A Deep Denoising Sound Coding Strategy for Cochlear Implants

Tom Gajecki, Yichi Zhang and Waldo Nogueira

***Abstract*—Cochlear implants (CIs) have proven to be successful at restoring the sensation of hearing in people who suffer from profound sensorineural hearing loss. CI users generally achieve good speech understanding in quiet acoustic conditions. However, their hearing ability degrades drastically in noisy backgrounds. To address this problem, current CI systems are delivered with front-end speech enhancement processors that can be beneficial for the listener, however, these perform well only in certain noisy environments, leaving quite some room for improvement in more challenging conditions. In this work, we propose substituting the CI sound coding strategy with a deep neural network (DNN) that performs end-to-end processing by taking the raw audio as input and providing a denoised electrodogram, i.e., the electrical stimulation patterns that are applied to the electrodes across time. We specifically design a DNN to emulate a common CI sound coding strategy; the advanced combination encoder (ACE). The DNN is designed not only to faithfully emulate the coding of acoustic signals that ACE would perform but also to remove unwanted interfering noises, when present, without sacrificing processing latency. The models were optimized using CI-specific loss functions and evaluated using objective instrumental measures and listening tests in CI participants. Results show that the proposed models achieve higher signal-to-noise ratio improvement and objective intelligibility scores when compared to the baseline models. Furthermore, the proposed deep learning-based sound coding strategy improved speech intelligibility in eight CI users.**

***Index Terms*—Cochlear implants, Deep neural networks, End-to-end, Speech enhancement**

I. Introduction

A cochlear implant (CI) is a surgically implanted neuroprosthetic device that restores the sensation of hearing in people who suffer from profound sensorineural hearing loss. The CI sound coding strategy is responsible for computing the electric stimulation current levels from the audio captured by the CI sound processor microphone. There are several CI sound coding strategies used in the industry [1], a widely used sound coding strategy is the continuous interleaved sampling (CIS) [2]. CIS uses a filter bank that decomposes the incoming sound into different analysis sub-band signals, which are used to modulate electric pulses that stimulate the auditory nerve. The set of pulses is sent to the available electrodes to stimulate the auditory nerve across time in an interleaved way. Other strategies perform band selection that pick the most perceptually relevant bands for stimulation. Band selection reduces power consumption without compromising speech intelligibility, the reason why is also widely used in the CI industry. Common criteria to select relevant bands are based on magnitude, used in the advanced combination encoder (ACE) [3], or on psychoacoustic masking, used in the PACE/MP3000 sound coding strategy [4]. When these CI sound coding strategies are used, electrodes located near the base of the cochlea represent higher frequencies, whereas those located in the most apical region transmit low-frequency information. In this work, we focus specifically on the ACE sound coding strategy, although the presented approach could be generalized to any available sound coding strategy, as all of them generate electrodograms (i.e., the normalized amplitudes that are subsequently mapped to the current levels that each electrode will deliver to the auditory nerve over time).

In general, a CI with its corresponding sound coding strategy can help the user to understand speech in quiet conditions, however, it fails to do so when loud interfering signals (i.e., at low signal-to-noise ratios; SNRs), such as noise or other talkers, are present [5]. In order to overcome the limitations of CIs in noisy conditions, many speech enhancement techniques have been proposed to improve speech intelligibility, such as spectral contrast enhancement [6], [7], spectral subtraction [8], Wiener filtering [9] and time-frequency masking [10]. However, recently the community is leaning towards more modern data-driven approaches, such as deep learning models, to perform single-channel speech enhancement [11]–[14]. These more recent source separation and speech enhancement methods commonly use time-frequency representations of the incoming input signals to perform feature extraction [15], [16]. These approaches can yield quite good performance but bypass potential rich sources of information, such as phase information, limiting speech separation performance. To overcome this problem, end-to-end deep learning-based approaches that directly work in the time domain have been recently proposed. For example, [17] proposed a fully-convolutional time-domain audio separation network (Conv-TasNet), a deep learning framework for end-to-end time-domain speech separation. This framework addresses the shortcomings of separation in the frequency domain, achieves state-of-the-art performance, and is suitable for low-latency applications. Thus, approaches that perform end-to-end processing are getting more and more attention in the community, which makes them an attractive solution to reduce interfering noise for CI processing. A front-end approach, however, may not fully exploit the CI processing characteristics.

In order to optimize speech enhancement for CIs, it may be beneficial to design algorithms that consider the CI processing scheme. Thus, there has been some work done specifically

TG, YZ, and WN are with the department of Otolaryngology, Medical University of Hanover and Cluster of Excellence, Hearing4all, Hannover, 30625, Germany. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project ID 446611346 (PI: Waldo Nogueira). Thanks to all the CI participants.

for CIs, where DNNs are included in the CI signal path [14], [18]. These approaches reduce noise, for example, by directly applying masks in the filter bank used by the CI sound coding strategy. Recently, inspired by the aforementioned Conv-TasNet, [19] proposed a deep learning-based end-to-end CI sound coding strategy; termed Deep ACE. Deep ACE estimates denoised CI electrodograms from raw audio data to improve speech intelligibility in CI users. Note that in this context, end-to-end implies a domain transformation (audio-to-electrodogram) designed to leverage the simplicity of the output to improve noise reduction for CIs. This may be of high importance due to the fact that electrodograms discard input phase information and also contain a high degree of redundancy in the signal. Deep ACE is designed to take advantage of such signal representation in order to extract global patterns from its characteristics, identifying which of them are the ones more likely to be embedding speech content.

In this work, we present an extensive empirical study of Deep ACE. We also introduce an improved version of it that uses CI-specific loss functions to enhance CI-specific hearing by penalizing wrong band selection. With this, we aim at achieving better noise reduction performance, to improve speech understanding in noisy conditions for CI users. Furthermore, we improve the generalization ability of the presented models. The models were evaluated using objective measures and behaviorally using listening tests in CI users to assess the potential benefits of Deep ACE with respect to the baseline algorithms and the clinical ACE sound coding strategy.

