A Fused Deep Denoising Sound Coding Strategy for Bilateral Cochlear Implants

Tom Gajecki and Waldo Nogueira, *Member, IEEE*

***Abstract*—Cochlear implants (CIs) provide a solution for individuals with severe sensorineural hearing loss to regain their hearing abilities. When someone experiences this form of hearing impairment in both ears, they may be equipped with two separate CI devices, which will typically further improve the CI benefits. This spatial hearing is particularly crucial when tackling the challenge of understanding speech in noisy environments, a common issue CI users face. Currently, extensive research is dedicated to developing algorithms that can autonomously filter out undesired background noises from desired speech signals. At present, some research focuses on achieving end-to-end denoising, either as an integral component of the initial CI signal processing or by fully integrating the denoising process into the CI sound coding strategy. This work is presented in the context of bilateral CI (BiCI) systems, where we propose a deep-learning-based bilateral speech enhancement model that shares information between both hearing sides. Specifically, we connect two monaural end-to-end deep denoising sound coding techniques through intermediary latent fusion layers. These layers amalgamate the latent representations generated by these techniques by multiplying them together, resulting in an enhanced ability to reduce noise and improve learning generalization. The objective instrumental results demonstrate that the proposed fused BiCI sound coding strategy achieves higher interaural coherence, superior noise reduction, and enhanced predicted speech intelligibility scores compared to the baseline methods. Furthermore, our speech-in-noise intelligibility results in BiCI users reveal that the deep denoising sound coding strategy can attain scores similar to those achieved in quiet conditions.**

***Index Terms*—Cochlear implants, sound coding strategy, deep neural networks, end-to-end, speech enhancement.**

I. Introduction

A COCHLEAR implant (CI) is a medical device surgically implanted to restore the sense of hearing in individuals with severe to profound sensorineural hearing loss. Notably, recent years have seen significant advancements in CI technology [1]. Consequently, individuals with bilateral hearing loss often receive implants on both sides [2]. Those who receive a CI in each ear are known as bilateral CI (BiCI) users, typically demonstrate improved speech understanding, sound localization, reduced listening effort, and enhanced quality of life in comparison to unilateral CI users (e.g., [3], [4], [5], [6]). However, their listening performance remains inferior to individuals with normal hearing (NH) (e.g., [7], [8], [9]).

The lower hearing performance in BiCI users could potentially stem from differences in electrode array insertion depth in each ear, differences between the electrode-nerve interfaces in each ear, and from the independent processing in each CI (e.g., [10], [11], [12], [13]). The computation of stimulation current levels over time and for individual electrodes (referred to as electrodograms) relies on audio captured by microphones embedded within each speech processor. The computation involves applying the CI sound coding strategy separately to each listening side. This method may lead to challenges, which include a potential lack of effective binaural integration [14] and the introduction of possible binaural artifacts [15]. Moreover, it might struggle with suppressing background noise or competing speech signals when present simultaneously in both ears [12]. Additionally, this approach might have limitations in fully transmitting interaural cues [16].

Typically, a CI in conjunction with its associated sound coding strategy enables the user to understand speech effectively in quiet environments. However, its effectiveness diminishes when encountering loud interfering signals, characterized by low signal-to-noise ratios (SNRs), such as background noise or multiple speakers talking simultaneously [17]. Several approaches have been proposed to enhance speech understanding in noisy environments for BiCIs. Some of these methods utilize traditional front-end processing techniques like binaural beamforming (e.g., [18], [19], [20]), while others integrate elements of the CI sound coding strategy and establish bilateral connections between certain processing components (e.g., [12], [21], [22]). These conventional approaches have proven effective in augmenting speech understanding in noise and sound source localization for BiCI users. However, with the advent of deep learning technology, the field is increasingly exploring the use of deep neural networks (DNNs) for speech enhancement (e.g., [23], [24], [25], [26], [27]). These methods have proven to be very successful at performing speech denoising while keeping speech quality and a high degree of generalization capabilities.

To optimize the enhancement of speech for CIs, it could prove advantageous to devise algorithms that take into account

Manuscript received 27 September 2023; revised 26 January 2024; accepted 14 February 2024. Date of publication 20 February 2024; date of current version 20 June 2024. This work was supported by German Research Foundation (DFG) under Project ID 446611346, led by Waldo Nogueira and Jörn Ostermann. *(Corresponding author: Waldo Nogueira.)*

Tom Gajecki is with the Department of Otolaryngology, Medical University of Hannover and Cluster of Excellence Hearing4all, Germany.

Waldo Nogueira is with the Department of Otolaryngology, Medical University of Hannover and Cluster of Excellence Hearing4all, Hannover 30625, Germany (e-mail: nogueiravazquez@mh-hannover.de).

Digital Object Identifier 10.1109/TBME.2024.3367530

the specific processing scheme of CIs. Consequently, there has been research dedicated to CIs, where DNNs are incorporated into their signal pathway [27], [28], [29], [30], [31]. These approaches target noise reduction by directly applying masks within the filter bank utilized by the CI sound coding strategy. Recently, drawing inspiration from the Conv-TasNet [23], an end-to-end CI sound coding strategy based on deep learning, termed "Deep ACE." was proposed [32], [33]. This approach was designed to replace the clinically available ACE sound coding strategy, but it could also be used to replace other commercially available ones. This method completely replaces the CI sound coding strategy with a DNN and achieves high speech understanding improvements in BiCI users (up to 22.8% improvement in word recognition score (WRS) in modulated background noise).

Presently, data-driven methodologies have primarily been employed with single CIs. However, these approaches are equally applicable to BiCIs. For instance, there is no inherent justification to assume that utilizing any monaural speech enhancement algorithm within a BiCI framework would not produce comparable auditory advantages as observed in the unilateral configuration. Nevertheless, this may not be the optimal approach, as independent processing could still potentially introduce artifacts that hinder effective binaural listening. A promising avenue to enhance the listening experience of BiCI users involves embracing multi-channel sound processing. Notably, diverse multi-channel front-end speech enhancement methods have been proposed. These strategies not only showcase effectiveness in enhancing speech denoising but also reveal an ability to maintain the integrity of essential binaural auditory cues (e.g., [34], [35]).

In recent advancements, there's a novel concept referred to as "Fusion Layers" introduced in [36]. These layers entail the exchange of information between two individual monaural speech-denoising algorithms. They achieve this by enabling Hadamard products between latent spaces at specific processing stages, drawing inspiration from multi-task learning methods, and emulating the inhibitory and excitatory mechanisms found in the human brain stem for binaural hearing [37]. Precisely, the fusion layers are designed to introduce non-linear elements into the learning model, enhancing the model's ability to fit training data effectively while improving generalization without impacting the number of trainable parameters. This approach of sharing features has proven to be highly effective in enhancing noise reduction compared to independent bilateral models, where processing is performed separately on each side.

In our study, we present a novel approach termed the "fused Deep ACE," which can naturally be extrapolated to other CI processing strategies. This approach integrates two Deep ACE algorithms through the utilization of fusion layers, enabling the sharing of latent representations from particular processing stages. We hypothesize that this bilateral sound coding strategy will result in improved speech understanding when contrasted with the conventional clinical approach. Additionally, we postulate that the fusion layer will capitalize on bilateral redundant information, potentially mitigating certain binaural artifacts and leading to the generation of more bilaterally coherent output electrodograms.

II. Methods & Materials

A. Algorithms

1) Bilateral Advanced Combination Encoder (ACE; Unprocessed): This is the main baseline algorithm used in this work and is based on a clinical BiCI setup, where each CI processes the sound independently using the ACE sound coding strategy. This setup does not perform any noise reduction and does not share any information between the listening sides. The ACE strategy begins by sampling the acoustic signal at 16 kHz, followed by applying a filter bank through a 128-point fast Fourier transform. This process introduces a 2 ms algorithmic latency, dependent on the channel stimulation rate (CSR). Estimations of desired envelopes are calculated for each spectral band (E_k) corresponding to an electrode, with M representing the total channels.

