convolution followed by a ReLU non-linearity, a 3×3 convolution with dilation factor i followed by a ReLU non-linearity, and finally another 1×1 convolution:

$$\mathbf{b}_{i,1}^s = ReLU(conv_{i,1}(\mathbf{o}_{i-1})) \tag{10}$$

$$\mathbf{b}_{i,2}^{s} = ReLU(conv_{i,2}(\mathbf{b}_{i,1}^{s})) \tag{11}$$

$$\mathbf{p}_{i,3}^{s} = \mathbf{c}_{i}^{s} = conv_{i,3}(\mathbf{b}_{i,2}^{s}) \tag{12}$$

The final convolutional step is utilized to get back the same input shape which allows the residual and skip connections to be added. This step increases the total number of parameters in the network without increasing the receptive field. Batch norm is applied at the end of the convolutional step. Overall the receptive field (RF) in both frequency and time can be calculated as follows: 13

$$RF(N, S, k) = k + S\sum_{i=0}^{N}(k-1)2^{i} \tag{13}$$

where k is the kernel size.

We use square kernels so the receptive fields have the same dimension in both the frequency and time domain in terms of bins but are different in terms of meaning and measure.

In the last step of the process, as depicted in the *mask* step (panel e) of Fig. 8, the output of the last stack, $\mathbf{o}_{N}^{S}$ is reshaped by a 1 × 1 convolution from a shape of $(C+1) \times F \times L$ to a shape of $2 \times F \times L$, where the first dimension represents the concatenation of real and imaginary parts.

The mask, **M** is obtained by first applying an hyperbolic tangent to the output of that convolution and then summing real and imaginary parts properly:

$$\widetilde{\mathbf{M}} = tanh(conv(\mathbf{o}_{N}^{S})). \tag{14}$$

$$\mathbf{M} = \widetilde{\mathbf{M}}(0, :, :) + i\widetilde{\mathbf{M}}(1, :, :) \tag{15}$$

where the operation $(j, :, :)$ represents the tensor slicing that selects only the j^{th} element in the first tensor dimension and i represents the imaginary unit. Finally, as shown by Eq. 5, the mask will be used to separate the desired speech.

The model is relatively simple and has very few parameters, around half a million for the version used to obtain the results presented here.

4.2. Training scheme

The BISS system in this work (Fig. 1) uses the decoded speech envelope from the brain decoder as the informed input to the speech separation network. Ideally, one would train the neural network with the brain-decoded envelopes. Unfortunately, the EEG and iEEG data collected for attention decoding typically amounts to less than one hour of data for each subject. This amount of data is not enough to train an accurate speech separation model

which has millions of parameters. Such a model would require in the order of tens of hours of recorded speech (Hershey et al., 2016; Luo and Mesgarani, 2018, 2019).

To address this problem, we decouple the training of the speech separation model from the training of the brain decoder model. The separately trained models are then fused at test time. In order to do this, the speech separation model is trained with the ground truth speech envelope extracted from the audio using same envelope calculation as the one used for the attention decoding model. This guarantees that the attention decoding model will provide an envelope which is most correlated with the desired speech to extract.

Most of the EEG data was collected in Denmark using Danish audiobooks while the iEEG data was collected in New York using English audiobooks. Since we propose to use a single model to extract desired speech from either EEG or iEEG, the training dataset for the speech separation model consists of a mixture of English and Danish utterances (i.e. the model is not language-specific). The English material used for training are the Wall Street Journal (WSJ) utterances in the WSJ-mix2 dataset often used for source separation benchmarks (Hershey et al., 2016; Luo and Mesgarani, 2018b). The Danish utterances are taken from Danish audiobooks used for the EEG study (Fuglsang et al., 2020). Note that the training data is completely separated from the testing data, i.e. the audio tracks used in the attention decoding for both EEG and iEEG are not part of the training dataset. The overall training dataset comprises 22 hours of data. We create mixed sentences on-the-fly at training time as a data augmentation method to effectively increase the amount of data used in training.

When estimating the frequency-domain masks for speech separation, the mean squared error (MSE) is generally used as the cost function. However, the estimated masks are usually smeared, limiting the separation quality (Wang and Chen, 2018). In this work, we propose to use a time-domain optimization method with a frequency domain solution (Ceolini and Liu, 2019) by embedding both the STFT and iSTFT procedure into the training pipeline. Because these operations are differentiable, we can use the normal backpropagation algorithm to train the model. The cost function used to optimize the model is SI-SDR. Optimizing the SI-SDR has shown very good results in time domain separation (Luo and Mesgarani, 2018, 2019) due to the fact that the model directly optimizes the measure which is used to evaluate its performance (Roux et al., 2019). The SI-SDR metric (SDR for simplicity) can be calculated directly from the time domain signals as follows:

$$\mathbf{s}_{target} = \frac{\langle \hat{\mathbf{s}}_d, \mathbf{s}_d \rangle \mathbf{s}_d}{\|\mathbf{s}_d\|^2} \tag{16}$$

$$\mathbf{e}_{noise} = \hat{\mathbf{s}}_d - \mathbf{s}_{target} \tag{17}$$

$$SI - SDR = 10 \log_{10} \frac{\|\mathbf{s}_{target}\|^2}{\|\mathbf{e}_{noise}\|^2} \tag{18}$$

The neural network model is trained using the Adam optimizer with default settings (Kingma and Ba, 2014) and early stopping as a regularizer.

4.3. Statistical analysis

Where relevant, we report the statistical significance based on a two-tailed Mann - Whitney *U* test. Significance is indicated by ns if 5.00e–02 < $p \leq$ 1.00e+00, * if 1.00e–02 < $p \leq$ 5,00e–02, ** if 1.00e–03 < $p \leq$ 1.00e–02. *** if 1.00e–04 < $p \leq$ 1.00e–03 and **** if $p \leq$ 1,00e–04. All the violin plots and box plots in the results of Section 2 report median, interquartile range (IQR) together with minimum and maximum values.

Acknowledgments

Funding

This work was partially supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 644732, the SNSF grant No. 200021172553, the grant from the National Institutes of Health NIDCD-DC014279, and the National Science Foundation CAREER award.