II. Methods & Materials

A. Advanced combination encoder cochlear implant sound coding strategy

The acoustic signal is first captured by the CI microphone and sampled at 16 kHz. Then, a filter bank implemented as a 128-point fast Fourier transform (FFT), commonly with a 32-point hop size, is applied, introducing a 2 ms algorithmic latency (this will depend on the channel stimulation rate; CSR). Next, an estimation of the desired envelope is calculated for each spectral band $E_k, (k = 1, ..., M)$. Each spectral band is mapped to an electrode and represents one channel. The total number of channels/electrodes is denoted by M. In this study, the band selection block sets $N = 8$ out of $M = 22$ envelopes by selecting the ones with the largest amplitudes; let A_0 be the set of available envelopes and $\tilde{Z}_N$ the set containing the N selected bands. The first element of $\tilde{Z}_N$ will be selected as follows:

$$\tilde{Z}_1 := \{\tilde{a}(z) \in A_0 : \tilde{a}(z) \geq a(z) \in A_0\}. \tag{1}$$

Now let be $A_1 := A_0 \backslash \tilde{Z}_1$, where the "$\cdot\backslash\cdot$" operator represents the subtraction of the elements belonging to the group on the right side from the group on the left side. The ith band will be added to the set of selected bands by choosing the one with the highest amplitude from the remaining elements of the set A_i like so:

$$\tilde{Z}_{i+1} := \{\tilde{a}(z) \in A_i : \tilde{a}(z) \geq a(z) \in A_i\}, \quad i = \{0, ..., N-1\}. \tag{2}$$

After eight iterations we obtain $\tilde{Z}_8$ which will contain the desired bands. Note that the elements of $\tilde{Z}_N$ cannot have the same value, that is, for any two elements z_i and $z_j \in \tilde{Z}_N$, $z_i \neq z_j$. The selected bands are subsequently non-linearly compressed by a loudness growth function (LGF) given by:

$$p_k = log(1 + \rho((E_k - s)/(m - s))/log(1 + \rho)), s \leq E_k \leq m. \tag{3}$$

The output of the LGF (p_k) represents the normalized stimulation patterns used to stimulate the auditory nerve and constitute the electrodogram. For values of E_k below base level s, p_k is set to zero, and for values of E_k above saturation level m, p_k is set to one. For a detailed description of the parameters s, m and ρ, refer to [20]. Finally, the last stage of the sound coding strategy maps p_k into the subject's dynamic range between threshold levels (THLs) and most comfortable levels (MCLs) for electrical stimulation. The *N* selected electrodes are stimulated sequentially for each audio frame, representing one stimulation cycle. The number of cycles per second thus determines the CSR. A block diagram representing the previously described processes is shown in Figure 1a; ACE.

B. Speech enhancement algorithms

1) Wiener filter (baseline 1): Here, we use a classic front-end signal processing method based on Wiener filtering, a widely used technique for speech denoising that relies on a priori SNR estimation [21] (Figure 1b; Wiener+ACE). Different variations of this algorithm are used in commercially available single-channel noise reduction systems included in CIs [22], [23]. Therefore, this classic algorithm is an appropriate baseline to use when developing new speech enhancement methods in the context of CIs [18].

2) Conv-TasNet (baseline 2): The DNN-based baseline system used in this study is the well-known conv-TasNet (which we will refer to as TasNet for simplicity) [17]. This system performs end-to-end audio speech enhancement and feeds the denoised signal to ACE, where further processing is performed to obtain the electrodograms (Figure 1c; TasNet + ACE). The TasNet structure has proven to be highly successful for single-speaker speech enhancement tasks, improving state-of-the-art algorithms, and obtaining the highest gains with modulated noise sources [24].

3) Deep ACE 1.0: This architecture is an optimization of the previously developed Deep ACE strategy described in [19]. Deep ACE 1.0 takes the raw audio input captured by the microphone and estimates the output of the LGF (Figure 1d; Deep ACE 1.0). By predicting the electrodogram $p_k \in [0, 1]$, the strategy is not only independent of individual CI fitting parameters, but it also retains the 2 ms total algorithmic delay introduced by the standard ACE strategy. The enhancer module in Deep ACE is similar to the one in TasNet+ACE, differing only in the activation function used in the encoder and in the output dimensionality of the decoder (for more details refer to [19]). The code for training and evaluating Deep ACE 1.0 can be found online[1].

[1]https://github.com/APGDHZ/DeepACE

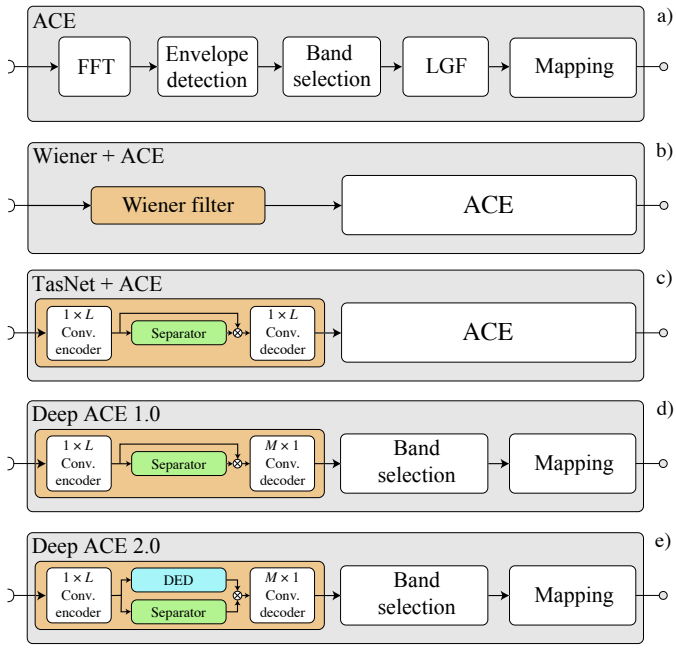

Fig. 1. Block diagrams of the four different signal processing systems. In c) and d) L refers to the *length of the filters* used in the encoder and decoder and M to the number of available electrodes for CI stimulation. Output white circles indicate the CI external emitting coil.

4) Deep ACE 2.0: This architecture is an optimization of the previously developed Deep ACE strategy described in [19]. Specifically, in this model, we introduce a deep envelope detector (DED) in the residual path of Deep ACE 1.0, which takes the role of the envelope detection block in the original ACE (see the blue block in Figure 1e; Deep ACE 2.0). This module is designed to perform dimensionality reduction between the encoder and the decoder modules in order to match the number of bands to be stimulated and also to extract other valuable features contained in the encoded signal. This operation is required for implementation reasons, specifically for the used loss function (see Section II-B6, eq. 8). For example, an input signal encoded using D filters (i.e., channels used for the input 1-D convolutions), will be reduced to M channels by the DED. This dimensionality reduction is performed by means of three stacked 1-D convolution layers. The code for training and evaluating Deep ACE 2.0 can be found online[2].