In this study, we select the N most energetic envelopes out of M based on their amplitudes. These selected envelopes undergo non-linear compression via a loudness growth function (LGF). The LGF output (p_k) represents the normalized stimulation amplitude for electrode k to stimulate the auditory nerve. Lastly, we map each p_k within the subject's dynamic range, spanning from threshold to comfortable stimulation levels giving the output current stimulation patterns I_k. These N-selected electrodes are stimulated sequentially for each audio frame, defining one stimulation cycle, and the CSR is determined by the cycles per second.

2) Bilateral Deep ACE: This condition closely resembles the baseline scenario (referred to as bilateral ACE) in that it lacks any exchange of information between the listening sides. However, it diverges from clinical ACE sound coding strategies by adopting the newly developed Deep ACE approach, as detailed in [32], [33]. More specifically, Deep ACE substitutes the conventional clinical ACE method with a DNN that takes in raw audio as its input and generates the denoised LGF output p_k.

In the initial step, Deep ACE encodes the left and right signals $X_{\{l,r\}}$ into a latent representation using a 1-D convolution layer. This operation can be mathematically expressed as a matrix multiplication:

$$X'_{\{l,r\}} = \Theta(X_{\{l,r\}} \cdot \mathbf{E}), \tag{1}$$

where $\mathbf{E}_{\{l,r\}} \in \Re^{(\mathbf{F}\times\mathbf{L})}$ are the left and right encoder basis functions and $\Theta(\cdot)$ is the antirectifier activation function used in Deep ACE, and F and L the number and length (in samples) of the filters used, respectively. The signal is then sent to a deep envelope detector (DED) that performs dimensionality reduction (from F to M) and to the separator module that will generate a deeper latent representation for each side $X''_{l,r} = \zeta(X'_{\{l,r\}}) \in \Re^{(1\times\mathbf{S})}$, where $\zeta(\cdot)$ is the learned function by the separator and $\mathbf{S}$ is the number of skip connections [23]. Then the DED and separator outputs are fed into a masker that will remove the noisy components of the encoded mixture. Specifically, during the mask generation, the fused output (after the separator) undergoes a series of sequential transformations. It first passes through a PReLU activation layer, followed by a 1D convolution to match its dimensionality with the encoded

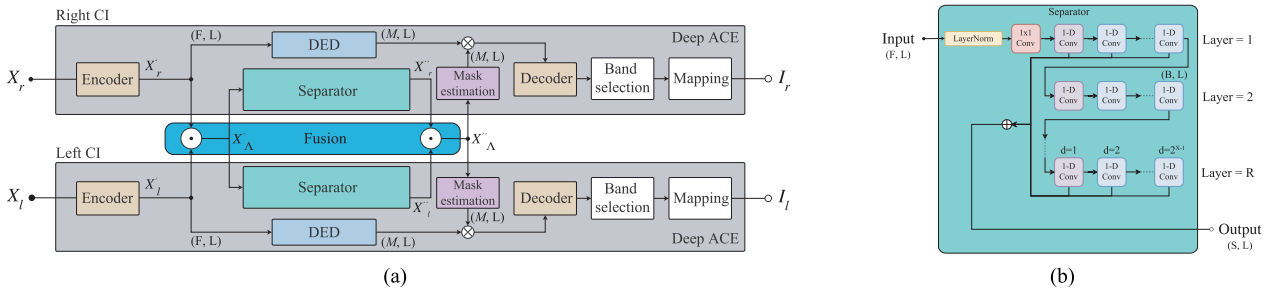

Fig. 1. Block diagram of the proposed fused Deep ACE (left panel; (a)) and a diagram of the separator module (right panel; (b)). The fused model takes the right and left time-domain noisy speech signals (X_r and X_l respectively) and produces the respective denoised current stimulation patterns for each listening side I_r and I_l for each stimulation frame. The fusion model performs element-wise dot product between the latent representations generated in each of the Deep ACE models and the DED is used for dimensionality reduction. For a brief explanation and values of the variables shown refer to Table I.

input. Finally, it proceeds through a soft masking process to estimate binary probabilities indicating the presence of speech content. This resultant mask is then applied to the DED output, effectively eliminating noisy components. Finally, the masked signals will be decoded through a transposed 1-D convolution to obtain p_k for each CI.

3) Fused Deep ACE: In this study, we introduce an approach involving integrating two monaural Deep ACE [33] models, with one model associated with each listening side. This integration is achieved through the utilization of fusion layers [36]. These fusion layers are influenced by the principles of multi-task learning, where model weights are shared across different models to address interconnected tasks. The function of these layers involves conducting element-wise dot products on tensors that depict latent representations at identical processing stages. More precisely, we combine the latent representations generated within each Deep ACE model, both following the encoding stage and subsequent to the separator modules as follows:

$$X'_\Lambda = \rho(X'_l, X'_r)$$
$$X''_\Lambda = \rho(X''_l, X''_r), \tag{2}$$

where $\rho(\cdot)$ is the Hadamard product operator. The outcome of these two fusion operations results in a model that performs "double fusion." These fused signals are fed into the separator and masker modules the same way as in the bilateral Deep ACE condition. In this model the consistency in dimensionality persists throughout each phase of the fused model, aligning with the structure of the independent model (bilateral deep ACE). A visual representation of this model's structure can be observed in Fig. 1. It is important to note that the band selection and mapping blocks function autonomously. This means that the bands chosen on one side might not align with those selected on the other side. Similarly, the mapping block operates separately for each electrode, with unique threshold and comfortable levels assigned to individual electrodes.

B. Model Training Setup

Training the models was conducted over a maximum of 100 epochs, employing batches consisting of two 4-second-long audio segments. The initial learning rate was initialized to 1e-3.

TABLE I
HYPERPARAMETERS USED TO TRAIN THE DEEP LEARNING MODELS

Parameter	Value	Description
F	64	Size of encoding
L	32	Window size
B	64	Bottleneck size
S	32	Skip connection size
X	8	Number of dilations
R	3	Number of layers

In case the validation set accuracy displayed no enhancement over a span of 3 consecutive epochs, the learning rate was reduced by half. To ensure regularization, early stopping with a patience of 5 epochs was implemented, safeguarding against overfitting. Only the model displaying the highest performance was retained. Model optimization was facilitated using the Adam optimizer [38]. The model's training employs a cost function based on mean-squared error (MSE) and binary cross entropy (BCE) for each listening side (for a detailed description of this cost function refer to [33]).

The hyperparameter configuration used in this study was slightly modified with respect to the ones shown in [32], specifically the separator module was bigger and the deep-envelope-detector (DED) was also increased in size. The model hyperparameters are shown in Table I.

C. Audio Material

In this work, we used a total of three different speech datasets and three noise types to assess the models' performance and generalization abilities. All these audio sets will be described in this section. As a preprocessing stage, all audio material was set to mono and resampled at 16 kHz. The corresponding electrodograms were obtained by processing all audio data with the ACE sound coding strategy at an output CSR of 1,000 pulses per second. All audio signals were generated by convolving source signals with binaural room impulse responses (BRIRs; [39]) and summing. BRIRs were generated for hearing aids located in each listening side [1] and consisted of 4 different rooms of different sizes and acoustic properties (see Table II), using the front microphone.

[1] https://github.com/IoSR-Surrey/RealRoomBRIRs

TABLE II
ROOM ACOUSTICAL PROPERTIES, INCLUDING RT_{60}, INITIAL TIME DELAY GAP (ITDG), DIRECT-TO-REVERBERANT RATIO (DRR), AND CLARITY INDEX Cte

Room	RT_{60}[s]	ITDG [ms]	DRR [dB]	C_{te} [dB]
A	0.32	8.72	6.09	16.5
B	0.47	9.66	5.31	11.4
C	0.68	11.9	8.82	17.4
D	0.89	21.6	6.12	9.53

1) Speech Data:

a) LibriVox corpus [40]: This speech data was originally designed for end-to-end speech translation, however, in this study, we mix the speech material with noise to train our models for speech denoising. The speech data consists of fluent spoken sentences with a total duration of 18 hours. The quality of audio and sentence alignments was checked by a manual evaluation, showing that speech alignment is in general very high. In fact, the sentence alignment quality is comparable to well-used parallel translation data.

b) TIMIT corpus [41]: This corpus contains broadband recordings of 630 people speaking the eight major dialects of American English, each reading ten phonetically rich sentences. In this work, files from 112 male and 56 female speakers in the test set were selected.

c) HSM corpus [42]: Speech intelligibility in quiet and in noise was measured utilizing the Hochmair, Schulz, Moser (HSM) sentence test, based on a dataset composed of 30 lists with 20 everyday sentences each (106 words per list).

