AN END-TO-END DEEP LEARNING SOUND CODING STRATEGY FOR COCHLEAR IMPLANTS

Tom Gajecki & Waldo Nogueira

Department of Otolaryngology,
Medical University Hannover and Cluster of Excellence Hearing4all,
Hannover, 30625, Germany
gajecki.tomas@mh-hannover.de, nogueiravazquez.waldo@mh-hannover.de

ABSTRACT

Cochlear implant (CI) users struggle to understand speech in noisy conditions. In this work, we propose an end-to-end speech coding and denoising sound coding strategy that estimates the electrodograms from the raw audio captured by the microphone. We compared this approach to a classic Wiener filter and TasNet to assess its potential benefits in the context of electric hearing. The performance of the network is assessed by means of noise reduction performance (signal-to-noise-ratio improvement) and objective speech intelligibility measures. Furthermore, speech intelligibility was measured in 5 CI users to assess the potential benefits of each of the investigated algorithms. Results suggest that the speech performance of the tested group seemed to be equally good using our method compared to the front-end speech enhancement algorithm.

Index Terms— Cochlear Implant, Deep Learning, Sound Coding Strategy, Speech Enhancement

1. INTRODUCTION

A cochlear implant (CI) is a surgically implanted medical device that can restore hearing to a profoundly deaf person. In general, CI users achieve good speech intelligibility in quiet conditions. When compared to normal-hearing listeners, however, CI users need significantly higher signal-to-noise ratios (SNRs) to achieve the same speech intelligibility [1]. This fact motivates researchers to investigate different speech enhancement techniques to improve the SNR of the incoming signal in acoustically challenging conditions [2].

The CI sound coding strategy is responsible for computing the electric stimulation currents from the audio captured by the microphone's CI sound processor. It uses a filter bank that decomposes the incoming sound into different analysis subband signals, which are used to encode electric pulses to stimulate the auditory nerve.

Previous research has shown that single-channel noise reduction algorithms can be used as front-end processors prior to the sound coding strategy to improve speech intelligibly of CI users [3, 4]. Single-channel noise reduction algorithms convert the signal into the spectral domain and apply masks to emphasize the signal in the frequency bands containing high SNRs while attenuating the noisy bands to obtain an overall enhancement of the target signal [5, 6]. The algorithms are based on statistical signal processing methods that include spectral subtraction and Wiener filtering [7, 8]. The challenge of estimating accurate masks for CI users while minimizing distortions on speech signals still remains a challenge. Classic Wiener filters as the ones used in commercial CI sound processors provide limited or no benefit under non-stationary conditions.

Alternatively, improved speech intelligibility for speech in non-stationary noise can be achieved using directional algorithms, but assume that the target speech and masking noise are spatially separated [5]. More recently, data-driven approaches based on deep neural networks (DNNs), have been successfully developed to improve speech understanding in non-stationary background conditions for CI listeners [9, 10]. These algorithms, however, act as front-end processors and are not well integrated into the CI sound coding strategy. In order to optimize speech enhancement for CIs, it may be beneficial to design algorithms that consider the CI processing scheme. There has been also work done more specific to CIs, where DNNs are included in the CI's signal path [11, 12]. These approaches perform noise reduction, for example, by using DNNs that inform the CI about which bands to select or by directly applying denoising masks in the filterbank used by the sound coding strategy of the CI. However, these approaches rely on the spectrum of the sound as input to the DNN or on extra spectro-temporal features [11]. In the last years, several audio processing models were recently proposed that operate directly on time-domain audio signals, for speech denoising or audio source separation [13, 14, 15, 16]. These end-to-end approaches offer advantages as fewer assumptions related to the magnitude and phase of the spectrum are required, while obtaining high performance.

