INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Joint compensation of multi-talker noise and reverberation for speech enhancement with cochlear implants using one or more microphones

Clément Gaultier, Tobias Goehring

Cambridge Hearing Group, MRC Cognition and Brain Sciences Unit, University of Cambridge, UK

clement.gaultier@mrc-cbu.cam.ac.uk, tobias.goehring@mrc-cbu.cam.ac.uk

Abstract

Following speech in noisy and reverberant situations is difficult for cochlear implant (CI) users. This study investigates single- and multi-microphone deep neural network (DNN) speech enhancement algorithms on the joint task of denoising and dereverberation. The DNN algorithms were trained and tested on simulated sound scenes from behind-the-ear hearing devices. Performance was assessed using objective measures and a listening study for reverberant mixtures of speech in multi-talker babble noise. We compare results for signal distortion, predicted intelligibility and speech reception thresholds measured in a listening experiment with 15 typically hearing participants using cochlear implant simulations. Objective metrics indicated listening benefits for both single- and multi-microphone approaches while the listening study results confirmed significant improvements in speech intelligibility for the multi-microphone approaches, holding strong promise to benefit CI listeners.
Index Terms: speech enhancement, dereverberation, cochlear implants, multichannel, speech intelligibility

1. Introduction

Whether using single- or multiple-microphone approaches, deep learning-based speech enhancement algorithms have been actively investigated in recent years. On the one hand, end-to-end learning and advances in time-frequency masking by deep-neural network architectures allowed for improved performance of single-microphone time-domain approaches on speech enhancement or separation tasks [1]. On the other hand, so called neural beamformers have been developed to tackle the problem of multi-microphone speech enhancement or separation by capitalising on spatial information. Many recent studies featuring such algorithms were evaluated on increasingly realistic sound scenes including reverberation. The algorithms were trained to predict the echoic speech targets, which can yield improvements in speech intelligibility due to noise removal only. However, interactions between noise and reverberation are likely to negatively affect speech perception, especially for people with hearing difficulties and those listening with cochlear implants [2, 3]. This limitation is calling for more advanced approaches able to not only account for noise but also for reverberation artefacts. In this study we investigated the performance of different algorithms based on the Dual-Path Recurrent Neural Network (DP-RNN) [4] architecture on the joint task of denoising and dereverberation (*i.e.* predicting the anechoic speech target) and evaluated their application to cochlear implants via objective metrics and a listening study using cochlear implant simulations.

C. Gaultier was supported by funding from the Fondation Pour l'Audition [grant number: FPA RD-2021-1]. T. Goehring was supported by Career Development Award MR/T03095X/1 from the Medical Research Council UK.

Cochlear implants (CI) are sensory prostheses that restore a sense of hearing to people with severe to profound sensorineural hearing loss [5]. CIs use a surgically-implanted electrode array in the cochlea to bypass the acoustic path in the auditory system with direct electrical stimulation of the spiral ganglion cells. While most CI recipients achieve good speech understanding in quiet conditions, background noise and reverberation negatively affect their speech perception. Previous work reported significant improvements in speech intelligibility for CI recipients in noisy situations with deep neural network algorithms [6, 7, 8] and for typically hearing listeners using CI simulations [9, 10]. However, these previous studies only used single-microphone approaches and considered background noise without reverberation. Traditional multi-microphone approaches, or beamformers, provided improvements in conditions where speech and background noise were spatially separated [11, 12], but so far no DNN-based approaches have been developed nor tested for CIs. Other studies solely focused on the detrimental effects of reverberation on CI speech perception and its mitigation [13] while ignoring the interactions with background noise. It is important to establish whether more realistic situations containing both background noise and reverberation can be mitigated by advanced speech enhancement algorithms with one or more microphones for cochlear implant listeners.

This paper is organised into the following sections: Section 2 describes the different speech enhancement approaches investigated in this study as well as the experimental evaluation. A systematic comparison between the single- and multi-microphone algorithms is performed and results are detailed in section 3 for signal-based objective metrics and speech intelligibility scores measured in a listening study with typically hearing volunteers using cochlear implant simulations. Finally, we discuss the results and limitations in section 4.

2. Single- and multi-microphone speech enhancement

In this section we briefly recap the signal model and the two speech enhancement approaches to be compared before describing the parameters used in the experimental validation.