Appendix

Appendix A. Informed speech separation (ISS)

We compare our informed speech separation (ISS) approach with permutation invariant training (PIT) which is the state-of-art method for training deep-learning based speech separation models. In this case, the *hint* input in our model comes from the envelope of the ground truth attended speech. The results in Fig. A1 show that ISS gives significantly better results (p = 7.8461e–09) than PIT for the causal setting. In contrast, the non-causal setting results show no significant difference (p = .1101) between ISS and PIT. The ISS algorithm produces significantly better results under non-causal settings (p = 5.0211e–09) over causal settings. The causal setting gives an absolute median difference of ≈ 0.9 dB a value that still indicates good separation quality for practical applications.

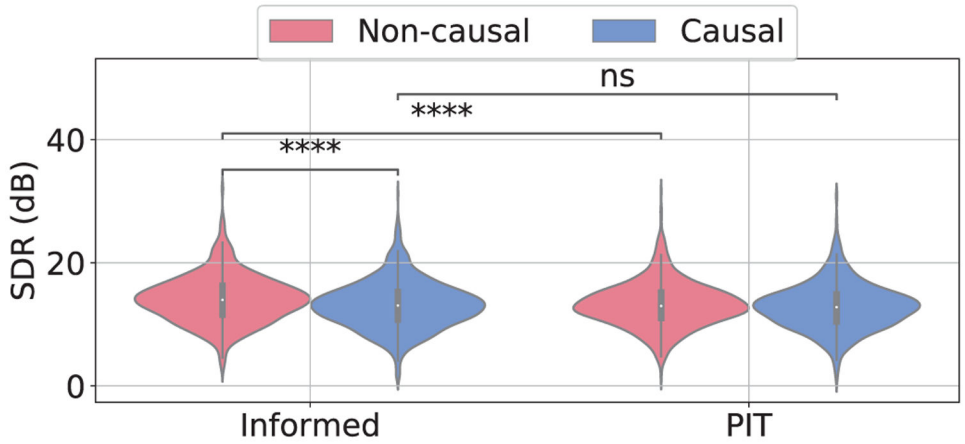

Fig. A1.
Comparison of ISS and PIT using the desired speech envelope as the hint for ISS. ISS: median = 13.0 dB (causal), median = 13.92 dB (non-causal). PIT: median = 12.79 dB

(causal), median = 12.98 dB (non-causal). Significance is indicated by **ns** if 5.00e–02 < $p \leq$ 1.00e+00, * if 1.00e–02 < $p \leq$ 5.00e–02, ** if 1.00e–03 < $p \leq$ 1.00e–02, *** if 1.00e–04 < $p \leq$ 1.00e–03 and **** if $p \leq$ 1.00e–04 using Mann - Withney U test.

Note that the model trained with PIT has around 1 million parameters and the model size scales almost linearly with the number of speakers in the mixture. On the other hand, the ISS model has only 0.5 million parameters and this number does not have to scale with the number of speakers in the mixture. Similarly, the number of operations to compute one spectrogram column mask is around 14 MOps for the PIT model and 7 MOps for the ISS model which makes the ISS model cheaper to compute for real-time applications. The number of parameters and number of operations are calculated based on the final settings of the model chosen for the best trade-off between size and performance (4.1.2). The final settings are shown in Table 1 and give rise to a receptive field with a span of 3.9 s in time and a span of 7900 Hz in frequency.

Appendix B. Speech in noise with EEG

To show that the BISS framework can successfully be applied across tasks of speaker separation and speech enhancement, we also looked at the possibility of reducing noise in attended speech using EEG signals. This is an easier task to solve than speaker separation. Mainly, this is due to the fact that the noise and speech have different frequency distributions and are easier to separate than 2 overlapping speakers. In particular, speech enhancement models that use neural networks can easily be trained without the need to use PIT: if one assumes one only speaker, there is no confound to resolve on which is the desired signal to extract. In this appendix, we show the results for BISS applied to speech enhancement using EEG. We used EEG recorded from a NH subject listening to speech in stationary speech-shaped background noise (data from (Hjortkjær et al., 2020)). The network is the same used above but it is trained with more added noise in the input, with respect to the model used for speaker separation. The *hint* to the network is still the envelope reconstructed from the EEG of the subject. Fig. B1 show the results for a particular subject and shows that the training scheme is effective in increasing the robustness of the network to the non-perfect reconstructed envelope. As we can see, compared to iEEG in speaker separation, even a low amount of noise helps the network in making use of the hint to separate the desired voice. Moreover we can see from the Fig. B2 that the method is successfully applied to all the subjects. Differently from the speaker separation task, we can see that for speech enhancement, the linear trend between Pearson's r and output SDR is less evident than the one present for speaker separation. This is due to the fact that the task is much easier to solve and that even a reconstructed envelope with a low reconstruction quality is informative enough for the model to separate the desired speaker.