Here we propose an end-to-end speech coding and enhancement method that takes the raw audio captured by the CI's microphone and outputs the levels that are encoded to the CI electrodes to stimulate the auditory nerve. Our approach is based on an end-to-end single-channel-speech enhancement algorithm [13]. This new approach is meant to completely bypass the CI sound coding strategy while providing the listener with signals as natural as the original sound coding strategy. Moreover, the end-to-end strategy may outperform a front-end DNN combined with a CI sound coding strategy, because the output electrodogram has reduced dynamic range, lower amplitude resolution, has no phase information and is more redundant [17] than the raw audio signal, therefore, may be easier to model.

The organization of the manuscript is as follows: section 2 presents the methods and materials, section 3 the evaluation of the speech enhancement algorithms using objective instrumental measures, and speech intelligibility tests in CI users. Section 4 presents the results and we conclude the manuscript in Section 4.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project ID 446611346. Thanks to all the CI participants that took part in this study. Also thank you to Sina Tahmasebi for his help recruiting the CI users for the listening experiments.

2. METHODS & MATERIALS

2.1. Investigated Algorithms

Advanced combination encoder (ACE): An audio signal is captured by the CI's microphone at 16 kHz. Next, a filter bank implemented as a fast Fourier transform (FFT) is applied to the signal. After that, an estimation of the desired envelope is calculated for each spectral band $E_k, (k = 1, ..., M)$. Each band is allocated to one electrode and represents one channel. For each frame of the audio signal, out of the M channels, N channels with the highest amplitudes are selected. Typical values for M and N are 22 and 8, respectively. The selected bands are then non-linearly compressed through the loudness growth function (LGF): $p_k = log(1+\rho((E_k-s)/(m-s))/log(1+\rho))$ if $s \leq E_k \leq m$. For values of E_k below base level s, p_k is set to 0, and for values of E_k above saturation level m, p_k is set to 1. For a detailed description of the parameters s, m and ρ, refer to [18]. Finally, the last stage of the sound coding strategy maps p_k into the subject's dynamic range between their thresholds (THLs) and most comfortable levels (MCLs) for electrical stimulation. For each frame of the audio signal, N electrodes are stimulated sequentially, representing one stimulation cycle. The number of cycles per second thus determines the channel stimulation rate (CSR). A block diagram representing the previously described processes is shown in Fig. 1a. The graphical representation of the current applied to each electrode over time is called an electrodogram (Fig. 2).

Baseline speech enhancement algorithm (Wiener): Here we use a front-end signal processing method based on Wiener filtering, a widely used technique for speech denoising that relies on a priori SNR estimation [8] (Figure 1b; Wiener + ACE). We used this algorithm as a front-end baseline for its similarity to commercially available noise reduction systems [19, 20]. Also, this method has been previously used as a baseline in studies that investigated state-of-the-art speech enhancement methods [15, 11].

Baseline end-to-end speech enhancement algorithm (TasNet): The DNN based baseline system used in this study is the well-known TasNet [13]. This system performs end-to-end audio speech enhancement and feeds the denoised signal to ACE, where further processing is performed to obtained 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, obtaining the highest gains with modulated noise sources [21].

End-to-end sound coding strategy for CIs (Deep ACE): Here we propose a new strategy that combines the ACE with the structure of TasNet [13]. Deep ACE takes the raw audio input and estimates the output of the LGF (Figure 1d). By predicting $p_k \in [0, 1]$, the network does not depend on individual fitting parameters and can generalize to different CI listeners.

The enhancer module of the deep ACE is similar to the one in TasNet+ACE, differing mainly in the activation functions used in the encoder and the output dimensionality of the decoder (Figure 1 c and d). The activation function used in deep ACE is given by:

$$\phi(x) = \begin{cases} \alpha x, & \text{if } x \geq 0. \\ -\beta x, & \text{otherwise,} \end{cases} \quad (1)$$

where $\{(\alpha, \beta) \in \mathbb{R}^{0+} \times \mathbb{R}^{0+}\}$ are trainable scalars that control the positive and negative slope of the rectifier. This activation function guarantees that the coded signal is represented by numbers greater than zero. The other difference between the enhancer blocks of TasNet + ACE and deep ACE is related to the output dimensionality. The TasNet enhancer module estimates an output in the time domain, every temporal convolutional window, whereas deep ACE will estimate the LGF output in the frequency domain, ready to perform band selection. The code for training and evaluating deep ACE can be found online[1]