2.1. Noisy reverberant signal model

We consider the following discrete signal model:

$$\mathbf{y}^i = \mathbf{x}^i \star \mathbf{h}^i + \mathbf{n}^i \tag{1}$$

with $\mathbf{y}^i \in \mathbb{R}^{\mathsf{L}}$ a reverberant noisy mixture sensed from the i^{th} microphone of a sound recording system. $\mathbf{x}^i \in \mathbb{R}^{\mathsf{M}}$ denotes the direct path signal of a speech source of interest at the i^{th} microphone. In our model we consider the case of a body-worn recording system (hearing devices for instance) where the

10.21437/Interspeech.2023-975

direct-path signal $\mathbf{x}$ (anechoic target) encompasses head-torso-pinna reflections. $\mathbf{h}^i \in \mathbb{R}^{\mathsf{N}}$ represents the acoustic path also called Room Impulse Response in case of indoor sound propagation between the speech source and the i^{th} microphone. The latter encompasses the effect of the room (reverberation) on the propagation. Finally, $\mathbf{n}^i \in \mathbb{R}^{\mathsf{L}}$ is the noise recorded on the i^{th} microphone. $\star$ denotes the convolution operation.

2.2. Model architectures

A single-microphone algorithm based on the Dual-Path Recurrent Neural Network (DP-RNN) [4] was compared against an algorithm using multiple microphone inputs based on the DP-RNN and Filter-and-Sum (FaSNet) architectures [14]. These methods are described in the following.

2.2.1. Single-microphone setup: Encoder-Masker-Decoder

Recent advances in single-microphone speech enhancement and speech separation research promoted end-to-end deep-learning approaches over more traditional time-frequency methods utilizing filterbanks based on the short-time frequency transform (STFT). Results suggested that substantial improvements in speech separation could be achieved by replacing STFTs with convolutional encoder-decoder layers [1]. Retaining an *Encoder-Masker-Decoder* architecture allows to perform end-to-end speech separation or enhancement directly on the time-domain speech signal. This is the case for the DP-RNN Time domain Audio Separation framework (DP-RNN TasNet)[4] used in this work. The DP-RNN TasNet is formed of three main blocks: a 1-D convolutional encoder that forms a 2-D representation of the single-microphone input mixture, a masking network (DP-RNN) formed of two RNNs to process local and global information and a 1-D transposed convolutional decoder to transform the masked speech representation back to the time domain. The masking network estimates the corresponding mask values to be applied to the encoder output for each time frame of the input speech mixture to obtain the speech source signal. The DP-RNN is trained end-to-end, such that all three blocks are optimised jointly to estimate the anechoic target speech contrary to previous work that estimated the reverberant target speech. We used the model parameters as described in [4] but switched to unidirectional mode for the RNN that processes global information as well as channel-wise normalization to more closely align with causal processing required for real-time processing in hearing devices such as cochlear implants. This normalisation method only requires inputs at the current time-step and no look-ahead for future information.

2.2.2. Multi-microphone setup: Neural Beamformer

Deep-learning based spatial filtering approaches raised interest in recent years and were successfully applied to automatic speech recognition [15]. Neural beamformers make use of multi-microphone inputs to derive a mapping between input mixtures and a source signal of interest. We used the method presented in [16] based on the Filter-and-Sum Network (FaSNet) [14]. This method also relies on DP-RNN blocks but additionally learns time-domain beamforming filters using information across multiple microphones. Firstly, a latent representation of the input mixtures is formed. This representation is established by concatenating a learned encoder output as in DP-RNN TasNet with normalized cross-correlation (NCC) features between microphones. The next stage of the FaSNet consists of DP-RNN blocks which map the latent representation to beamforming filters (instead of mask values) that are then applied to the input mixture to obtain the corresponding speech source signal. The model was parameterised as in [16] but adjusted to unidirectional processing for the global RNN and channel-wise normalization. As the original FaSNet algorithm works in a frame-based fashion and requires a symmetrical contextual window with past and future samples centered around the current frame, we further decreased the frame size and contextual window to 1ms and 6ms, respectively, in order to reduce the latency required to estimate the beamforming filter. Therefore, the adjusted FaSNet only requires 4ms of future information for the processing in line with requirements for hearing devices. Similar to the single-microphone DP-RNN, the FaSNet was trained end-to-end to estimate the anechoic target speech.

2.3. Experimental setup

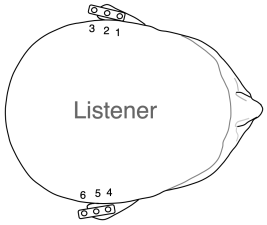

Figure 1: *Schematic showing the location of the behind-the-ear multi-microphone hearing devices with microphones 1-6.*