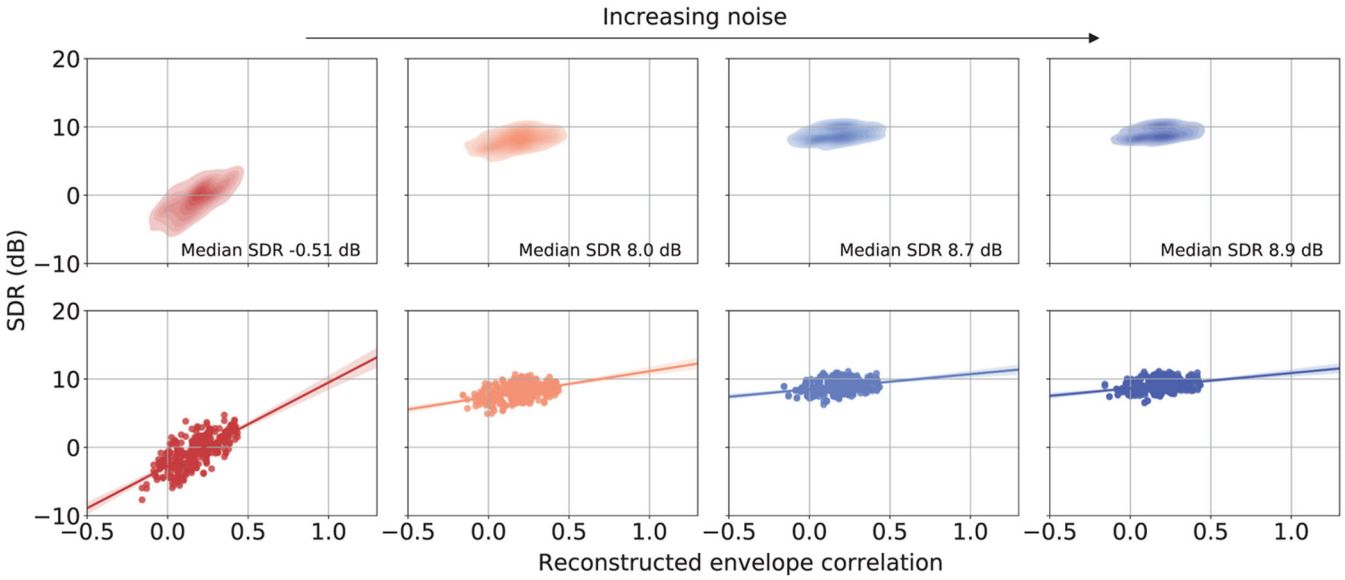

Fig. B1.

Results from Subject 23 in the EEG recordings. The x-axis indicates r_{speech}, y-axis indicates SDR in dB. The panels in the top row show the density distribution of the points using kernel density estimate with Gaussian kernels. The panels in the bottom row show each utterance separately and a linear fit obtained using linear regression. The shaded area represents the 95% confidence interval of the regression. The panels from left to right show results from increasing the σ of the noise during training (from σ = 0.0 to σ = 0.6 with steps of 0.2).

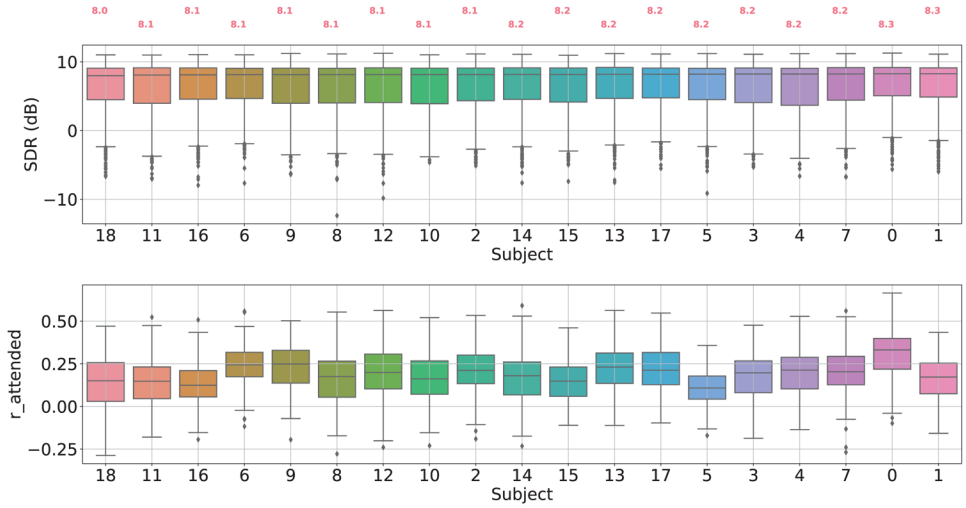

Fig. B2.

Performance for all EEG subjects considering only valid trials (r_{diff} > 0.0) on the speech in noise task. Median values for SDR are highlighted above the top panel.

Appendix C. Noise distribution in AAD

To make the speech separation model more robust to the degraded quality of the envelope reconstructed from the brain signals, we employ a training scheme known as curriculum learning (Braun et al., 2017). This scheme consist in increasing progressively, over training epochs, the difficulty of the task by introducing progressively more noise in the training. In order for this scheme to be effective, one needs to ensure that the noise injected during training is of the same distribution of the noise that will be present at test time. Here, to justify the choice of the training scheme used in Section 4, we show the empirical distribution of the noise in the reconstructed envelope, which is represented by the error between the original envelope and the envelope reconstructed with AAD. This is exactly the noise that the network will be faced with when trained with the clean envelope and tested with the (noisy) reconstructed one. Fig. C1 shows the distribution of error for both EEG and iEEG. As expected, the distribution of error for the EEG reconstruction has a bigger standard deviation with respect to the standard deviation of the iEEG reconstruction error. This follows the results showing that the quality of the reconstruction in higher for the iEEG data.