5) Model training setup: The deep learning models were trained for a maximum of 100 epochs on batches of two 4-second-long audio segments captured by a single CI. The initial learning rate was set to 1e-3. The learning rate was halved if the accuracy of the validation set did not improve during 3 consecutive epochs, early stopping with 5-epoch patience was applied as a regularization method, and only the best-performing model was saved. For optimization, Adam [25] was used to optimize the desired cost function, which depended on the algorithm's training.

6) Model training objectives: In the case of the TasNet+ACE algorithm, the optimizer was used to maximize the scale-invariant (SI) SNR [26] at the output of the TasNet. The SI-SNR between a given signal with T samples, $\boldsymbol{x} \in \mathbb{R}^{1\times T}$ and its estimate $\hat{\boldsymbol{x}} \in \mathbb{R}^{1\times T}$ is defined as:

$$\text{SI–SNR}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = 10 \cdot \log_{10}\left(\frac{||\gamma \boldsymbol{x}||^2}{||\gamma \boldsymbol{x} - \hat{\boldsymbol{x}}||^2}\right), \gamma = \frac{\hat{\boldsymbol{x}}^\top \boldsymbol{x}}{||\boldsymbol{x}||^2}. \quad (4)$$

In the case of both variations of the Deep ACE models, the decoder module will estimate the output at the LGF of ACE. For each of the Deep ACE variants, we will use a different cost function, which will be based on the mean-squared error (MSE) between the predicted and target signals, as in [19]. The MSE across electrodes between an F-frame target signal, $\boldsymbol{p} \in \mathbb{R}^{M\times F}$ and its estimate $\hat{\boldsymbol{p}} \in \mathbb{R}^{M\times F}$, is defined as:

$$\text{MSE}(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \frac{1}{M}\sum_{k=1}^{M}(\boldsymbol{p}_k - \hat{\boldsymbol{p}}_k)^2. \quad (5)$$

In Deep ACE 1.0 we extend the loss function used in [19] with an additional additive term. Specifically, inspired by the fact that CI users are sensitive to wrong band selection [27], we combine the MSE defined in 5 with a punishment term for incorrectly selected frequency bands, which we will refer to as weighted-MSE (wMSE), defined as:

$$\text{wMSE}(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \frac{1}{M}\Big(\sum_{k\in\tilde{Z}_N}(\boldsymbol{p}_k - \hat{\boldsymbol{p}}_k)^2 + w\cdot\sum_{k\in\tilde{Z}_{\overline{N}}}(\boldsymbol{p}_k - \hat{\boldsymbol{p}}_k)^2\Big), \quad (6)$$

where $\tilde{Z}_N$ is the set of bands with non-zero values, $\tilde{Z}_{\overline{N}}$ is the set of the remaining bands, and w is the weighting multiplicative factor, empirically determined through experiments and set to $w = 10$. The theoretical basis for this modification comes from experiments [28], which shows that CI users are more sensitive to distortion of the band selection compared to distortion of the signal envelopes. That is, CI users can tolerate large distortions in speech segments if the band selection is not corrupted.

With Deep ACE 2.0, we also aim at penalizing wrong band selection with a second loss function. However, in this case, the second loss function targeted the output of the mask estimation in the networks, and was used to minimize the binary cross entropy (BCE; computed in nats) between the predicted mask and the generated ideal binary mask as follows:

$$\text{BCE} = -\frac{1}{M}\sum_{k=1}^{M} e_k \cdot \log(P(e_k)) + (1 - e_k)\cdot\log(1 - P(e_k)), \quad (7)$$

where $e_k \in \{0, 1\}$ indicates if electrode k is being stimulated ($e_k = 1$) or not ($e_k = 0$), and $P(e_k)$ is the predicted probability that electrode k is being stimulated. The two loss functions (eq. 5 and eq. 7) were combined using weighting factors as follows:

$$\mathcal{L} = w_{MSE} \cdot \text{MSE} + w_{BCE} \cdot \text{BCE}. \quad (8)$$

The weighting factors w_{MSE} and w_{BCE} were set by empirical testing to 15, and 1, respectively. Note that this loss function required that the estimated mask had the same dimensionality as the LGF output, the reason why a DED had