2.3.1. Sound scenes and datasets

Training data consisted of 6000 simulated sound scenes (4s each) from behind-the-ear hearing devices. A diagram representing a listener equipped with such a system is represented in Figure 1 with microphones, shown as small circles and numbered from 1 to 6. We used speech and impulse response recordings from the first Clarity Enhancement Challenge (CEC1) training dataset [17], [18] as well as noise recordings from the WSJ0 Hipster Ambient Mixtures (WHAM!) training dataset [19]. Reverberated noisy speech mixtures were generated at signal-to-noise ratios (SNRs) sampled uniformly from -20dB to $+20$dB. We convolved the noise recordings with the averaged diffuse parts of three randomly chosen room impulse responses similarly to [20] to model a diffuse noise. Evaluation data consisted of 270 simulated sound scenes using unseen test data that was different from the training data in the relevant aspects to robustly assess model performance. We used Bamford-Kowal-Bench (BKB) sentences [21] (English, spoken by a male talker as in [7]) as target speech, room impulse responses from the CEC1 test dataset and random excerpts from a multi-talker babble recording (16-talkers babble, Auditec®, St Louis, MO) to generate the reverberated noisy speech mixtures at SNRs from -20 dB to +20 dB.

2.3.2. Model training and configuration

Contrary to previous speech enhancement or speech separation studies, which used the reverberated clean speech as target, here the target signal for the algorithm training was the anechoic clean speech. The algorithms were trained to predict the direct path target speech $\mathbf{x}^1$ from the front left microphone of the sensing system (microphone 1 in Figure 1). This represents a joint task for the algorithm to compensate for the effects of both noise and reverberation at the same time. We trained the models for 100 epochs using the Adam optimizer, with a learning rate of 10^{-3} and a batch size of 2. All algorithms were trained to maximize the scale-invariant Signal-to-Distortion Ratio (SI-SDR) [22] using gradient clipping with a maximum l_2 norm of 5. We selected the models with the best validation loss for the

Table 1: *Quantitative results for signal-based measures in 3 SNR conditions.*

	Unprocessed			Single microphone			2 microphones			6 microphones			Ideal Ratio Mask		
SNR (dB)	-4	0	10	-4	0	10	-4	0	10	-4	0	10	-4	0	10
SI-SDR (dB)	−5.29	−2.63	4.62	−7.28	0.49	8.28	6.18	8.00	10.06	9.13	10.80	12.69	10.65	11.90	13.58
eSTOI	0.45	0.55	0.77	0.44	0.58	0.82	0.66	0.76	0.88	0.77	0.84	0.92	0.88	0.91	0.94
NCM	0.29	0.49	0.80	0.21	0.51	0.84	0.76	0.83	0.91	0.85	0.91	0.96	0.94	0.97	0.98

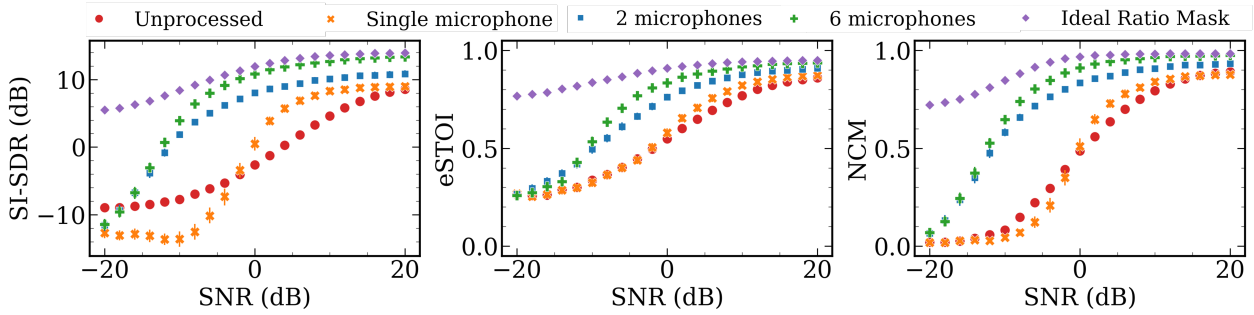

Figure 2: *Comparison of signal-based metrics across SNR for all conditions.*

testing.[1] We trained three different models with access to one or more microphones. The first model was adapted from [4] with access to a single microphone (front-left, microphone 1 in Figure 1). The second model was adapted from [16] with access to 2 unilateral microphones (front-left and rear-left, microphones 1 and 3 in Figure 1). The third model was trained with access to all 6 bilateral microphones (1 to 6 on Figure 1). The first two models thus relied on unilateral information without or with spatial information while the third model used bilateral and spatial information. We implemented the models by adapting the baseline implementations of [4] and [16] from the Asteroid toolbox [23].