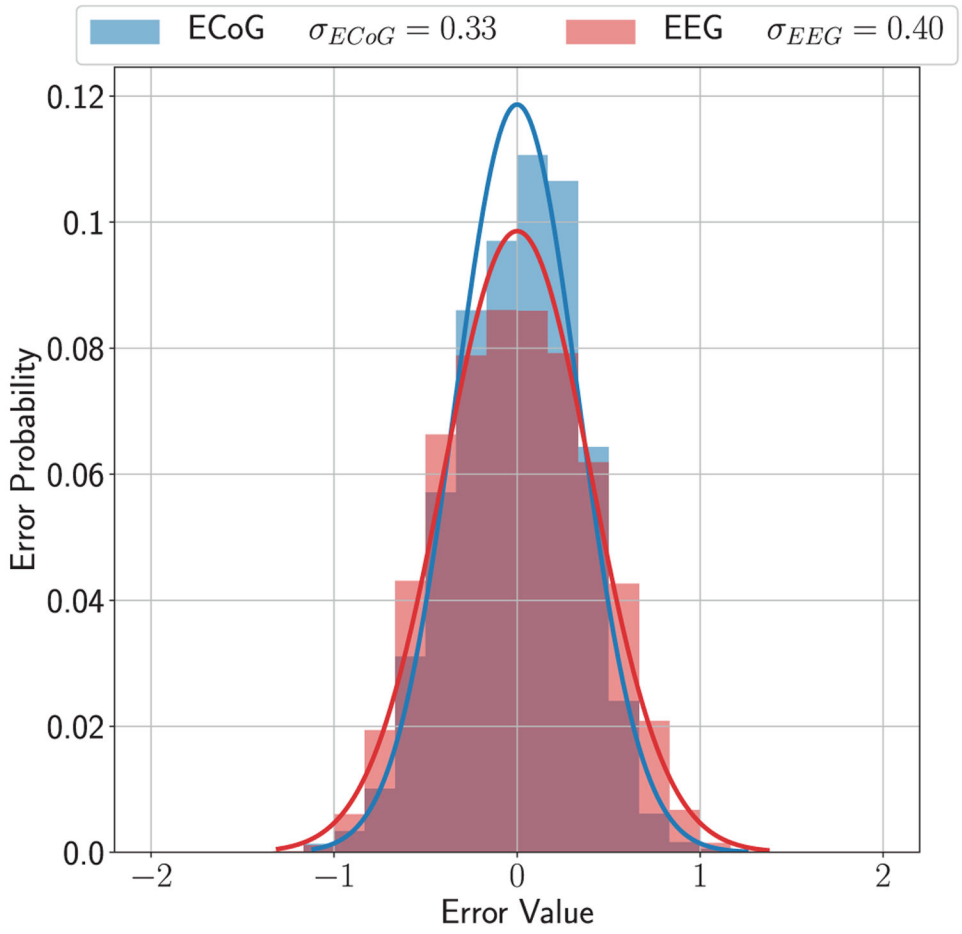

Fig. C1.

Distribution of errors between the reconstructed attended envelope and the original attended envelope for both EEG and iEEG.

References

Akbari H, Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N, 2019. Towards reconstructing intelligible speech from the human auditory cortex. Sci. Rep 9 (1), 874, [PubMed: 30696881]

Aroudi A, Doclo S, 2019. Cognitive-driven binaural lcmv beamformer using eeg-based auditory attention decoding. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 406–410,

Barfuss H, Mueglich M, Kellermann W, 2016. HRTF-based robust least-squares frequency-invariant polynomial beamforming. In: 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 1–5.

Braun S, Neil D, Liu S-C, 2017. A curriculum learning method for improved noise robustness in automatic speech recognition. In: 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, pp. 548–552,

Bregman AS, Pinker S, 1978. Auditory streaming and the building of timbre.. Canadian Journal of Psychology/Revue canadienne de psychologie 32 (1), 19.

Ceolini E, Kiselev I, Liu S-C, 2020. Evaluating multi-channel multi-device speech separation algorithms in the wild: a hardware-software solution. IEEE/ACM Trans Audio Speech Lang Process 28, 1428–1439.

Ceolini E, Liu S, 2019. Combining deep neural networks and beamforming for real-time multi-channel speech enhancement using a wireless acoustic sensor network. In: 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6.

Chen Z, Luo Y, Mesgarani N, 2017. Deep attractor network for single-microphone speaker separation. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 246–250. doi:10.1109/ICASSP.2017.7952155.

Cherry EC, 1953. Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am 25 (5), 975–979.

de Cheveigné A, Wong DD, Di Liberto GM, Hjortkjaer J, Slaney M, Lalor E, 2018. Decoding the auditory brain with canonical component analysis. Neuroimage 172, 206–216. [PubMed: 29378317]

Clark JL, Swanepoel DW, 2014. Technology for hearing loss–as we know it, and as we dream it.. Disability and rehabilitation. Assistive technology 9 5. 408–13 [PubMed: 24712413]

Conn PM, 2006. Handbook of models for human aging. Academic Press.

Dijkstra K, Brunner P, Gunduz A, Coon WG, Ritaccio AL, Farquhar J, Schalk G, 2015. Identifying the attended speaker using electrocorticographic (ecog) signals.. Brain computer interfaces 2 4, 161–173. [PubMed: 26949710]

Ding N, Simon JZ, 2012. Emergence of neural encoding of auditory objects while listening to competing speakers.. Proceedings of the National Academy of Sciences of the United States of America 109 29. 11854–9 [PubMed: 22753470]

Doclo S, Gannot S, Moonen M, Spriet A, 2008. Acoustic beamforming for hearing aid applications. Handbook on array processing and sensor networks 269–302.

Doclo S, Kellermann W, Makino S, Nordholm S, 2015. Multichannel signal enhancement algorithms for assisted listening devices: exploiting spatial diversity using multiple microphones. IEEE Signal Process. Mag 32, 18–30.

Ephrat A, Mosseri I, Lang O, Dekel T, Wilson K, Hassidim A, Freeman WT, Rubinstein M, 2018. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. ACM Trans. Graph 37, 112:1–112:11.

Fuglsang S, Märcher-Rørsted J, Dau T, Hjortkjær J, 2020. Effects of sensorineural hearing loss on cortical synchronization to competing speech during selective attention.. J. Neurosci.

Fuglsang SA, Dau T, Hjortkjær J, 2017. Noise-robust cortical tracking of attended speech in real-world acoustic scenes. Neuroimage 156, 435–444. [PubMed: 28412441]

Gannot S, Vincent E, Golan SM, Ozerov A, 2017. A consolidated perspective on multimicrophone speech enhancement and source separation. IEEE/ACM Trans. Audio Speech Lang. Process 25, 692–730.

Geirnaert S, Francart T, Bertrand A, 2020. An interpretable performance metric for auditory attention decoding algorithms in a context of neuro-steered gain control. IEEE Trans. Neural Syst. Rehabil. Eng 28 (1), 307–317. doi:10.1109/TNSRE.2019.2952724. [PubMed: 31715568]

Golumbic EMZ, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE, 2013. Mechanisms underlying selective neuronal tracking of attended speech at a cocktail party. Neuron 77 (5), 980–991. doi:10.1016/j.neuron.2012.12.037. [PubMed: 23473326]

Han C, Luo Y, Mesgarani N, 2019. Online deep attractor network for real-time single-channel speech separation. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 361–365. doi:10.1109/ICASSP.2019.8682884.

Han C, O'Sullivan J, Luo Y, Herrero J, Mehta AD, Mesgarani N, 2019. Speaker-independent auditory attention decoding without access to clean speech sources. Sci. Adv 5 (5). doi:10.1126/sciadv.aav6134.