2) Noise Data:

a) Environmental noises; DEMAND [43]: The environmental noises recorded to create this dataset are split into six categories; four are indoor noises and the other two are outdoor recordings. The indoor environments are further divided into domestic, office, public, and transportation; the open-air environments are divided into streets and nature. There are 3 environment recordings per category.

b) Synthetic noises; SSN [44] and ICRA7 [45]: To evaluate the different algorithms, in this work we also use stationary speech-shaped noise (SSN) and non-stationary modulated seven-speaker babble noise (ICRA7) as synthetic interferers.

3) Training, Evaluation and Testing Data:

The training set was composed of speech from the LibriVox corpus and noise from the DEMAND dataset. Specifically, 30 male (M) and female (F) speakers were randomly selected from the speech corpus, and two environments were randomly selected from each of the noise categories. For validation, 20% of the training data was used. The testing phase involved the utilization of the HSM speech dataset, coupled with synthetic noises employed as interfering signals.

During the training, validation, and objective testing phases, the speech and noise signals were spatially separated, and positioned on opposite sides in relation to the listener's frontal orientation. The specific placements of these signals were randomly chosen within a range of 0 to $\pm 90°$. However, it's essential to note that for the listening experiments, the placements of target speech and interfering noise source were not selected randomly. In those experiments, the speaker consistently remained in front of the listener, while the noise source was consistently situated at $\pm 55°$ azimuth, effectively masking the better-performing CI. Speech and noise signals were mixed at SNR values ranging uniformly from -5 to 10 dB calculated at the better SNR side (note here that at large noise azimuths the CI situated ipsilateral to the noise source might encounter a significant SNR decrease, potentially up to around 10 dB). The processed clean speech signals were also included in the listening experiments to assess whether the proposed model introduced perceptually relevant distortions.

D. Objective Evaluation

To objectively evaluate the performance of each examined algorithm, we gauge the extent of noise reduction accomplished, establish electrode-wise correlation coefficients between the denoised and clean signals, and determine speech intelligibility through the application of the modified binaural short-time objective intelligibility (MBSTOI) index [46]. Notably, in this study, our focus is on investigating comprehensive CI processing, consequently prompting the computation of the MBSTOI index from synthesized electrodograms ($\boldsymbol{p}$) derived using a vocoder. This results in the utilization of a specific variant of MBSTOI referred to as vocoder-MBSTOI (V-MBSTOI).

1) SNRi:

To assess the amount of noise reduction performed by each of the tested algorithms we compute the SNR improvement (SNRi). This measure is calculated in the electrodogram domain and compares the original input SNR to the one obtained after denoising, and is given by:

$$\mathrm{SNRi} = 10 \cdot \log_{10}\left(\frac{\sum_{k=1}^{M} ||\boldsymbol{p}_k^n - \boldsymbol{p}_k^c||^2}{\sum_{k=1}^{M} ||\boldsymbol{p}_k^d - \boldsymbol{p}_k^c||^2}\right), \tag{3}$$

where $\boldsymbol{p}_k$ represents the LGF output of band k and the superscripts n, c, and d are used to denote the noisy, clean, and denoised electrodograms, respectively.

a) V-MBSTOI: To estimate the speech intelligibility performance expected from each of the algorithms, the V-MBSTOI score [47], [48], [49] was used. This metric relies directly on MBSTOI [46], which is modeled based on normal hearing binaural speech performance. Specifically, the purpose of this metric is to evaluate the potential relative variations in speech performance that could be achieved in behavioral experiments, rather than providing an exact estimation of an individual's CI performance. The V-MBSTOI score ranges from 0 to 1, where the higher score represents a predicted higher speech performance.

b) Linear cross-correlation: To characterize potential distortions and artifacts introduced by the tested algorithms, the linear correlation coefficients (LCCs) between the clean ACE electrodograms ($\boldsymbol{p}^c$) and the denoised electrodograms ($\boldsymbol{p}^d$) were computed. The LCCs were first computed channel-wise (i.e., one correlation coefficient was computed for each of the 22 channels) to assess channel output degradation caused by the denoising process. The LCC_k for band k is computed based on

the Pearson correlation coefficient [50] as follows:

$$\mathrm{LCC}_k = \frac{\mathrm{cov}\left(\boldsymbol{p}_k^c, \boldsymbol{p}_k^d\right)}{\sigma_{\boldsymbol{p}_k^c} \cdot \sigma_{\boldsymbol{p}_k^d}}, \tag{4}$$

where $\mathrm{cov}(X, Y)$ is the covariance between X and Y, and $\sigma_{\boldsymbol{p}_k}$ is the standard deviation of the values in the corresponding electrodogram $\boldsymbol{p}_k$.

We also present the LCCs as a function of the noise azimuth LCC_θ, which is computed as follows:

$$\mathrm{LCC}_\theta = \frac{\mathrm{cov}\left(\boldsymbol{p}_\theta^c, \boldsymbol{p}_\theta^d\right)}{\sigma_{\boldsymbol{p}_\theta^c} \cdot \sigma_{\boldsymbol{p}_\theta^d}}, \tag{5}$$

where $\boldsymbol{p}_\theta^c$ and $\boldsymbol{p}_\theta^d$ are the LCCs averaged across electrodes for a noise source coming from azimuth θ.

c) Electric interaural coherence: Similar to the LCCs, we also use the electric interaural coherence (EIC). Here we compute the channel-wise LCCs between between the right electrodograms ($\boldsymbol{p}^r$) and the left electrodograms ($\boldsymbol{p}^l$) as follows:

$$\mathrm{EIC}_k = \frac{\mathrm{cov}\left(\boldsymbol{p}_k^r, \boldsymbol{p}_k^l\right)}{\sigma_{\boldsymbol{p}_k^r} \cdot \sigma_{\boldsymbol{p}_k^l}}. \tag{6}$$

We also present the EIC as a function of the noise azimuth LCC_θ, which is computed as follows:

$$\mathrm{EIC}_\theta = \frac{\mathrm{cov}\left(\boldsymbol{p}_\theta^r, \boldsymbol{p}_\theta^l\right)}{\sigma_{\boldsymbol{p}_\theta^r} \cdot \sigma_{\boldsymbol{p}_\theta^l}}. \tag{7}$$

E. Behavioral Evaluation

To validate the objective instrumental measures and to assess their impact on actual BiCI hearing, we perform two behavioral experiments namely, a speech intelligibility experiment and a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA). The speech intelligibility experiments are designed to investigate the benefits of the proposed denoising algorithms when compared to the clinical setups, and the MUSHRA [51] will help understand how BiCIs rate the quality of the performed denoising.

The stereo signals were transmitted through direct stimulation using a bilaterally synchronized RF GeneratorXS interface from Cochlear Ltd. (Sydney, Australia) in conjunction with MATLAB software (Mathworks, Natick, MA) via the Nucleus Implant Communicator V.3, also from Cochlear Ltd. All testing procedures were conducted on a personal computer equipped with customized MATLAB software. Before commencing experiments involving subjects, a hardware check was carried out by analyzing the signals generated by the research interface using an oscilloscope. The stimulation signals were characterized by cathodic-phase leading, biphasic pulses presented in a monopolar configuration (MP1+2). This stimulation mode utilizes two extracochlear electrodes: one ball electrode positioned under the temporalis muscle and another plate electrode on the receiver case. These pulses consistently featured an 8-μs phase gap and 25-μs phase duration, and they were presented in a base-to-apex sequence.