[2]https://github.com/APGDHZ/DeepACE2.0

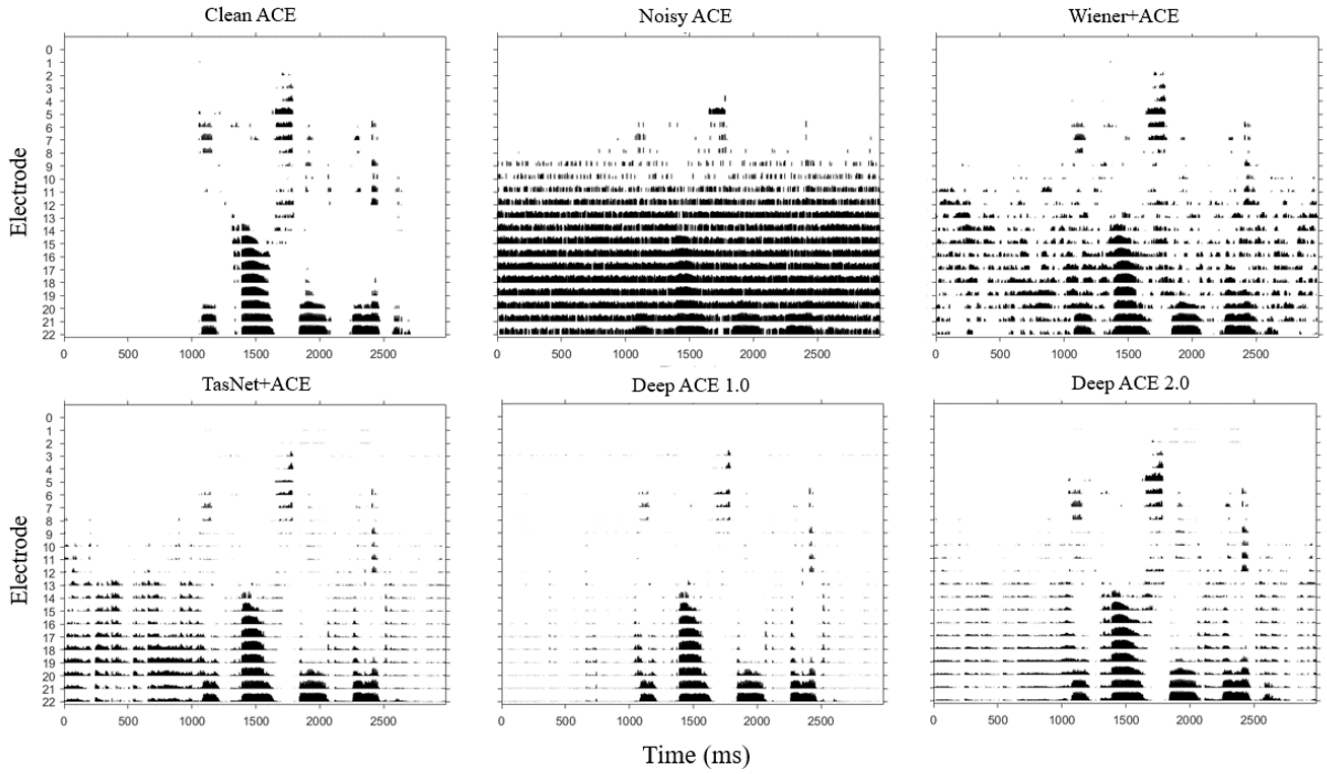

Fig. 2. Visual representations of the logarithmic growth function output in each electrode over time (electrodograms; $\boldsymbol{p}_k$) corresponding to the clean, noisy, and processed noisy speech signals processed by the different algorithms. Electrode numbers increase as the mapped frequencies decrease.

to be included in the residual path of Deep ACE 1.0 to perform dimensionality reduction (see Figure 1).

C. Audio material

In this work, a total of three different speech datasets and three noise types were used to assess the models' performance and generalization abilities. As a preprocessing stage, all samples were set to mono files and resampled to 16 kHz. Speech and noise signals were mixed at SNR values ranging from -5 to 15 dB, in 5 dB steps. The corresponding electrodograms were obtained using the ACE strategy with 1000 pulses per second (pps).

1) Speech material:

a) ***LibriVoxDeEn corpus*** *[29]:* 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 used from this corpus consisted 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.

b) ***HSM corpus*** *[30]:* It consists of 30 lists of 20 (German) everyday sentences (106 words per list). All sentences are spoken once by a male and a female speaker, yielding 600 speech files for each gender.

c) ***TIMIT corpus*** *[31]:* It 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.

2) Noise material:

a) ***Environmental noises****; DEMAND [32]:* The environmental noises recorded for this dataset are divided 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.

b) ***Synthetic noises****; SSN [33]* and ICRA *[34]:* In order to assess with which kind of noises the networks performed the best, we included synthetic stationary speech-shaped noise (SSN) and non-stationary modulated seven-speaker babble noise (ICRA).

3) Training, evaluation and testing datasets: The speech was obtained from the LibriVoxDeEn corpus and noise from the DEMAND dataset for training the models. Specifically, 30 male (M) and female (F) speakers were randomly selected from the speech corpus, and from each of the noise categories in the DEMAND noise dataset, two random environments were randomly selected. For validation, 20% of the training data was used. For the testing, the rest of the data described in section II-C was used. A detailed description of how the different datasets were created for each of the experiments is shown in Table I. The clean speech was also included as test data to investigate whether the here presented model can achieve performance close to ACE in quiet conditions without introducing artifacts.

TABLE I
DATASETS USED TO TRAIN, VALIDATE AND TEST THE MODELS

Use	Speech	# Speakers	Duration	Noise
Training	LibriVox	30M/30F	18 h.	DM
Test set 1	LibriVox	3M/3F	4.5 h.	DM/SSN/ICRA7
Test set 2	TIMIT	112M/56F	4.2 h.	DM/SSN/ICRA7
Test set 3	HSM	1M/1F	1.5 h.	DM/SSN/ICRA7
Behavioral test	HSM	1M	6.5 h.	SSN/ICRA7

D. Evaluation

1) Objective evaluation: To assess objectively the performance of each of the tested algorithms we will compute the amount of noise reduction, electrode-wise correlation coefficients between the denoised and clean signals, and a speech intelligibility index based on the STOI. Note that in this work we investigate end-to-end CI processing, so all of these objective measures will be computed from the electrodogram ($\boldsymbol{p}_k$) domain.

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

$$\mathrm{SNRi} = 10 \cdot \log_{10} \left(\frac{\sum_{k=1}^{M} ||\boldsymbol{p}_k^n - \boldsymbol{p}_k^c||^2}{\sum_{k=1}^{M} ||\boldsymbol{p}_k^d - \boldsymbol{p}_k^c||^2} \right), \tag{9}$$

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

*b) **LCC**:* To characterize potential distortions and artifacts introduced by the tested algorithms, the linear correlation coefficients (LCC) between the clean ACE electrodograms ($\boldsymbol{p}_k^c$) and the denoised electrodograms ($\boldsymbol{p}_k^d$) were computed. This channel-wise calculation allowed for the evaluation of the distortions introduced as a function of the electrode band. The LCC_k for band k is computed as follows:

$$\mathrm{LCC}_k = \frac{\mathrm{cov}\left(\boldsymbol{p}_k^c, \boldsymbol{p}_k^d\right)}{\sigma_{\boldsymbol{p}_k^c} \cdot \sigma_{\boldsymbol{p}_k^d}}, \tag{10}$$

where $\mathrm{cov}(X, Y)$ is the covariance between X and Y, and σ is the covariance of the values of denoised and clean electrodogram divided by the product of standard deviations of the values of them.

*c) **STOI**:* To estimate the speech intelligibility performance expected from each of the algorithms, the STOI score [35] was used. Although this metric is based on measurements done in normal hearing listeners, it has also been used in CI studies as a reference and can serve as a useful reference to compare relative speech enhancement performance between algorithms. The STOI score ranges from 0 to 1, where the higher score represents a predicted higher speech performance. In this work, the perceived intelligibility was estimated using the clean unprocessed speech as a reference and the vocoded denoised speech as the processed signal. The vocoded speech was obtained from the electrodograms ($\boldsymbol{p}_k$), which were used to modulate noise in the corresponding frequency bands. Each noise band was recombined to create the time domain vocoded speech.