Hershey JR, Chen Z, Le Roux J, Watanabe S, 2016. Deep clustering: Discriminative embeddings for segmentation and separation. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 31–35. doi:10.1109/ICASSP.2016.7471631.

Hjortkjær J, Märcher-Rørsted J, Fuglsang SA, Dau T, 2020. Cortical oscillations and entrainment in speech processing during working memory load. European Journal of Neuroscience 51 (5), 1279–1289,

Horton C, Srinivasan R, D'zmura M, 2014. Envelope responses in single-trial eeg indicate attended speaker in a 'cocktail party'.. J. Neural. Eng 11 4, 046015, [PubMed: 24963838]

Kerlin JR, Shahin AJ, Miller LM, 2010. Attentional gain control of ongoing cortical speech representations in a "cocktail party". J. Neurosci 30 (2), 620–628. doi:10.1523/JNEUROSCI.3631-09.2010. [PubMed: 20071526]

Kingma DP, Ba J, 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980,

Kiselev I, Ceolini E, Wong D, d. Cheveigne A, Liu SC, 2017. Whisper: Wirelessly synchronized distributed audio sensor platform. In: 2017 IEEE 42nd Conference on Local Computer Networks Workshops (LCN Workshops), pp. 35–43. doi:10.1109/LCN.Workshops.2017.62.

Liu Y, Wang D, 2019. Divide and conquer: a deep casa approach to talker-independent monaural speaker separation. IEEE/ACM Trans. Audio Speech Lang. Process 27 (12), 2092–2102. [PubMed: 33748322]

Luo Y, Mesgarani N, 2018. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 696–700.

Luo Y, Mesgarani N, 2019. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. IEEE/ACM Trans. Audio Speech Lang. Process 27. doi:10.1109/TASLP.2019.2915167.

Mesgarani N, Chang EF, 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485 (7397), 233–236. [PubMed: 22522927]

Mesgarani N, David SV, Fritz JB, Shamma SA, 2009. Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. J. Neurophysiol 102 (6), 3329–3339. [PubMed: 19759321]

Miran S, Akram S, Sheikhattar A, Simon JZ, Zhang T, Babadi B, 2018. Real-time tracking of selective auditory attention from m/eeg: a bayesian filtering approach. Front Neurosci 12, 262. doi:10.3389/fnins.2018.00262. [PubMed: 29765298]

van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K, 2016. Wavenet: A generative model for raw audio. Arxiv.

Oreinos C, Buchholz JM, 2013. Measurement of a full 3d set of hrtfs for in-ear and hearing aid microphones on a head and torso simulator (hats). Acta Acustica united with Acustica 99 (5), 836–844.

O'Sullivan J, Chen Z, Herrero J, McKhann GM, Sheth SA, Mehta AD, Mesgarani N, 2017. Neural decoding of attentional selection in multi-speaker environments without access to clean sources. J. Neural. Eng 14 (5), 056001. doi:10.1088/1741-2552/aa7ab4. [PubMed: 28776506]

Peelle JE, Wingfield A, 2016. The neural consequences of age-related hearing loss. Trends Neurosci. 39, 486–497. [PubMed: 27262177]

Power AJ, Foxe JJ, Forde E-J, Reilly RB, Lalor EC, 2012. At what time is the cocktail party? a late locus of selective attention to natural speech. European Journal of Neuroscience 35 (9), 1497–1503. doi:10.1111/j.1460-9568.2012.08060.x.

Reindl K, Meier S, Barfuss H, Kellermann W, 2014. Minimum mutual information-based linearly constrained broadband signal extraction. IEEE/ACM Trans. Audio Speech Lang. Process 22, 1096–1108.

Roux JL, Wisdom S, Erdogan H, Hershey JR, 2019. Sdr half-baked or well done? In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 626–630.

Schwartz O, Gannot S, Habets EAP, 2017. Multispeaker lcmv beamformer and postfilter for source separation and noise reduction. IEEE/ACM Trans. Audio Speech Lang. Process 25 (5), 940–951. doi:10.1109/TASLP.2017.2655258.

Van Eyndhoven S, Francart T, Bertrand A, 2017. EEG-Informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses. IEEE Trans. Biomed. Eng 64 (5), 1045–1056. doi:10.1109/TBME.2016.2587382. [PubMed: 27392339]

Wang D, Chen J, 2018. Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans. Audio Speech Lang. Process 26 (10), 1702–1726. [PubMed: 31223631]

Wang J, Chen J, Su D, Chen L, Yu M, Qian Y, Yu D, 2018. Deep extractor network for target speaker recovery from single channel speech mixtures. INTERSPEECH.

Williamson DS, Wang Y, Wang D, 2016. Complex ratio masking for monaural speech separation. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 24 (3), 483–492.

Wong DDE, Fuglsang SA, Hjortkjær J, Ceolini E, Slaney M, de Cheveigné A, 2018. A comparison of regularization methods in forward and backward models for auditory attention decoding. Front. Neurosci 12, 531. doi:10.3389/fnins.2018.00531. [PubMed: 30131670]

Wong DDE, Hjortkjær J, Ceolini E, Nielsen SV, Griful SR, Fuglsang S, Chait M, Lunner T, Dau T, Liu S-C, de Cheveigné A, 2018. A closed-loop platform for real-time attention control of simultaneous sound streams. ARO Midwinter meeting (abstract), ARO Midwinter meeting (abstract).

Xiao X, Chen Z, Yoshioka T, Erdogan H, Liu C, Dimitriadis D, Droppo J, Gong Y, 2019. Single-channel speech extraction using speaker inventory and attention network. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 86–90.

Yu D, Kolbæk M, Tan Z-H, Jensen J, 2017. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 241–245.

Zhao L, Benesty J, Chen J, 2014. Design of robust differential microphone arrays. IEEE/ACM Trans. Audio Speech Lang. Process 22, 1455–1466.