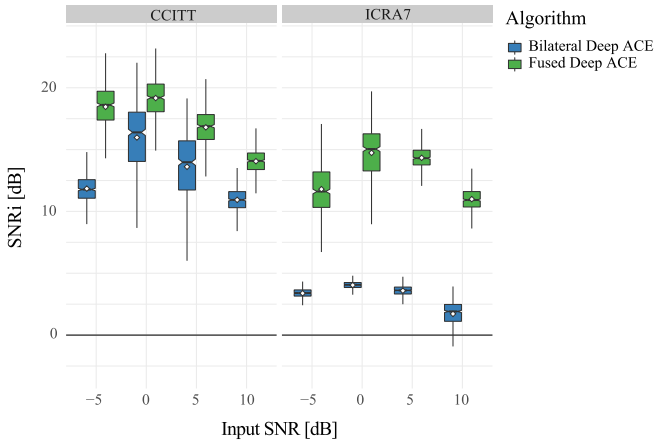

Fig. 2. Box plots showing the mean SNRi scores across listening sides in dB for the tested algorithms in CCITT and ICRA7 noises for the different SNRs using the HSM speech dataset. The black horizontal bars within each box represent the median for each condition, the diamond-shaped marks indicate the mean, and the top and bottom extremes of the boxes indicate the $Q_3 = 75\%$ and $Q_1 = 25\%$ quartiles, respectively. The box length is given by the interquartile range (IQR), used to define the whiskers that show the variability of the data above the upper and lower quartiles (the upper whisker is given by $Q_3 + 1.5{\cdot}$IQR and the lower whisker is given by $Q_1 - 1.5{\cdot}$IQR [52]).

1) Speech Understanding Experiment: Speech intelligibility in noisy environments was assessed using the HSM sentence set [42]. To conduct this assessment, each speech token underwent digital downsampling from 44.1 kHz to 16 kHz. During testing, subjects were presented with sentences from the front in a simulated acoustic setting, which included background interference noise (either CCITT or ICRA7) originating from a 55-degree azimuth angle, masking their self-reported better ear. The noise azimuth was selected to be 55 degrees because this angle corresponded to the point where electrical interaural coherence (EIC; described in Section II-D1c) was at its minimum (see Fig. 8), thus maximizing the impact on speech understanding.

Before the speech tests began, a training phase was implemented, comprising two sets of 20 sentences presented in quiet conditions. This training allowed listeners to adapt to the fitting parameters specific to the study and familiarize themselves with the sound delivery through the research interface.

Subjects were instructed to verbally repeat the sentences as accurately as possible during the tests. Two observers were present during the tests: one managed the software interface, while the other recorded the number of correctly identified words by marking them in a printed list corresponding to the sentences. Each listening condition was evaluated twice using different sentence lists, and the final score was computed as the average number of correctly identified words across these repetitions. The subjects were unaware of the specific conditions being tested, and an audiologist, blind to the test conditions, conducted the speech intelligibility assessments.

2) MUSHRA: This test is aimed at assessing how well-presented speech sentences are perceived in comparison to a specified reference using MUSHRA. The scores provided by the listener will range from 0 (poor) to 100 (excellent). In the

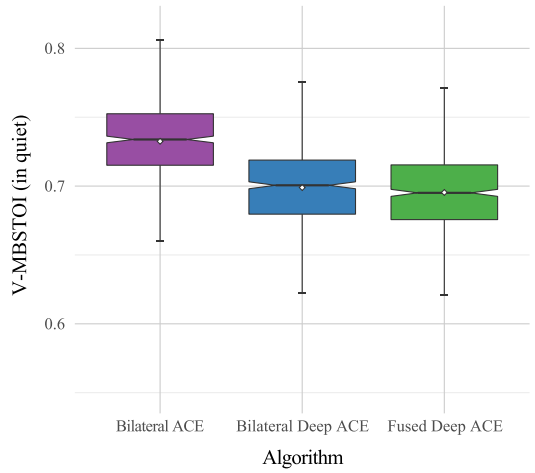

Fig. 3. Box plots showing the V-MBSTOI scores for the tested algorithms in quiet for the different SNRs using the HSM speech dataset. The black horizontal bars within each box represent the median for each condition, the diamond-shaped marks indicate the mean, and the top and bottom extremes of the boxes indicate the $Q_3 = 75\%$ and $Q_1 = 25\%$ quartiles, respectively. The box length is given by the interquartile range (IQR), used to define the whiskers that show the variability of the data above the upper and lower quartiles (the upper whisker is given by $Q_3 + 1.5{\cdot}$IQR and the lower whisker is given by $Q_1 - 1.5{\cdot}$IQR [52]).

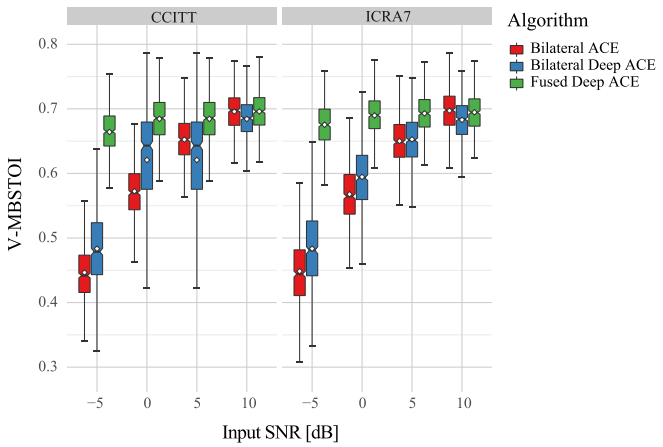

Fig. 4. Box plots showing the V-MBSTOI scores for the tested algorithms in CCITT and ICRA7 noises for the different SNRs using the HSM speech dataset. The black horizontal bars within each box represent the median for each condition, the diamond-shaped marks indicate the mean, and the top and bottom extremes of the boxes indicate the $Q_3 = 75\%$ and $Q_1 = 25\%$ quartiles, respectively. The box length is given by the interquartile range (IQR), used to define the whiskers that show the variability of the data above the upper and lower quartiles (the upper whisker is given by $Q_3 + 1.5{\cdot}$IQR and the lower whisker is given by $Q_1 - 1.5{\cdot}$IQR [52]).

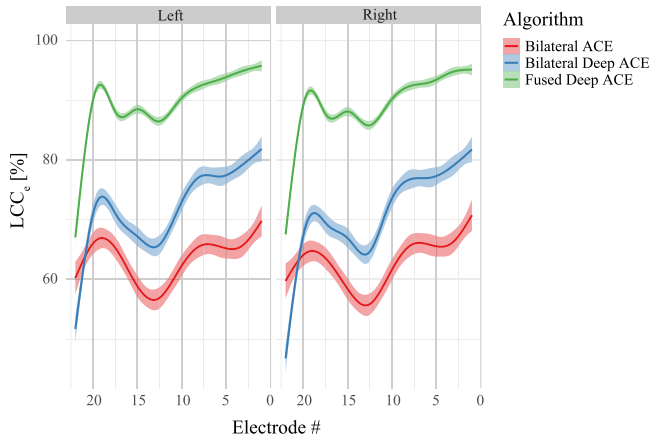

Fig. 5. Polynomial regressions showing the channel-wise LCCs between processed and clean electrodograms for the different algorithms, noises, and listening sides using the HSM dataset. Shaded areas represent the 95% confidence level interval [52]. Higher electrode numbers represent lower frequencies.

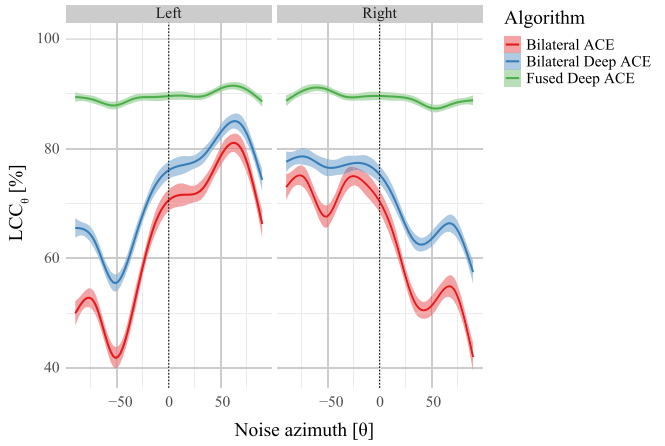

Fig. 6. Polynomial regressions showing the linear cross-correlations between processed and clean electrodograms for the different algorithms averaged across electrodes as a function of the azimuth, noises, and listening sides using the HSM dataset. Shaded areas represent the 95% confidence level interval [52]. Higher electrode numbers represent lower frequencies.

context of this study, the primary goal of this experiment was to establish a relative score for the quality of speech denoising concerning the clean speech signal generated by the clinical sound coding strategy ACE. To create a reference point, we derived an anchor by applying a low-pass filter with a cut-off frequency of 3.5 kHz to the noisy, unprocessed mixture. Two primary conditions were examined: one with clean audio and the other in a noisy environment (using both CCITT and ICRA7 noise profiles).