TABLE II
LISTENER DEMOGRAPHICS AND ETIOLOGY

ID	Age	Gender	Tested side	SSN SNR	ICRA7 SNR
BI01	70	M	R	0	5
BI02	40	M	R	0	5
BI03	74	M	R	0	0
BI04	70	M	R	5	5
BI05	36	F	R	0	0
BI06	62	M	R	5	5
BI07	86	M	R	10	10
BI08	57	F	R	5	5

2) Behavioral evaluation:

*a) **Participant demographics**:* Eight postlingually deafened CI users participated in this listening test. All participants were native German speakers and had been implanted for several years. They were invited to come to the German Hearing Center of the Hannover Medical School (MHH) for a 3-hours test and the travel cost were all covered. The experiment was granted ethical approval by the MHH ethics commission. A synopsis of the patient-related data is shown in Table II.

*b) **Experimental setup**:* The test material was processed to obtain all electrodograms. The channel stimulation rate used in this work to train and evaluate the models was 1000 pps. Stimuli were delivered to the participants' CI via direct stimulation through the RF GeneratorXS interface (Cochlear Ltd.) with MATLAB via the Nucleus Implant Communicator V.3 (Cochlear Ltd.).

During the test, the participants were asked to sit in a quiet room and remove their own implants, the stimulation was delivered to the implanted receiving coil. The speech material was entirely based on the HSM sentence corpus (see Table I). Subjects were asked to repeat the presented sentence out loud as accurately as possible. All the correctly repeated words were marked, and for each test list, there was a final score that represented the percentage of understood words (word recognition score; WRS). Each listening condition was tested twice with different sentence lists, and the lists were never repeated. All the conditions were blinded to the subjects.

III. RESULTS

1) Objective instrumental results:

*a) **SNRi**:* Figure 3 illustrates the SNR improvement obtained with each of the algorithms. The results of TIMIT and LibriVoxDeEn corpora were also investigated and similar trends were observed across both datasets. The two Deep ACE algorithms achieved steady improvement in the testing datasets.

It can also be seen that the two Deep ACE models outperformed TasNet+ACE and Wiener+ACE in almost every condition, especially in low SNR environments. However, note that higher SNR improvement does not necessarily imply that the speech will be more clearly perceived, as some of the speech content may have been removed or greatly attenuated. For example, Figure 2 shows that the Deep ACE 1.0 model achieved the best noise reduction performance for unseen noise. However, some of the bands that originally contained part of the speech signal no longer do, which could potentially affect speech understanding. The Deep ACE 2.0 model, on

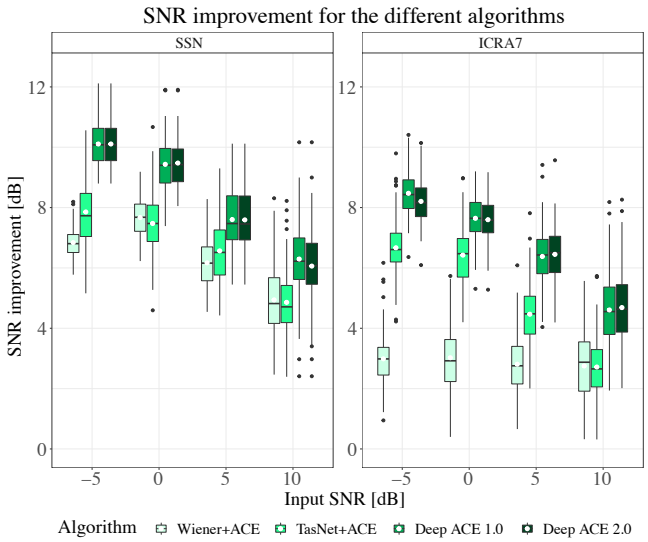

Fig. 3. SNR improvement in dB for the tested algorithms in SSN and ICRA7 noise for the different SNRs using the HSM dataset. The black horizontal bars within each of the boxes represent the median for each condition, the circle-shaped marks indicate the mean improvement, the top and bottom extremes of the boxes indicate the 75% and 25% quartiles, respectively, and whiskers indicate the variability outside the upper and lower quartiles.

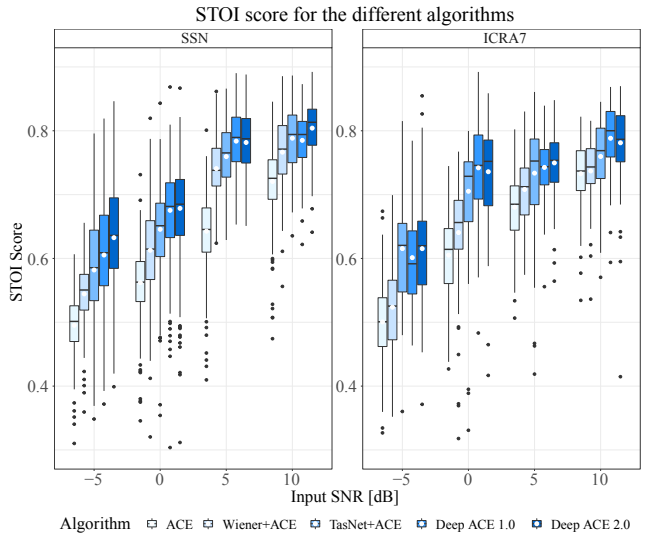

Fig. 5. STOI scores obtained by the tested algorithms in SSN and ICRA7 noise for the different SNRs using the HSM dataset. The black horizontal bars within each of the boxes represent the median for each condition, the circle-shaped marks indicate the mean improvement, the top and bottom extremes of the boxes indicate the 75% and 25% quartiles, respectively, and whiskers indicate the variability outside the upper and lower quartiles.