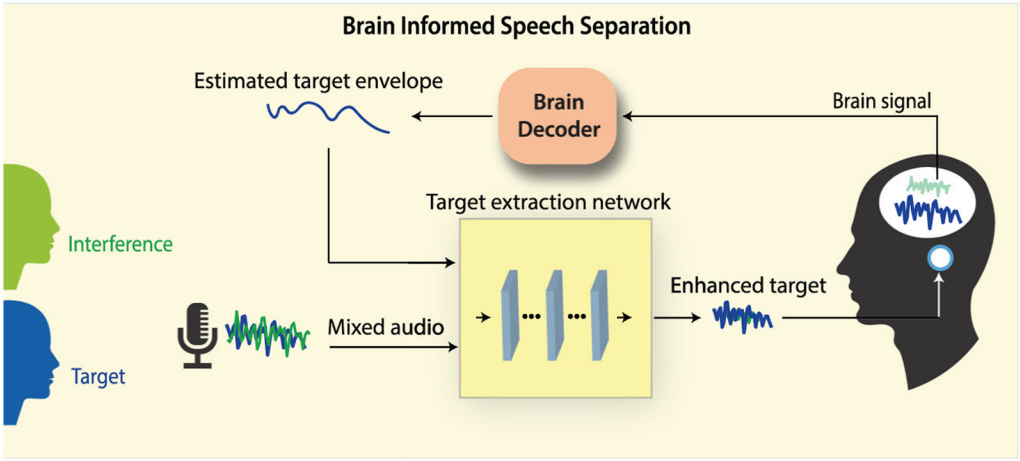

Fig. 1.

BISS schematic. A subject attends to one (blue) out of two simultaneous talkers. The decoding algorithm (the Brain Decoder) estimates the envelope of the attended speech based on recorded iEEG or EEG brain signals. The speech separation neural network model receives two inputs: (1) the speech mixture and (2) the decoded envelope (*hint*). These inputs are used by the model to separate and enhance the speech of the attended talker. The output of the model is (3) the enhanced speech which is fed to the HA of the subject in this closed-loop setup.

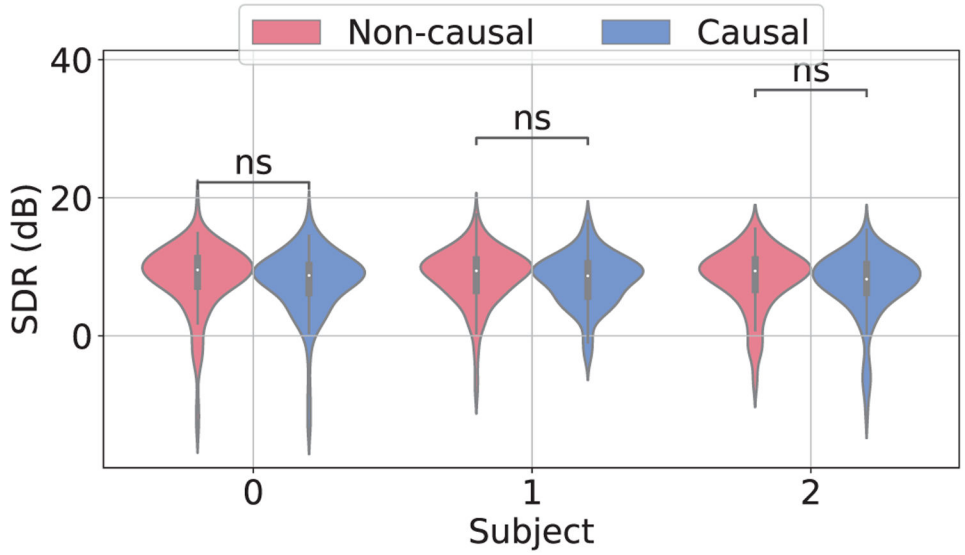

Fig. 2.

Results for BISS using envelopes decoded from iEEG data. Violin plots are presented for each subject separately and for the different model settings causal and non-causal. Significance is indicated by **ns** if $p > 0.05$ using Mann - Whitney U test.

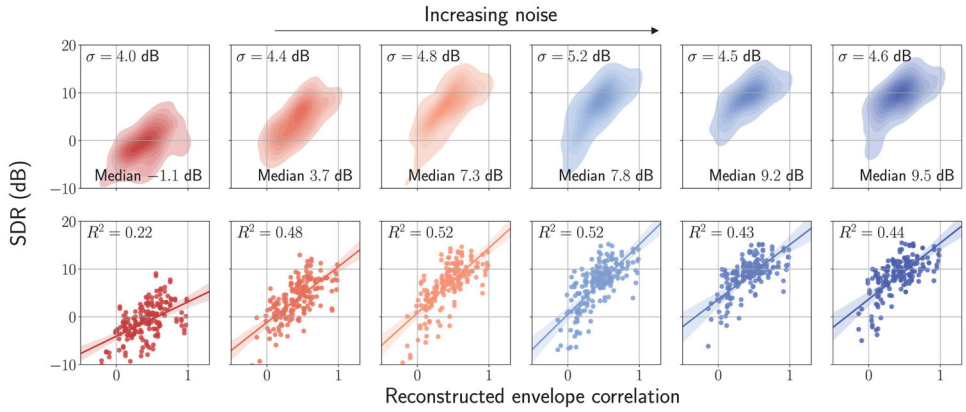

Fig. 3.

Results from Subject 0 in the iEEG recordings as a function of noise variance during curriculum training. The x-axis indicates $r_{diff} = r_{attended} - r_{unattended}$, and the y-axis indicates SDR improvement in dB. The panels in the top row show the density distribution of the points using kernel density estimate with Gaussian kernels. The panels in the bottom row show each utterance separately and a linear fit obtained using linear regression. The shaded area represents the 95% confidence interval of the regression. The panels from left to right show results from increasing the σ of the noise during training (from σ = 0.0 to σ = 0.5 with steps of 0.1).