In the clean condition, we compared the reference clean ACE to the anchor and the clean speech signals processed separately by the independent BiCI strategy and the fused Deep ACE sound coding strategy. This comparison aimed to determine if there were discernible differences between clinical processing in a quiet setting and the proposed algorithms.

In the noisy condition, we compared the reference clean ACE to the anchor, the two proposed algorithms, and the unprocessed ACE signal in a noisy environment. Within each MUSHRA block, corresponding to each primary condition, eight sentences were assessed. The sentences were delivered at different SNRs, with two each at -5 dB, 0 dB, 5 dB, and 10 dB.

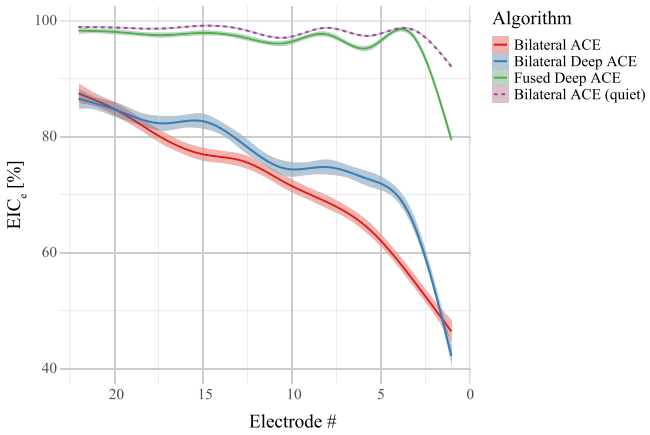

Fig. 7. Polynomial regressions showing the EIC for each electrode pair averaged across noises and listening sides using the HSM dataset. Shaded areas represent the 95% confidence level interval [52]. Higher electrode numbers represent lower frequencies.

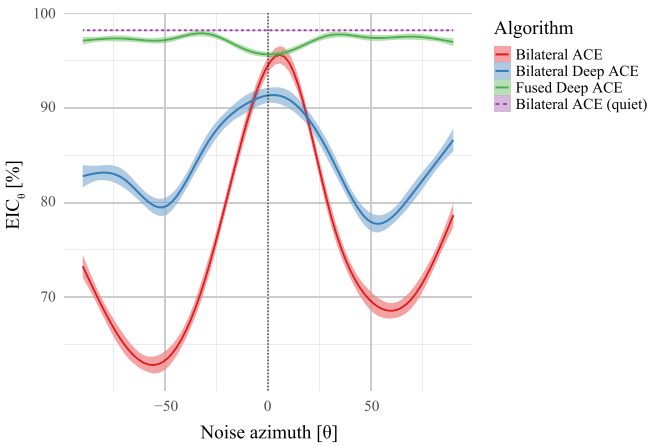

Fig. 8. Polynomial regressions showing the EIC for each azimuth averaged across electrodes, noises, and listening sides using the HSM dataset. Shaded areas represent the 95% confidence level interval [52]. Higher electrode numbers represent lower frequencies.

III. Results

A. Objective Instrumental Results:

a) SNRi: Fig. 2 shows box plots showing the mean SNRi scores across listening sides in dB for the tested algorithms in CCITT and ICRA7 noises for the different SNRs using the HSM speech dataset. In this context, the fused model demonstrates superior performance compared to the independent model. Specifically, the independent model struggles to effectively denoise speech when ICRA7 noise is present. This finding contrasts with the outcomes documented in [33]. The discrepancy might be attributed to this study's approach, as highlighted in Section II-C, where the SNR calculation is performed at the better SNR side. Consequently, significant SNR drops potentially up to 10 dB occur on the side ipsilateral to the noise. Tackling speech denoising without utilizing information from the opposite side appears to pose a considerable challenge in such scenarios.

b) V-MBSTOI: Fig. 3 illustrates the V-MBSTOI scores obtained by the evaluated algorithms in quiet. It can be seen here that the denoising algorithms do not introduce a significant drop in the V-MBSTOI scores relative to the bilateral ACE condition.

Fig. 4 presents the V-MBSTOI scores achieved by the assessed algorithms under different speech and noise conditions. Generally, the denoised signals exhibit higher scores using the denoising algorithms compared to the bilateral ACE, and the improvement is roughly proportional to the input SNR (calculated at the better SNR side).

However, it is noteworthy that the bilateral Deep ACE model falls short of the fused speech denoising method, indicating that the artifacts in the latter are comparatively smaller. Additionally, the V-MBSTOI scores computed across various input SNRs exhibit less variability for the fused Deep ACE model when compared to the bilateral Deep ACE and bilateral ACE counterparts. This suggests that the fused Deep ACE model may demonstrate greater robustness in scenarios with low input SNRs.

c) Linear cross-correlation: Fig. 5 illustrates the calculated LCCs across CI electrode numbers for each listing side (averaged across various noise conditions). The data reveals that the fused Deep ACE model exhibits superior performance in terms of channel-wise LCCs. Furthermore, the bilateral Deep ACE model falls between the bilateral ACE and fused Deep ACE algorithms, indicating that the fusion operation contributes to the enhancement of speech in BiCI listening. Fig. 6 depicts the computed linear cross-correlations with respect to the noise azimuth, considering an average across all electrodes. The data indicates that when the noise source aligns with the same side as the tested CI (where LCCs are measured), the correlation tends to decrease, as anticipated due to the lower SNRs. In contrast, for the fused Deep ACE model, the LCCs appear to remain relatively constant regardless of the azimuth of the interfering noise signal. This observation suggests that the fused Deep ACE model effectively utilizes the fusion operation by leveraging redundant information present on both listening sides.

d) Electric interaural coherence: Fig. 7 visually represents the calculated EIC as a function of the CI electrode numbers for each listening side, with the data averaged across various noise conditions. The results demonstrate that the fused Deep ACE model surpasses the other models in terms of channel-wise EIC. Additionally, the bilateral Deep ACE model occupies an intermediate position between the bilateral ACE and fused Deep ACE algorithms, suggesting that the fusion operation plays a vital role in maintaining the integrity of the speech signal across all frequencies. In Fig. 8, the calculated EIC data unveils notable trends based on the noise azimuth. The fused deep denoising model consistently maintains speech correlation, regardless of the noise source's location, showcasing its capacity to sustain speech intelligibility across various noise scenarios. Conversely, the unprocessed condition exhibits higher coherence when the noise originates from the listener's front. However, a shift occurs with the bilateral Deep ACE model, which displays greater coherence when noise is in front but reverses this trend when the noise source widens to azimuths beyond 25 degrees. This pattern suggests that the bilateral Deep ACE model may have

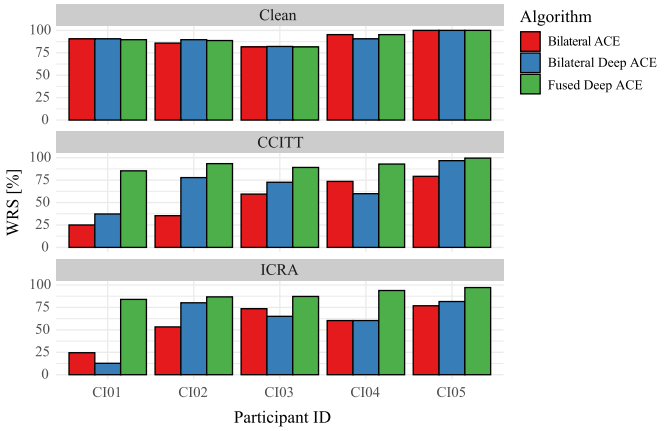

Fig. 9. Bar plots showing the mean individual word recognition scores by subject for the HSM sentence test under CCITT (left panel) and ICRA7 (right panel) noises for all tested algorithms.