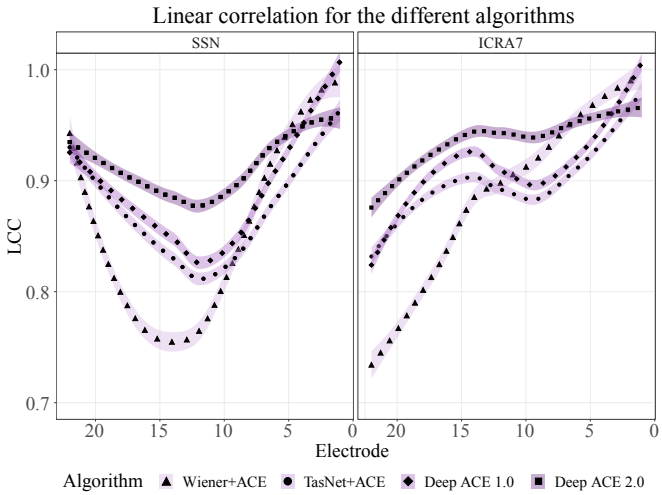

Fig. 4. Linear correlation coefficients between processed and clean electrodograms across bands of different algorithms in SSN and ICRA7 noises using the HSM dataset. Shaded areas represent the standard deviation. Higher electrode numbers represent lower frequencies.

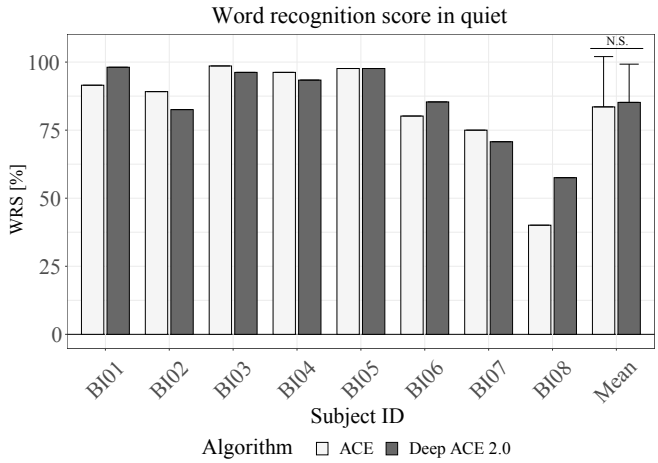

Fig. 6. Mean individual and group mean percentage of correctly understood words by subject for the HSM sentence test in quiet. Error bars indicate the standard deviation.

the contrary, could not remove all the noise but retained more speech information.

b) ***LCC:*** A band-wise linear correlation was used to assess the similarity between the original clean and processed electrodograms with the different algorithms. Figure 4 shows the obtained LCCs as a function of the correlated band. It can be seen that the Wiener+ACE condition shows the lowest correlation for the lower frequency bands and that Deep ACE 2.0 shows, in general, the highest LCCs.

c) ***STOI:*** Figure 5 illustrates the STOI scores obtained by the tested algorithms in different speech and noise conditions. All four algorithms showed improvement with respect to ACE. Three deep learning-based algorithms outperformed the Wiener+ACE. Among them, the here proposed Deep ACE 1.0 and 2.0 models were better than TasNet+ACE and they achieved similar performance. The same trend could be seen for TIMIT and LibriVoxDeEn corpora. In general, the STOI scores were consistent with the previous SNR improvement. As for clean speech (without background noise), the mean STOI scores of the HSM corpus obtained by ACE, TasNet+ACE, and the proposed Deep ACE 1.0 and 2.0 were 0.807, 0.789, 0.796, and 0.803, respectively.

2) Behavioral results: Figure 6 shows the WRS in quiet measured in eight CI subjects. The unprocessed ACE and Deep ACE 2.0 were evaluated for this condition to test whether the latter introduced any artifacts that could compromise the integrity of the clean speech signals. Results show no significant differences between the WRS obtained with ACE and Deep ACE 2.0, implying that Deep ACE 2.0 was able to properly code the incoming clean speech.

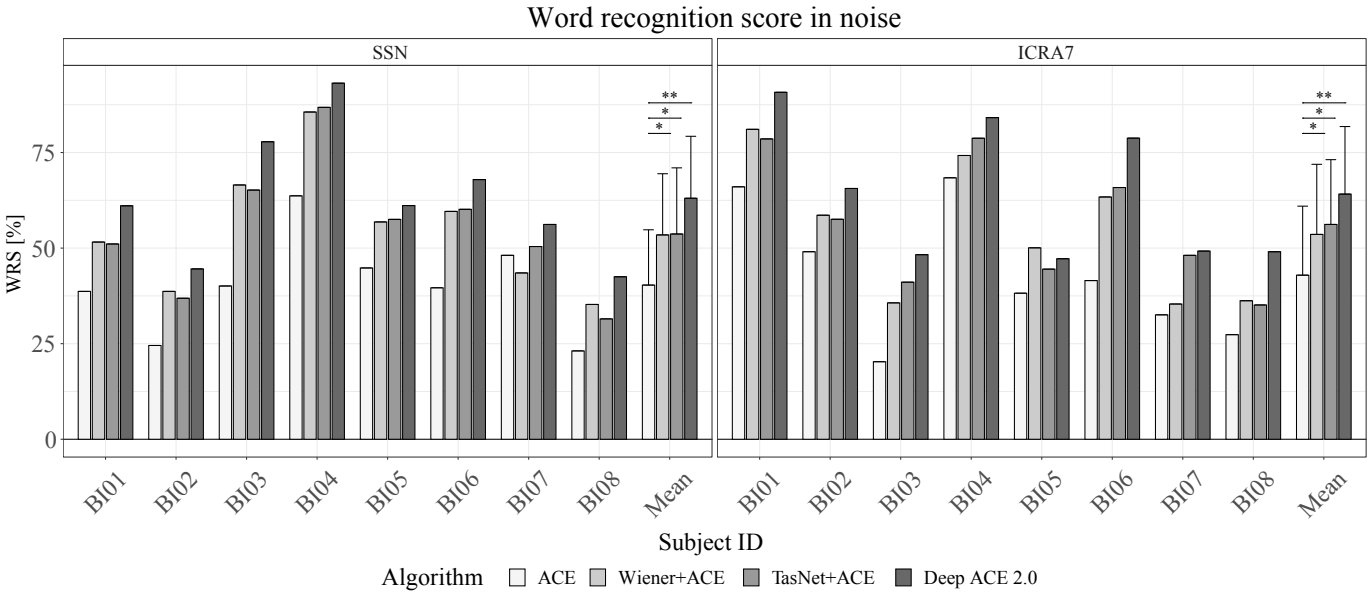

Fig. 7. Mean individual and group mean percentage of correctly understood words by subject for the HSM sentence test in noise for SSN and ICRA7 noises for all tested conditions. Error bars indicate the standard deviation.