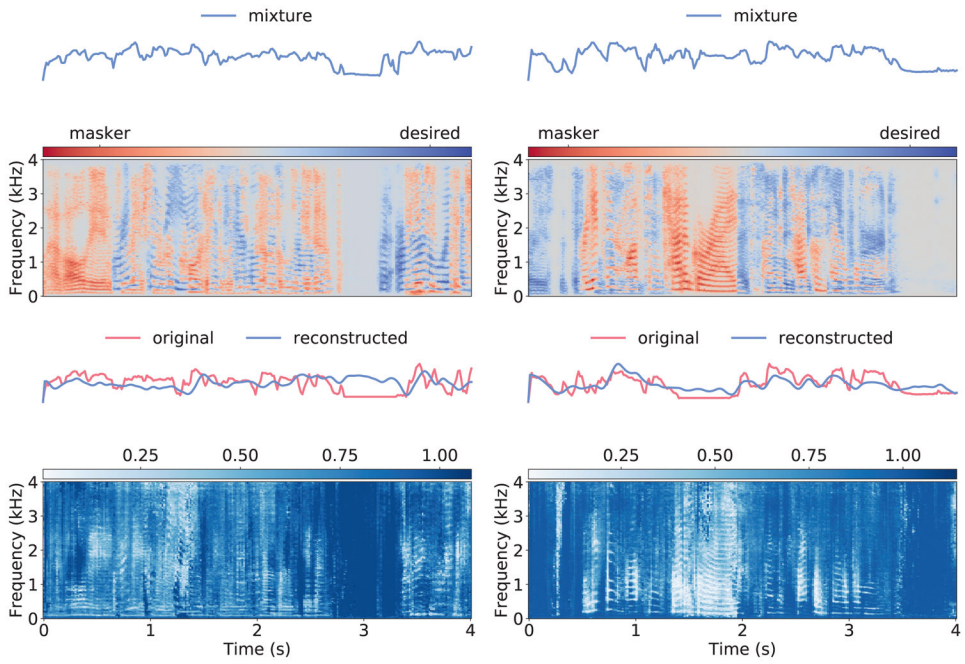

Fig. 4.

Two examples of estimated masks. a) failed mask with a correlation of −0.13 and an SDR improvement of −10.4 dB. b) successful mask with a r of 0.69 and an SDR of 9.2 dB. From top to bottom the panels represent: Mixture envelope, mixture spectrogram with desired speaker highlighted in blue and masker speaker highlighted in red, original and reconstructed desired speech envelopes, mask estimated by the model based on the decoded envelope.

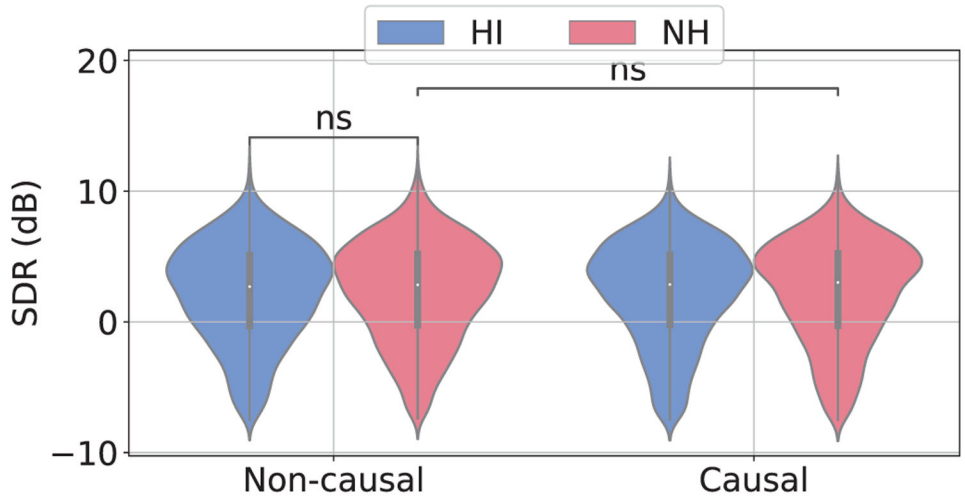

Fig. 5.

Separation results from the BISS model using envelopes reconstructed from EEG data. Highlighted are the performance for each of the two groups, NH (21 subjects) and HI (20 subjects) for both the causal and non-causal models. Each subject was tested on 128 non-overlapping utterances of 4 seconds. The y-axis shows the separation quality in terms of SDR improvement in dB. Significance is indicated by **ns** if $p > 0.05$ using Mann - Whitney U test.

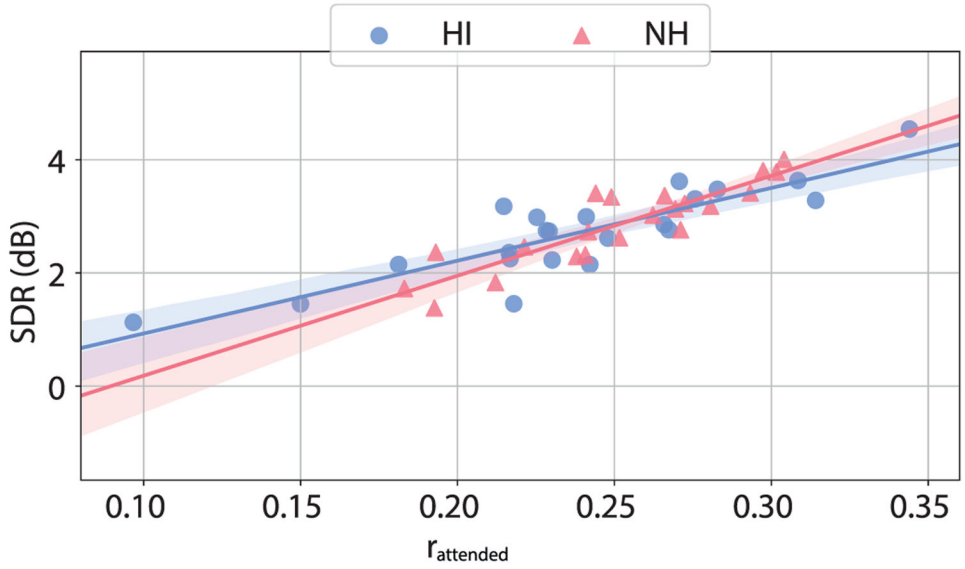

Fig. 6.