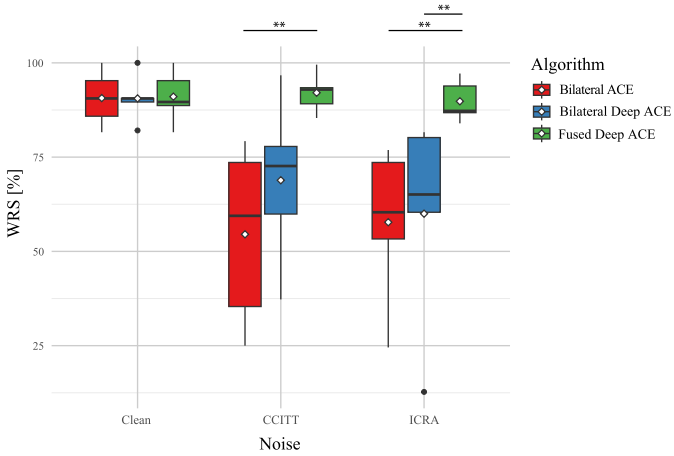

Fig. 10. Box plots of the group word recognition score measured in the five tested BiCI subjects for the three noise conditions. The black horizontal bars within each box represent the median for each condition, the diamond-shaped marks indicate the mean, and the top and bottom extremes of the boxes indicate the $Q_3 = 75\%$ and $Q_1 = 25\%$ quartiles, respectively. The box length is given by the interquartile range (IQR), used to define the whiskers that show the variability of the data above the upper and lower quartiles (the upper whisker is given by $Q_3 + 1.5{\cdot}$IQR and the lower whisker is given by $Q_1 - 1.5{\cdot}$IQR [52]). Asterisks on top of the significance bar indicate the significance level (* $p<0.05$, ** $p<0.01$, *** $p<0.001$). Black dots indicate observations that fall beyond the whisker range (outliers).

TABLE III
CI PARTICIPANT INFORMATION AND EXPERIMENT SETTINGS

ID	Age	Gender	N-of-M	CCITT SNR	ICRA7 SNR
BI01	71	M	8-of-20	5 dB	5 dB
BI02	72	F	5-of-22	0 dB	5 dB
BI03	60	F	8-of-19	0 dB	0 dB
BI04	70	M	8-of-19	0 dB	0 dB
BI05	75	M	8-of-22	–5 dB	–5 dB

limitations in handling denoising when target and interfering signals are co-located.

B. Behavioral Results

a) Speech intelligibility: Fig. 9 shows the bar plots of the individual WRS obtained by each of the tested BiCI listeners for the three tested conditions (i.e., clean, CCITT, and ICRA7). The tested SNR for each individual and noise type is shown in Table III.

Fig. 10 displays box plots illustrating the mean WRS measured in the five BiCI subjects across three noise conditions: clean, CCITT, and ICRA7. A Kruskal-Wallis test did not reveal any significant differences in mean speech intelligibility scores for the clean condition ($H(2) = 0.04$, $p = 0.98$). However, in the case of the CCITT noisy condition ($H(2) = 7.46$, $p = 0.02$) and the ICRA noisy condition ($H(2) = 9.57$, $p = 0.008$), the subsequent non-parametric Kruskal-Wallis tests did detect significant differences.

Subsequent pairwise comparisons, conducted using Wilcoxon signed-rank tests, indicated a significant distinction between the unprocessed ($M = 54.52\%$, $SD = 23.65\%$) and the fused deep ACE condition for the CCITT noise ($M = 92.07$, $SD = 5.27\%$) conditions ($p = 0.008$). Similarly, significant differences were observed between the unprocessed condition ($M = 57.74\%$, $SD = 20.90\%$) and the fused Deep ACE condition ($M = 89.81\%$, $SD = 5.49\%$) in the ICRA7 noise condition ($p = 0.008$). Additionally, in the ICRA7 noise condition, significant differences were found between the bilateral Deep ACE condition ($M = 60\%$, $SD = 28\%$) and the fused Deep ACE condition ($p = 0.008$).

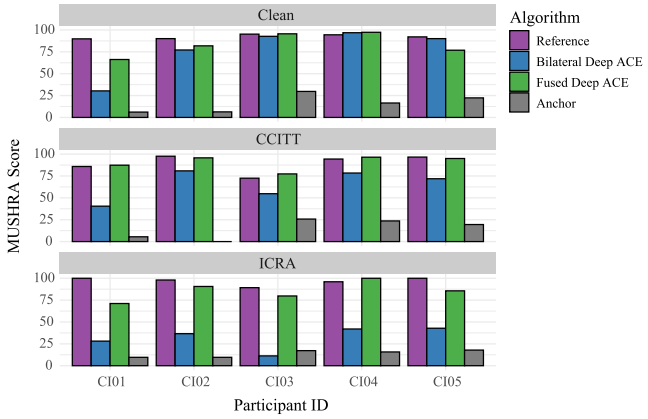

Fig. 11. Bar plots showing the mean individual MUSHRA scores for the HSM sentence test in quiet (left panel), in CCITT noise (center panel), and ICRA7 noise (right panel) noises for all tested algorithms.

b) MUSHRA: Fig. 11 shows the bar plots of the individual MUSHRA scores obtained by each of the tested BiCI listeners for the three tested noise conditions (i.e., clean, CCITT, and ICRA7). Fig. 12 illustrates box plots depicting the group MUSHRA scores obtained from five BiCI subjects under three distinct noise conditions: clean, CCITT, and ICRA7. Three separate Kruskal-Wallis tests, one for each noise condition, unveiled significant differences in the mean MUSHRA scores. Specifically, there were significant differences observed in the quiet condition ($H(3) = 10.99$, $p = 0.01$), the CCITT noise

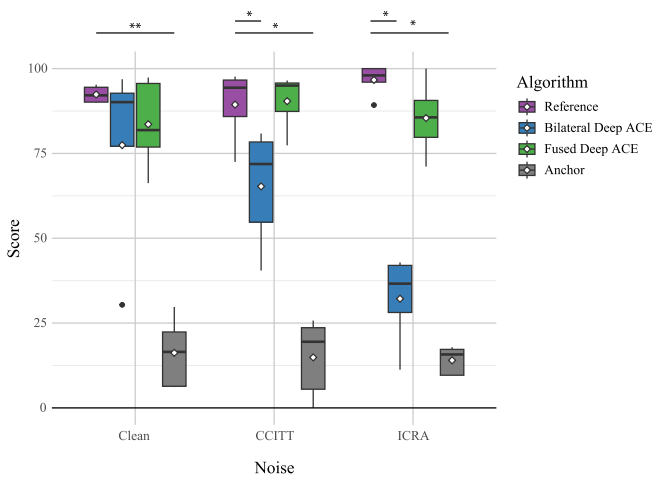

Fig. 12. Box plots of the group MUSHRA score measured in the five tested BiCI subjects for the three noise conditions. The black horizontal bars within each box represent the median for each condition, the diamond-shaped marks indicate the mean, and the top and bottom extremes of the boxes indicate the $Q_3 = 75\%$ and $Q_1 = 25\%$ quartiles, respectively. The box length is given by the interquartile range (IQR), used to define the whiskers that show the variability of the data above the upper and lower quartiles (the upper whisker is given by $Q_3 + 1.5{\cdot}$IQR and the lower whisker is given by $Q_1 - 1.5{\cdot}$IQR [52]). Asterisks on top of the significance bar indicate the significance level (* $p<0.05$, ** $p<0.01$, *** $p<0.001$). Black dots indicate observations that fall beyond the whisker range (outliers).

condition ($H(3) = 14.5, p = 0.002$), and the ICRA7 noise condition ($H(3) = 16.02, p = 0.001$).