Figure 7 shows the WRS in noise measured using the different algorithms for SSN and ICRA7 background noises. All three noise reduction algorithms were able to substantially improve speech intelligibility in both tested noises when compared to the unprocessed ACE. Among them, the Deep ACE 2.0 significantly outperformed the other two, achieving a mean improvement of about 20%.

In order to compare the WRS obtained with each of the three algorithms, the benefit in WRS with respect to ACE was computed (δWRS = WRS$_{denoised.}$ $-$ WRS$_{ACE}$). Figure 8 shows the benefit in speech understanding processed by different algorithms in SSN and ICRA7 noises. The statistical analysis reveals that Deep ACE 2.0 obtains significantly higher scores than the other algorithms in SSN. However, for the ICRA7 noise condition, Deep ACE 2.0 showed a significant benefit in WRS only when compared to the Wiener+ACE, although a trend towards higher scores is observed.

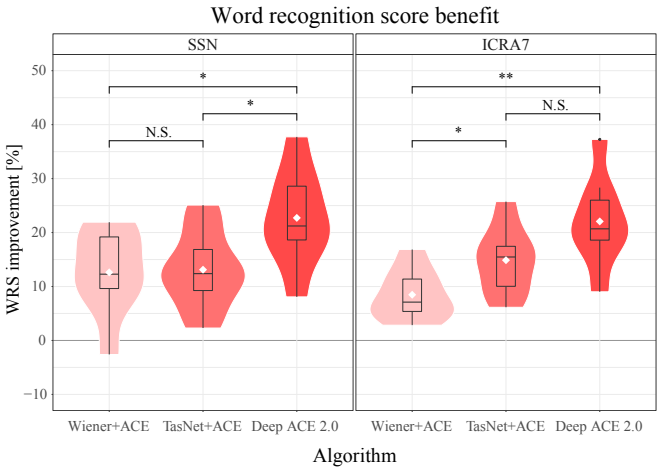

Fig. 8. Violin plots showing the WRS improvement by processing the noisy signals with the different algorithms compared to ACE. The black horizontal bars within each of the boxes represent the median for each condition, the diamond-shaped marks indicate the mean improvement, the top and bottom extremes of the boxes indicate the 75% and 25% quartiles, respectively, and whiskers indicate the variability outside the upper and lower quartiles.

IV. Discussion

Here we propose an end-to-end speech coding and denoising strategy for CIs. The vast majority of speech enhancement algorithms for CI users rely on front-end processing, for this reason, in this work we investigate end-to-end CI deep learning models that merge the preprocessing denoising stage with the CI sound coding strategy. This approach takes advantage of the simplicity of the output to be predicted, the electrodogram, which does not contain any phase information and may potentially facilitate noise reduction for CIs by removing unwanted redundancies in the input signal. Merging the noise reduction algorithm with the CI sound coding strategy also has the benefit of reducing processing latency, when compared to other front-end approaches. For example, when a front-end TasNet denoising block would introduce 4 ms of latency, whereas the here presented Deep ACE models introduce only 2 ms. This may be crucial for devices such as CIs that need to transmit signals with minimal latency. Besides, the Deep ACE 2.0 has a low computational complexity with 552k parameters in total, which makes it possible to deploy the model for real-time applications.

This study extends previous preliminary work [19], where Deep ACE was introduced for the first time, by optimizing its architecture, improving its speech enhancement performance, and improving its generalization capabilities. Specifically, Deep ACE 1.0 and 2.0 were optimized by adjusting the hyper-parameters, the topology (in the case of Deep ACE 2.0), and loss functions that targeted erroneous band selection, based on some hearing perception aspects of CI users. Although the implemented loss functions for the Deep ACE 1.0 and 2.0 models relied on the same idea of penalizing the non-

stimulating channels, Deep ACE 2.0 seemed to achieve a smoother noise reduction by preserving some noisy residuals.

The objective results indicate that in contrast to [19], the two proposed end-to-end CI speech enhancement models outperform both front-end baseline algorithms, in terms of SNR improvement and predicted speech intelligibility. Specifically, we observe that Deep ACE 2.0 obtained the highest performance whereas the Wiener+ACE obtains the poorest. The objective results also show that the trained models are resistant to unseen data, that is, they are capable of generalizing to new noise and speech signals. This is also an improvement achieved in this study with respect to [19], where part of the testing material was also included in the training set.

The behavioral speech tests show that the proposed end-to-end deep learning coding strategy Deep ACE 2.0 can be used to code clean speech, as shown in Figure 6, where no significant differences in speech understanding were observed between the original clean ACE and clean Deep ACE 2.0. Furthermore, speech understanding in noise measured in this study is consistent with the objective measures. That is, from Figure 9 we expected no significant differences in speech performance between TasNet+ACE and Wiener+ACE when using SSN background noise, but to see a significant improvement with the TasNet+ACE with respect to Wiener+ACE when using ICRA7 modulated noise, which happens to be the case (see Figure 8). However, it is important to point out that all tested algorithms tested behaviorally outperform the baseline ACE sound coding strategy for both SSN and ICRA7 noises. Although the observed improvement using the Wiener+ACE using ICRA7 background noise was not observed in [19], it is still consistent with other previous studies [9]. Also, this benefit in non-stationary noise may be related to the fact that this condition was tested mostly at positive SNRs (Table II).

V. Conclusions

In this work, an end-to-end deep learning speech denoising and coding strategy to enhance the speech intelligibility of CI users was presented. The inspiration is the recently proposed Deep ACE, which effectively reduces the noise. In this context, two new implementations were proposed to optimize the Deep ACE for the better hearing performance of CI users. These two models were then evaluated using objective instrumental measures and the better-performing model was behaviorally tested by means of listening tests with CI users. The performance was compared to the ACE and two front-end baseline models, namely Wiener+ACE and TasNet+ACE. Results showed that the optimized Deep ACE 1.0 and 2.0 models outperformed the other baseline models, and brought great and steady objective improvement. Furthermore, the results show that Deep ACE 2.0 also gave a behavioral benefit in WRS when compared to the baseline models. Based on these findings, the present study suggests that Deep ACE 2.0 has the potential as a method to improve speech understanding under noisy conditions for CIs.