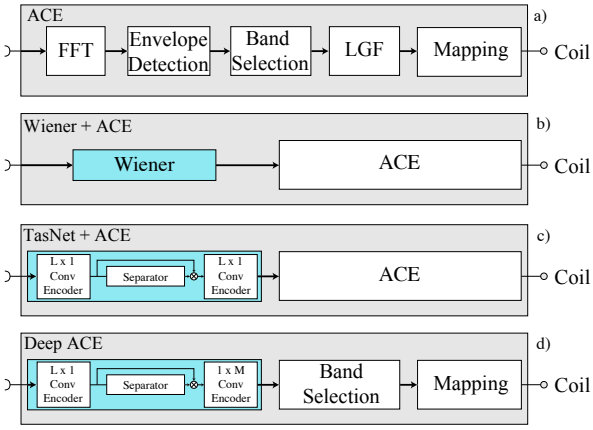

Fig. 1. Block diagrams of the four different signal processing systems. In c) and d) L refers to the length of filters used in the encoder block (refer to Table 1).

2.2. Datasets

Dataset 1: The audio dataset was provided by the 1st Clarity Enhancement Challenge [22]. It consists of 6,000 scenes including 24 different speakers. The development dataset, used to monitor the models' performance, consists of 2,500 scenes including 10 target speakers. Each scene corresponds to a unique target utterance and a unique segment of noise from an interferer, mixed at SNRs ranging from -6 to 6 dB. The three sets are balanced for the target speaker's gender. Binaural room impulse responses (BRIRs) were used to model a listener in a realistic acoustic environment. The audio signals for the scenes are generated by convolving source signals with the BRIRs and summing. BRIRs were generated for hearing aids located in each listening side, providing 3 channels each (front, mid, rear). From which only the front microphone was used.

Dataset 2: In addition to dataset 1, the Hochmair, Schulz, Moser (HSM) sentence test [23], composed of 30 lists with 20 everyday sentences each (106 words per list) was included. The HSM sentences were mixed with interfering multiple-speaker-modulated speech-weighted noise source (ICRA7) [24] and interfering Consultatif International Téléphonique et Télégraphique (CCITT) noise [25], at SNRs ranging from -5 to 5 dB. Speech and noise signals were convolved with a BRIR [26] and presented in a virtual acoustic scenario at a distance of 80 cm in front of the listener.

Train, validation and test datasets: All data were downsampled to 16 kHz. To train the models, the training set of dataset 1 was mixed with 30% of dataset 2. To validate and optimize the

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

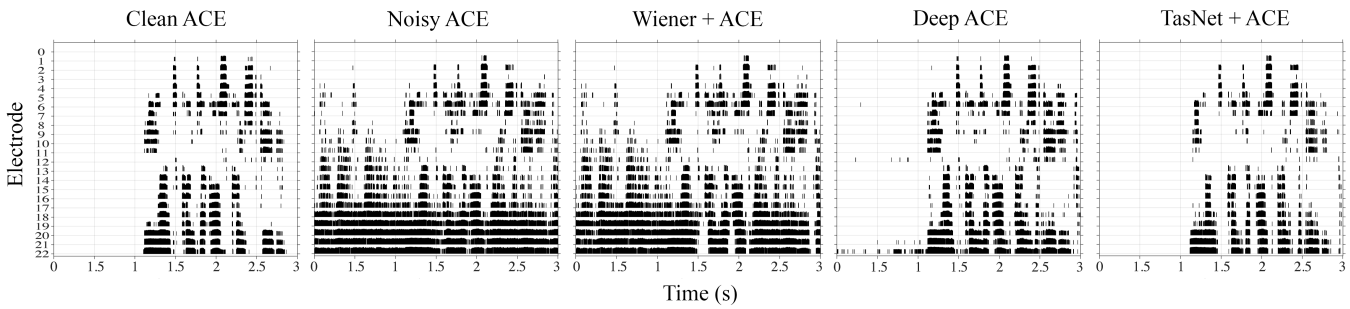

Fig. 2. Electrodograms for the clean and nosy speech produced by ACE (first two electrodograms) and the electrodograms produced by the enhancing algorithms.

models, the validation set of dataset 1 was used. Finally, for the final evaluations, the remaining 70% of dataset 2 was used.