Subsequent non-parametric Wilcoxon signed-rank pairwise comparisons further elucidated these differences. In the clean condition, the reference ($M = 92.37$, $SD = 2.46$) obtained higher scores compared to the anchor ($M = 16.23$, $SD = 10.25$) conditions ($p = 0.008$). In the CCITT noise condition, the reference ($M = 89.4$, $SD = 10.52$) received higher ratings than both the bilateral Deep ACE ($M = 65.28$, $SD = 17.2$; $p = 0.03$) and anchor ($M = 14.88$, $SD = 11.46$; $p = 0.01$) conditions. Finally, in the ICRA7 noise condition, the reference also achieved higher scores ($M = 96.65$, $SD = 4.46$) compared to the bilateral Deep ACE ($M = 32.17$, $SD = 13.09$; $p = 0.01$) and anchor ($M = 14.02$, $SD = 4.1$; $p = 0.01$) conditions.

IV. Discussion

In this study, we introduce and evaluate a novel deep learning-based strategy for sound coding in BiCIs. Our approach involves the integration of two monaural end-to-end deep denoising CI sound coding methods through fusion layers that facilitate the exchange of information between the listening sides. This exchange is achieved by combining specific latent representations generated in each monaural model. The presented fused Deep ACE model aims to replicate the ACE sound coding strategy while automatically eliminating unwanted interfering noise from the target speech, all while maintaining minimal processing latency. To be precise, this model introduces a 2 ms latency identical to the bilateral ACE setup, enabling the potential real-time application of the proposed approach. It is crucial to emphasize that the transmission of the latent representation must be considered. For instance, if we assume that only one hearing side transmits information to the contralateral side for fusion, with an encoded size of 64 and a skip connection size of 32, the data to be sent totals 192 bytes (assuming each parameter is represented by a 16-bit fixed-point variable), equivalent to 1,536 bits. Given a channel stimulation rate of 1,000 pulses per second, this information needs to be transmitted 1,000 times per second, resulting in a rate of 1,536 kbps. Additionally, factoring in the round trip where the information is sent and received back by the sender, the rate would double, reaching approximately 3 Mbps. This underscores the need for efficient coding schemes to facilitate the practical utilization of the fused Deep ACE model.

Initially, we assess the impact of fusion (fused Deep ACE) by comparing the effectiveness of speech denoising and performance with the bilateral version (bilateral Deep ACE). Furthermore, we compare our approach with the standard clinical BiCI setup, which does not incorporate any denoising (bilateral ACE). Our evaluation involves the testing of this method on speech and the assessment of speech enhancement quality in five BiCI users.

The objective instrumental measures reveal that in quiet environments, there are no discernible differences in speech intelligibility between the bilateral ACE setup, bilateral Deep ACE, and fused Deep ACE models (as shown in Fig. 3). However, in the context of speech denoising, both bilateral Deep ACE and fused Deep ACE models exhibit improvements in SNR, with the fused Deep ACE model achieving the highest. Surprisingly, the bilateral Deep ACE model performs less effectively when exposed to background ICRA7-modulated noise. This outcome is unexpected, given previous research indicating better results in a similar scenario (as reported in [33]). This discrepancy could be attributed to the lower SNR used in the current study. Nevertheless, both fused Deep ACE and bilateral Deep ACE models consistently outperform the unprocessed setup in terms of predicted speech intelligibility across all input SNRs.

To assess the extent of clean speech preservation after denoising, we employ objective measures, such as cross-channel and cross-noise azimuths' LCCs. These measures demonstrate that the fused Deep ACE model surpasses the bilateral Deep ACE model in terms of speech-denoising effectiveness and introduces fewer artifacts. This improvement is likely associated with the fused model's ability to exploit the redundancy of speech information shared between sides through the fusion layers. This result is consistent across channels and azimuths. Additionally, as expected, the bilateral Deep ACE model generally exhibits higher LCCs than the unprocessed condition, considering that the unprocessed signal retains all the original interfering noise. It is noteworthy that there is an asymmetry in the LCCs observed in both the bilateral Deep ACE and bilateral ACE conditions (as depicted in Fig. 6), with lower LCCs measured on the side ipsilateral to the noise source. This asymmetry is also present in the fused Deep ACE model, but to a lesser extent, possibly due to the sharing of speech information between sides.

In the context of BiCI listening, it is crucial to evaluate the retention of EIC after speech denoising, as low EIC has been shown to negatively affect speech intelligibility in BiCI users, as highlighted in [53]. Our assessment reveals that the fused

layer achieves the highest EIC scores when measured across azimuths and electrodes, outperforming the bilateral Deep ACE and bilateral ACE conditions. When observing EIC as a function of the azimuth (as shown in Fig. 8), all three conditions exhibit the highest EIC when speech and noise sources are co-located, aligning with expectations. In this scenario, the unprocessed condition achieves EIC scores closer to those of the fused Deep ACE model, surpassing the scores of the bilateral Deep ACE model. This shows that enhancing speech becomes easier even for the investigated models when interfering noise and target speech are spatially separated, potentially linked to the binaural unmasking phenomenon observed in human binaural hearing. This underscores the significance of higher SNR listening sides in BiCI speech denoising, particularly when speech information is shared between sides, as facilitated by the fusion layers in our approach.

The behavioral results in quiet conditions reveal no significant differences in speech intelligibility among the ACE, bilateral Deep ACE condition, and fused Deep ACE sound coding strategies, corroborating the findings from objective measures. This consistency is further confirmed by the MUSHRA test, where no discrepancies in scores are observed among the reference, bilateral Deep ACE, and fused Deep ACE conditions. However, in noisy speech conditions, speech intelligibility experiments showed that the fused Deep ACE model outperforms the bilateral ACE and bilateral Deep ACE conditions, while the bilateral Deep ACE condition surpasses ACE only in the presence of CCITT noise, failing to yield improvement when ICRA7 background noise is present. These results align with the observed SNR improvements in these conditions.

Furthermore, the MUSHRA scores indicate that all BiCI users were capable of identifying the reference and the anchor. In terms of denoising algorithms, the scores were consistently lower for the bilateral Deep ACE model compared to the reference, particularly for both CCITT and ICRA7 conditions. This concurs with the measured speech intelligibility results, implying that speech intelligibility may be significantly affected not only by the limited SNR improvement in this condition but also by the bilateral distortions introduced by the bilateral Deep ACE model.

V. Conclusion

This study underscores the potential of a fused deep learning-based BiCI sound coding strategy (fused Deep ACE) in enhancing speech, especially when speech and interfering noise sources are spatially separated. Notably, the approach's ability to retain interaural coherence compared to the bilateral Deep ACE model is highlighted. The proposed fused Deep ACE model achieved significant improvement in objective instrumental measures as well as in the listening experiments with BiCI participants. However, it is crucial to recognize that this approach may not be optimal in all listening conditions, as it may compromise binaural and spatial awareness, akin to the effects of front-end beamformers. Further research is warranted to strike a balance between achieving high speech denoising performance and maintaining spatial awareness through fusion layers, which may entail a trade-off, as outlined in [54]. Nevertheless, our presented approach exhibits promising speech-denoising performance and may prove beneficial in specific listening conditions.

References

[1] T. Lenarz, "Cochlear implant–state of the art," *Laryngorhinootologie*, vol. 96, no. 1, pp. 123–151, 2017.

[2] A. Kan and R. Y. Litovsky, "Binaural hearing with electrical stimulation," *Hear. Res.*, vol. 322, pp. 127–137, 2015.

[3] R. Y. Litovsky et al., "Simultaneous bilateral cochlear implantation in adults: A multicenter clinical study," *Ear Hear.*, vol. 27, no. 6, 2006, Art. no. 714.

[4] F. Asp et al., "Bilateral versus unilateral cochlear implants in children: Speech recognition, sound localization, and parental reports," *Int. J. Audiol.*, vol. 51, no. 11, pp. 817–832, 2012.

[5] K. C. Hughes and K. L. Galvin, "Measuring listening effort expended by adolescents and young adults with unilateral or bilateral cochlear implants or normal hearing," *Cochlear Implants Int.*, vol. 14, no. 3, pp. 121–129, 2013.