References

[1] J. Wouters, H. J. McDermott, and T. Francart, "Sound coding in cochlear implants: From electric pulses to hearing," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 67–80, 2015.

[2] B. S. Wilson, C. C. Finley, D. T. Lawson, R. D. Wolford, D. K. Eddington, and W. M. Rabinowitz, "Better speech recognition with cochlear implants," *Nature*, vol. 352, no. 6332, pp. 236–238, 1991.

[3] P. Seligman, H. McDermott *et al.*, "Architecture of the spectra 22 speech processor," *Annals of Otology, Rhinology and Laryngology*, vol. 104, no. suppl 166, pp. 139–141, 1995.

[4] W. Nogueira, A. Büchner, T. Lenarz, and B. Edler, "A psychoacoustic "nofm"-type speech coding strategy for cochlear implants," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 18, pp. 1–16, 2005.

[5] I. Hochberg, A. Boothroyd, M. Weiss, and S. Hellman, "Effects of noise and noise suppression on speech perception by ci users," *Ear and Hearing*, vol. 13, pp. 263–271, 1992.

[6] W. Nogueira, T. Rode, and A. Büchner, "Spectral contrast enhancement improves speech intelligibility in noise for cochlear implants," *The Journal of the Acoustical Society of America*, vol. 139, no. 2, pp. 728–739, 2016.

[7] T. Baer, B. C. Moore, and S. Gatehouse, "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality, and response times," *Journal of rehabilitation research and development*, vol. 30, pp. 49–49, 1993.

[8] L.-P. Yang and Q.-J. Fu, "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise," *The journal of the Acoustical Society of America*, vol. 117, no. 3, pp. 1001–1004, 2005.

[9] N. Guevara, A. Bozorg-Grayeli, J.-P. Bebear, M. Ardoint, S. Saaï, D. Gnansia, M. Hoen, P. Romanet, and J.-P. Lavieille, "The voice track multiband single-channel modified wiener-filter noise reduction system for cochlear implants: patients' outcomes and subjective appraisal," *International Journal of Audiology*, vol. 55, no. 8, pp. 431–438, 2016.

[10] R. Koning, N. Madhu, and J. Wouters, "Ideal time–frequency masking algorithms lead to different speech intelligibility and quality in normal-hearing and cochlear implant listeners," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 1, pp. 331–341, 2014.

[11] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1568–1578, 2016.

[12] Y.-H. Lai, Y. Tsao, X. Lu, F. Chen, Y.-T. Su, K.-C. Chen, Y.-H. Chen, L.-C. Chen, L. P.-H. Li, and C.-H. Lee, "Deep learning–based noise reduction approach to improve speech intelligibility for cochlear implant recipients," *Ear and hearing*, vol. 39, no. 4, pp. 795–809, 2018.

[13] Y. Hu and P. C. Loizou, "Environment-specific noise suppression for improved speech intelligibility by cochlear implant users," *The Journal of the Acoustical Society of America*, vol. 127, no. 6, pp. 3689–3695, 2010.

[14] N. Mamun, S. Khorram, and J. H. Hansen, "Convolutional neural network-based speech enhancement for cochlear implant recipients," in *Interspeech*, vol. 2019. NIH Public Access, 2019, p. 4265.

[15] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International conference on latent variable analysis and signal separation*. Springer, 2015, pp. 91–99.

[16] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[17] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[18] F. Bolner, T. Goehring, J. Monaghan, B. Van Dijk, J. Wouters, and S. Bleeck, "Speech enhancement based on neural networks applied to cochlear implant coding strategies," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6520–6524.

[19] T. Gajecki and W. Nogueira, "An end-to-end deep learning speech coding and denoising strategy for cochlear implants," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3109–3113.

[20] W. Nogueira, A. Büchner, T. Lenarz, and B. Edler, "A psychoacoustic" nofm"-type speech coding strategy for cochlear implants," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 18, pp. 1–16, 2005.

[21] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2. IEEE, 1996, pp. 629–632.

[22] S. J. Mauger, K. Arora, and P. W. Dawson, "Cochlear implant optimized noise reduction," *Journal of Neural Engineering*, vol. 9, no. 6, p. 065007, 2012.

[23] A. Büchner, M. Brendel, H. Saalfeld, L. Litvak, C. Frohne-Buechner, and T. Lenarz, "Results of a pilot study with a signal enhancement algorithm for HiRes 120 cochlear implant users," *Otology & neurotology*, vol. 31, pp. 1386–90, 2010.

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

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

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

[27] B. van Dijk, M. Moonen, J. Wouters *et al.*, "Understanding the effect of noise on electrical stimulation sequences in cochlear implants and its impact on speech intelligibility," *Hearing Research*, vol. 299, pp. 79–87, 2013.

[28] O. U. R. Qazi, B. van Dijk, M. Moonen, J. Wouters *et al.*, "Understanding the effect of noise on electrical stimulation sequences in cochlear implants and its impact on speech intelligibility," *Hearing Research*, vol. 299, pp. 79–87, 2013.

[29] B. Beilharz, X. Sun, S. Karimova, and S. Riezler, "Librivoxdeen: A corpus for german-to-english speech translation and speech recognition," in *Proceedings of LREC*, 2020.

[30] I. Hochmair-Desoyer, E. Schulz, L. Moser, and M. Schmidt, "The hsm sentence test as a tool for evaluating the speech understanding in noise of cochlear implant users." *The American journal of otology*, vol. 18, no. 6 Suppl, pp. S83–S83, 1997.

[31] V. Zue, S. Seneff, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech communication*, vol. 9, no. 4, pp. 351–356, 1990.

[32] J. Thiemann, N. Ito, and E. Vincent, "Demand: a collection of multi-channel recordings of acoustic noise in diverse environments," in *Proc. Meetings Acoust*, 2013, pp. 1–6.

[33] H. Fastl and E. Zwicker, "Psychoacoustics - facts and models," *Springer*, vol. 3rd edition, 2007.

[34] W. A. Dreschler, H. Verschuure, C. Ludvigsen, and S. Westermann, "Icra noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment: Ruidos icra: Señates de ruido artificial con espectro similar al habla y propiedades temporales para pruebas de instrumentos auditivos," *Audiology*, vol. 40, no. 3, pp. 148–157, 2001.

[35] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.