Separation performance using envelopes reconstructed from EEG for each subject. The performance of the NH group (21 subjects) and HI group (20 subjects) are shown for the model in the causal setting. Each subject was tested on 128 non-overlapping utterances of 4 seconds. The y-axis shows the separation quality in terms of SDR improvement in dB. Significance is indicated by **ns** if $p > 0.05$ using Mann - Whitney U test.

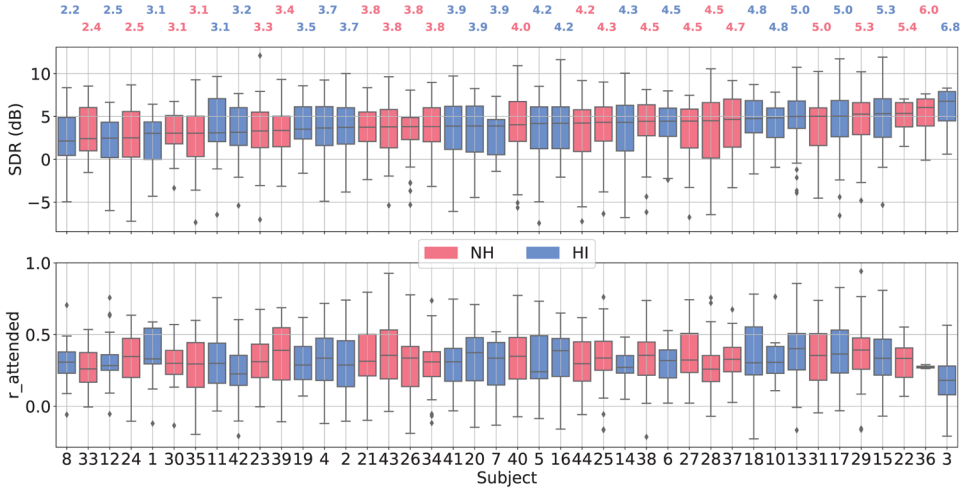

Fig. 7.

Performance for all EEG subjects considering only correctly decoded trials ($r_{diff} > 0.0$). Results for the HI and NH groups are shown in blue and red, respectively. Median values for SDR are highlighted above the top panel.

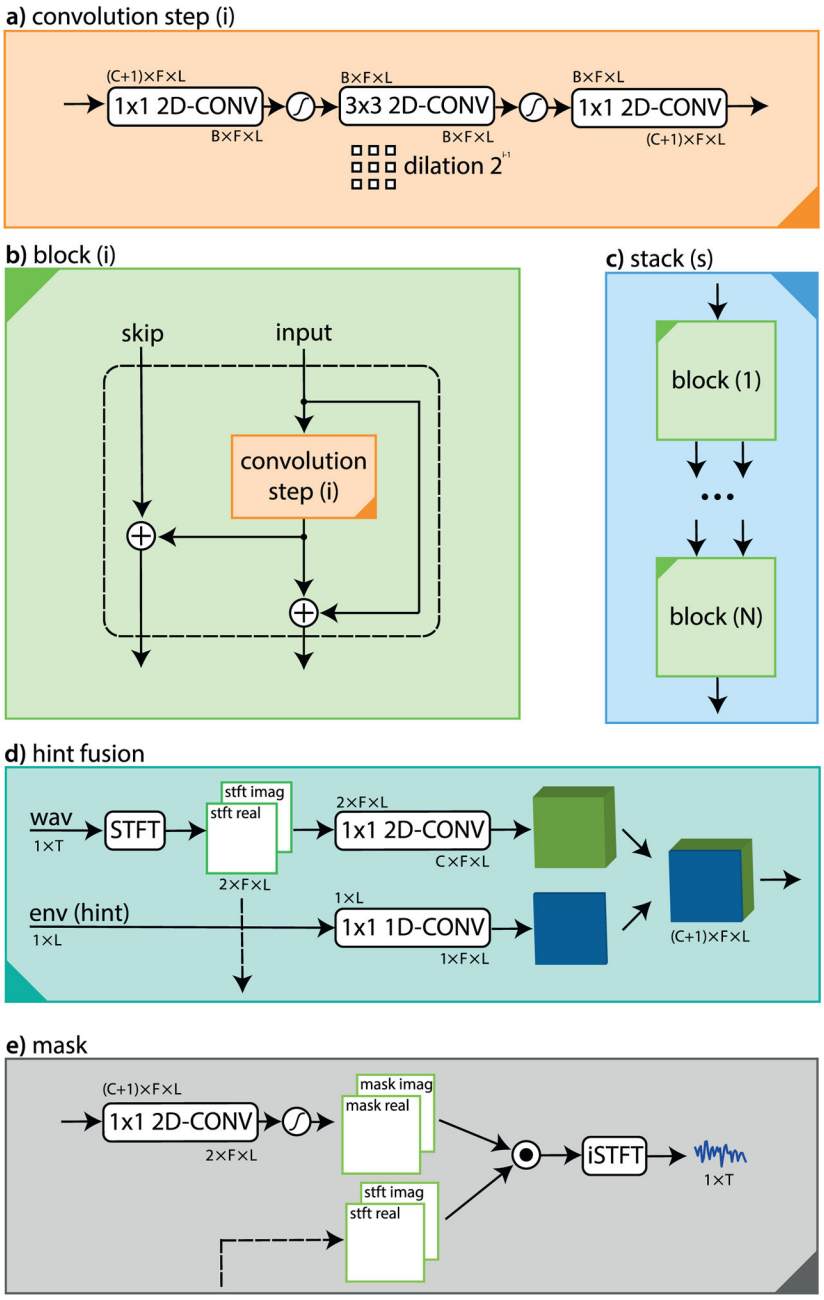

Fig. 8.

Architecture of the network used for BISS. Panels **a)** to **e)** show the details of each sub-module of the full architecture.

Table 1

Legend of symbols used in the model architecture.

Symbol	Description	Value
F	Number of frequency bins	257
L	Number of STFT time windows	257
T	Number of samples in the waveform	32000
C	Channels in the stack	32
B	Channels in the convolutional step	64
S	Number of stacks	2
N	Number of blocks	6
i	Index of each block (dilation factor)	-
s	Index of each stack	-
k	kernel size	3