[6] J. v. Schoonhoven et al., "The effectiveness of bilateral cochlear implants for severe-to-profound deafness in adults: A systematic review," *Otol. Neurotol.*, vol. 34, no. 2, pp. 190–198, 2013.

[7] P. C. Loizou et al., "Speech recognition by bilateral cochlear implant users in a cocktail-party setting," *J. Acoustical Soc. Amer.*, vol. 125, no. 1, pp. 372–383, 2009.

[8] J. Murphy et al., "Spatial hearing of normally hearing and cochlear implanted children," *Int. J. Pediatr. Otorhinolaryngol.*, vol. 75, no. 4, pp. 489–494, 2011.

[9] I. S. Kerber and I. B. U. Seeber, "Sound localization in noise by normal-hearing listeners and cochlear implant users," *Ear Hear.*, vol. 33, no. 4, pp. 445–457, 2012.

[10] R. v. Hoesel and R. S. Tyler, "Speech perception, localization, and lateralization with bilateral cochlear implants," *J. Acoustical Soc. Amer.*, vol. 113, no. 3, pp. 1617–1630, 2003.

[11] S. R. Dennison et al., "The impact of synchronized cochlear implant sampling and stimulation on free-field spatial hearing outcomes: Comparing the ciPDA research processor to clinical processors," *Ear Hear.*, vol. 43, no. 4, pp. 1262–1272, 2022.

[12] T. Gajecki and W. Nogueira, "The effect of synchronized linked band selection on speech intelligibility of bilateral cochlear implant users," *Hear. Res.*, vol. 396, 2020, Art. no. 108051.

[13] T. Gajecki and W. Nogueira, "Enhancement of interaural level differences for bilateral cochlear implant users," *Hear. Res.*, vol. 409, 2021, Art. no. 108313.

[14] A. Kan et al., "Effect of mismatched place-of-stimulation on binaural fusion and lateralization in bilateral cochlear-implant users," *J. Acoustical Soc. Amer.*, vol. 134, no. 4, pp. 2923–2936, 2013.

[15] M. J. Goupell et al., "Effect of mismatched place-of-stimulation on the salience of binaural cues in conditions that simulate bilateral cochlear-implant listening," *J. Acoustical Soc. Amer.*, vol. 133, no. 4, pp. 2272–2287, 2013.

[16] B. Williges et al., "Coherent coding of enhanced interaural cues improves sound localization in noise with bilateral cochlear implants," *Trends Hear.*, vol. 22, 2018, Art. no. 2331216518781746.

[17] I. Hochberg et al., "Effects of noise and noise suppression on speech perception by cochlear implant users," *Ear Hear.*, vol. 13, no. 4, pp. 263–271, 1992.

[18] N. Guevara et al., "The voice track multiband single-channel modified wiener-filter noise reduction system for cochlear implants: Patients' outcomes and subjective appraisal," *Int. J. Audiol.*, vol. 55, no. 8, pp. 431–438, 2016.

[19] M. Tammen and S. Doclo, "Deep multi-frame MVDR filtering for binaural noise reduction," in *Proc. IEEE Int. Workshop Acoustic Signal Enhancement*, 2022, pp. 1–5.

[20] R. M. Baumgärtel et al., "Comparing binaural pre-processing strategies II: Speech intelligibility of bilateral cochlear implant users," *Trends Hear.*, vol. 19, 2015, Art. no. 2331216515617917.

[21] E. A. Lopez-Poveda et al., "A binaural cochlear implant sound coding strategy inspired by the contralateral medial olivocochlear reflex," *Ear Hear.*, vol. 37, no. 3, pp. 138–148, 2016.

[22] E. A. Lopez-Poveda et al., "Intelligibility in speech maskers with a binaural cochlear implant sound coding strategy inspired by the contralateral medial olivocochlear reflex," *Hear. Res.*, vol. 348, pp. 134–137, 2017.

[23] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
[24] Y.-H. Lai et al., "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1568–1578, Jul. 2017.
[25] Y.-H. Lai et al., "Deep learning–based noise reduction approach to improve speech intelligibility for cochlear implant recipients," *Ear Hear.*, vol. 39, no. 4, pp. 795–809, 2018.
[26] Y. Hu and P. C. Loizou, "Environment-specific noise suppression for improved speech intelligibility by cochlear implant users," *J. Acoustical Soc. Amer.*, vol. 127, no. 6, pp. 3689–3695, 2010.
[27] N. Mamun, S. Khorram, and J. H. L. Hansen, "Convolutional neural network-based speech enhancement for cochlear implant recipients," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 4265–4269.
[28] F. Bolner et al., "Speech enhancement based on neural networks applied to cochlear implant coding strategies," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 6520–6524.
[29] W. Nogueira et al., "Development of a sound coding strategy based on a deep recurrent neural network for monaural source separation in cochlear implants," in *Proc. IEEE Speech Commun.; 12. ITG Symp.*, 2016, pp. 1–5.
[30] T. Goehring et al., "Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users," *Hear. Res.*, vol. 344, pp. 183–194, 2017.
[31] N. Zheng et al., "A noise-robust signal processing strategy for cochlear implants using neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 8343–8347.
[32] T. Gajecki and W. Nogueira, "An end-to-end deep learning speech coding and denoising strategy for cochlear implants," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 3109–3113.
[33] T. Gajecki, Y. Zhang, and W. Nogueira, "A deep denoising sound coding strategy for cochlear implants," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 9, pp. 2700–2709, Sep. 2023.
[34] C. Han, Y. Luo, and N. Mesgarani, "Real-time binaural speech separation with preserved spatial cues," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6404–6408.
[35] R. Gu et al., "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7319–7323.
[36] T. Gajecki and W. Nogueira, "Deep latent fusion layers for binaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3127–3138, 2023.
[37] D. R. Moore, "Anatomy and physiology of binaural hearing," *Audiology*, vol. 30, no. 3, pp. 125–134, 1991.
[38] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.
[39] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1867–1871, Sep. 2010.
[40] B. Beilharz et al., "LibriVoxDeEn: A corpus for German-to-English speech translation and speech recognition," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 3590–3594.
[41] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Commun.*, vol. 9, no. 4, pp. 351–356, 1990.
[42] I. Hochmair-Desoyer et al., "The HSM sentence test as a tool for evaluating the speech understanding in noise of cochlear implant users," *Amer. J. Otol.*, vol. 18, no. 6, pp. 83–83, 1997.
[43] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Meetings Acoust.*, 2013.
[44] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, vol. 22. Berlin, Germany: Springer, 2006.
[45] W. A. Dreschler et al., "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," *Audiology*, vol. 40, no. 3, pp. 148–157, 2001.
[46] A. H. Andersen et al., "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Commun.*, vol. 102, pp. 1–13, 2018.
[47] C. H. Taal et al., "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4214–4217.
[48] R. Hinrichs et al., "A subjective and objective evaluation of a codec for the electrical stimulation patterns of cochlear implants," *J. Acoustical Soc. Amer.*, vol. 149, no. 2, pp. 1324–1337, 2021.
[49] G. D. Watkins, B. A. Swanson, and G. J. Suaning, "An evaluation of output signal to noise ratio as a predictor of cochlear implant speech intelligibility," *Ear Hear.*, vol. 39, no. 5, pp. 958–968, 2018.
[50] D. A. Freedman, *Statistics and the Scientific Method*. Berlin, Germany: Springer, 1985.
[51] J. Liebetrau et al., "Revision of Rec. ITU-R Bs. 1534," in *Proc. Audio Eng. Soc. Conv.*, 2014.
[52] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2012. [Online]. Available: http://www.R-project.org/
[53] M. Cleary et al., "Effect of experimentally introduced interaural frequency mismatch on sentence recognition in bilateral cochlear-implant listeners," *JASA Exp. Lett.*, vol. 3, no. 4, 2023, Art. no. 044401.
[54] D. Marquardt, V. Hohmann, and S. Doclo, "Interaural coherence preservation in multi-channel wiener filtering-based noise reduction for binaural hearing aids," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2162–2176, Dec. 2015.