2.3. Model Training

The models were trained for a maximum of 100 epochs on batches of two 4 s 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 a patience of 5 epochs was applied as a regularization method, and only the best performing model was saved. For the model optimization, Adam [27] was used to optimize the desired cost function, which depended on the algorithm to be trained.

TasNet + ACE cost function: In the case of the TasNet + ACE algorithm, the optimizer was used to maximize the scale-invariant (SI) SNR [28] 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:

$$\text{SI–SNR}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = 10 \cdot log_{10}\Big(\frac{|\gamma \boldsymbol{x}^2||_2^2}{||\gamma \boldsymbol{x}^2 - \hat{\boldsymbol{x}}^2||_2^2}\Big), \gamma = \frac{\hat{\boldsymbol{x}}^\top \boldsymbol{x}}{||\boldsymbol{x}^2||_2^2}. \quad (2)$$

Deep ACE cost function: Because the enhancer module will estimate the output at the LGF of ACE, the optimizer will be used to minimize the mean-squared-error (MSE) between the predicted and target signals. The MSE between a given true signal with K frames, $\boldsymbol{p} \in \mathbb{R}^{K\times 1}$ and its estimate $\hat{\boldsymbol{p}} \in \mathbb{R}^{K\times 1}$ is defined as:

$$MSE(\boldsymbol{p}, \hat{\boldsymbol{p}}) = \frac{1}{M}\sum_{k=1}^{M}(\boldsymbol{p}_k - \hat{\boldsymbol{p}}_k)^2. \quad (3)$$

The models were trained and evaluated using a PC with an Intel(R) Xeon(R) W-2145 CPU @ 3.70GHz, 256 GB of RAM, and an NVIDIA TITAN RTX as the accelerated processing unit.

3. EVALUATION

3.1. Objective Instrumental Evaluation

SNR Improvement: For a given algorithm, the SNR (eq. 2 with $\gamma = 1$) improvement with respect to the unprocessed signal (noisy ACE) will be reported. This is simply computed as follows: $SNRi = SNR_{proc.} - SNR_{unproc.}$.

To obtain the processed signals in the time domain for each of the algorithms, the generated electrodograms were resynthesized using a sine vocoder with a THL of 100 and an MCL of 150 clinical units (refer to [18]). Then, equation 2 was applied to the corresponding vocoded signals.

STOI: The generated electrodograms were resynthesized using the vocoder (described in the previous subsection). The original noiseless, clean speech signals served as reference signals (raw speech signals captured by the microphone). The resynthesized audio waveforms and the reference signals were used to compute the short-time objective intelligibility (STOI) measure [29].

Hyper-parameter optimization: To assess which model size was the best to train the algorithms, we factorized the problem by examining the effect on the training error as a function of the skip connection size. We perform 5 independent training sessions for each of skip connection size. The models with the lowest errors were chosen. Table 1 shows the used hyper-parameters of the implemented models. For a detailed description of these hyper-parameters refer to [13].

Description	Value
Number of filters in autoencoder	64
Length of the filters	32
Number of channels in the bottleneck blocks	64
Number of channels in the skip-connections	32
Number of channels in the convolutional blocks	128
Kernel size in convolutional blocks	128
Number of convolutional blocks in each repeat	3
Number of repeats	2

Table 1. Hyper-parameters used for training the models.

Table 2 shows the validation loss and number of parameters of the resulting evaluated deep learning models.

Evaluated end-to-end models		
	Deep ACE	TasNet+ACE
Validation Loss	0.002	9.04 dB
# Parameters	462,444	447,296

Table 2. Validation loss and number of parameters of the evaluated deep learning models.

3.2. Listening Evaluation

Participants: 5 postlingually deafened CI users participated in the study. All participants were native German speakers and traveled to the Hannover Medical School (MHH) for a 2-hour listening test and their travel costs were covered. The experiment was granted with ethical approval by the MHH ethics commission. A synopsis of the pertinent patient-related data is shown in Table 3.

ID	Age [yrs]	Gender	Clinical CSR	SNR [dB]
BI01	63	M	900	0
BI02	69	M	900	0
BI03	69	M	900	0
BI04	52	F	900	0
BI05	85	M	900	5

Table 3. Listener demographics and etiology. The clinical CSR expressed in pulses per second (pps) is the one that participants were using in their clinical speech processors. The last column indicates the SNR at which eat subject was tested.

Test scenario: For the listening experiments in CIs, the remaining 70% of the test dataset, mixed with ICRA7 noise, was used. Stimuli were delivered via direct stimulation through the RF GeneratorXS interface (Cochlear Ltd., Sydney, Australia) with MATLAB (Mathworks, Natick, MA) via the Nucleus Implant Communicator V.3 (Cochlear Ltd., Sydney, Australia). The CSR used in this study to train and evaluate the models was 1000 pps. Speech intelligibility in noise was measured by means of the HSM sentence test [23]. Subjects were asked to repeat sentences out loud as accurately as possible. Each listening condition was tested twice with different sentence lists, then the final score was computed by taking the mean number of correct words for each condition. The conditions were blinded to the subjects.

4. RESULTS

4.1. Results from Objective Instrumental Measures

SNR improvement: Figure 3 shows the SNRi of the investigated algorithms w.r.t. ACE, for two different background noises and three input SNRs.

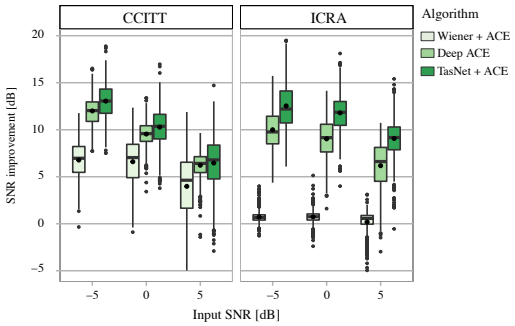

Fig. 3. SNRi in dB for the tested algorithms in CCITT noise and in ICRA noise for the different skip connection sizes.

STOI: The mean STOI scores obtained by the baseline CI sound coding strategy, by deep ACE and TasNet+ACE algorithms for speech signals with no interfering noise were 0.8, 0.78, and 0.79, respectively. Figure 4 shows the STOI results of the investigated algorithms, for two different noise interfere types and three input SNRs.

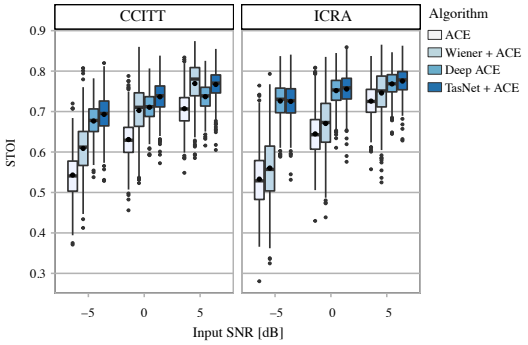

Fig. 4. STOI scores obtained by the tested algorithms in CCITT noise and in ICRA noise for the different skip connection sizes.

4.2. Listening Results

Figure 5 shows the percentage of understood words in quiet and mixed with ICRA7 noise at an input SNR indicated in Table 3.

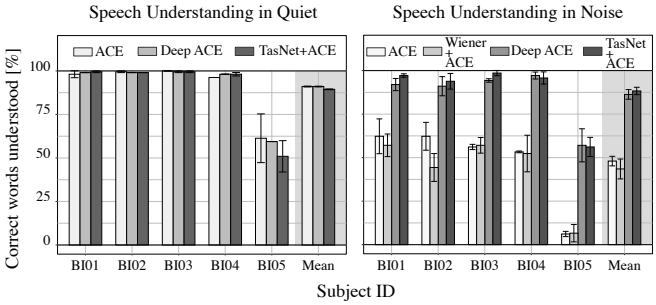

Fig. 5. Individual and mean percentage of correct understood words by subject for the HSM sentence test in quiet and in ICRA7 noise.

5. CONCLUSIONS

In this work, we have presented an adaptation of TasNet's model for speech denoising to a CI sound coding strategy; deep ACE. This approach allows reducing the processing complexity of the ACE sound coding strategy while performing noise reduction for CIs. We found that the proposed method and a front-end speech enhancement method based on TasNet do not affect speech understanding in quiet when compared to ACE. Moreover, we found that these two methods are capable of improving substantially speech intelligibility in noisy conditions when compared to ACE and common front-end Wiener filtering. The proposed method has the potential to completely substitute any CI sound coding strategy while keeping its general usage for every listener and to perform speech enhancement in noisy conditions.

6. REFERENCES

[1] 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.

[2] C. P. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[3] 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. 728–39, 2016.

[4] D. Wang and J. H. L. Hansen, "Speech enhancement for cochlear implant recipients," *The Journal of the Acoustical Society of America*, vol. 143, no. 4, pp. 2244–2254, 2018.

[5] A. Buechner, K. Dyballa, P. Hehrmann, S. Fredelake, and T. Lenarz, "Advanced beamformers for cochlear implant users: Acute measurement of speech perception in challenging listening conditions," *PLOS ONE*, vol. 9, no. 4, pp. 1–9, 04 2014.

[6] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *Latent Variable Analysis and Signal Separation*, Cham, 2015, pp. 91–99, Springer International Publishing.

[7] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

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

[9] W. Nogueira, T. Gajecki, B. Krueger, J. Janer, and A. Buechner, "Development of a sound coding strategy based on a deep recurrent neural network for monaural source separation in cochlear implants," in *Speech Communication; 12. ITG Symposium*, 2016, pp. 1–5.

[10] Y. Lai, Y. Tsao, X. Lu, F. Long Chen, Y. Su, K. Chen, Y. Chen, L. Chen, L. Po-Hung Li, and C. Lee, "Deep learning–based noise reduction approach to improve speech intelligibility for cochlear implant recipients," *Ear and Hearing*, vol. 39, pp. 795–809, 2018.

[11] 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)*, 2016, pp. 6520–6524.

[12] N. Mamun, S. Khorram, and J. H. L. Hansen, "Convolutional neural network-based speech enhancement for cochlear implant recipient," in *INTERSPEECH 2019*, Brno, Czech Republic, 2019.

[13] 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, pp. 1256–1266, 2019.

[14] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," 2018.

[15] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5069–5073.

[16] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *INTERSPEECH*, 2017.

[17] R. Hinrichs, T. Gajecki, J. Ostermann, and W. Nogueira, "A subjective and objective evaluation of a codec for the electrical stimulation patterns of cochlear implants," *The Journal of the Acoustical Society of America*, vol. 149, no. 2, pp. 1324–1337, 2021.

[18] 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. 101672, Nov 2005.

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

[20] 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, 12 2010.

[21] J. R. Hershey, Z. Chen, J. Le 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.

[22] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, G. Naylor J. F. Culling, E. Porter, and R. Viveros Muñoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in proceedings of the annual conference of the international speech communication association," in *INTERSPEECH 2021*, Brno, Czech Republic, 2021.

[23] I. J. Hochmair-Desoyer, E. Schulz, L. E. 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 6 Suppl, pp. S83, 1997.

[24] A. D. Wouter, H. Verschuure, C. Ludvigsen, and S. Westermann, "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.

[25] H. Fastl, "A background noise for speech audiometry," *Audiol. Acoust.*, , no. 26, pp. 2–13, 1987.

[26] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 298605, Jul 2009.

[27] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

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

[29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *in Proc. ICASSP, 2010*, pp. 4214–4217.