2.4. Cochlear implant simulation

Noise or tone vocoders are commonly used to simulate speech as obtained after CI processing [9, 7, 10]. In order to simulate speech perception with CIs, all tested signals used for comparison in section 3 were processed with the SPIRAL vocoder [24] which was developed and validated to encompass the effects of limited spectro-temporal resolution and increased spread of excitation with CIs [25]. SPIRAL was parameterized with 16 analysis filter bands to represent the 16 electrode channels as used in CIs manufactured by Advanced Bionics® (AB, Valencia, CA) and a current decay slope of -16 dB/octave to simulate the effects of spread of excitation observed typically with CIs [26]. In previous research, the SPIRAL vocoder and similar approaches have successfully been used to simulate CI speech perception in typically hearing listeners and produced speech performance well aligned with findings for CI listeners [7].

3. Performance comparison

We assessed performance of the three models on objective signal-based metrics and measured speech intelligibility scores in a listening study. We compared the three different speech enhancement approaches, as described in subsection 2.3.2 (Single microphone, 2 microphones and 6 microphones), against two control conditions: the noisy reverberant input speech mixture (Unprocessed condition) and the speech obtained by applying an STFT-based Ideal-Ratio-Mask (IRM) [27] on the front-left microphone signal ($\mathbf{y}^1$) with the anechoic target STFT magnitudes as reference.

3.1. Signal-based metrics results

Performance was assessed with two speech intelligibility prediction metrics: the extended Short Time Objective Intelligibility measure (eSTOI) [28] and the normalized covariance metric (NCM) [29]. eSTOI is a popular metric in the speech enhancement community and NCM was successfully used on vocoded speech signals in previous CI studies [30, 7]. For comparison purposes to other speech enhancement studies, we also report SI-SDR results. All measures were computed for 105 sentences from the evaluation dataset using the vocoded anechoic clean speech as the reference (Table 1). Figure 2 left, middle and right panels show average SI-SDR, eSTOI and NCM scores for the whole range of SNRs (−20 dB to +20 dB). Error bars represent 95% confidence intervals.

3.2. Listening study

To assess the potential of the speech enhancement models in restoring speech perception for people with cochlear implants, we measured speech intelligibility for typically-hearing volunteers listening to CI-simulated stimuli. Details of this listening study are given below.

3.2.1. Experimental procedure

Fifteen native English-speaking, volunteers with typical hearing (average age 27.4 years) participated in this study, as part of a larger research program that was approved by the National Research Ethics committee for the East of England. Before starting the test, the participants listened to 2 lists of 15 sentences from the BKB dataset to acclimatize to vocoded speech. After the acclimatization, Speech Reception Thresholds (SRTs) were measured, as the SNR at which 50% of the speech was understood correctly, for each of the 5 conditions listed in section 3. We used one list of 15 sentences per condition with a standard one-up / one down adaptive staircase procedure in which the participants had to repeat the speech stimuli they heard to be scored by the experimenter. The procedure started at 0 dB SNR and used a step size of 2 dB as described

[1]This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service, using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council.

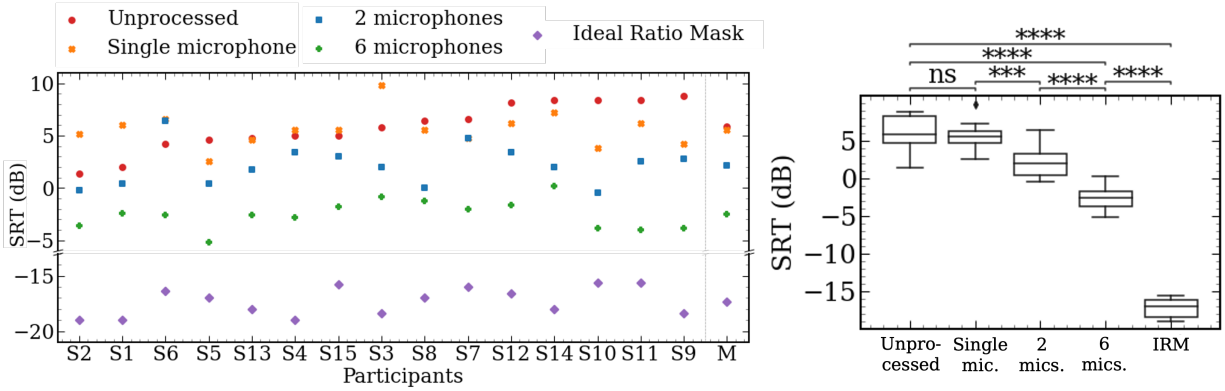

Figure 3: *Individual (left) and mean (right) speech reception thresholds for the 5 conditions compared in the listening study.*

in [31]. The condition order was randomized for each participant and the procedure was repeated in reverse order to compensate for any remaining learning effect throughout the session. We used a double-blinded design, in which both the participant and the experimenter were unaware as to which condition was tested. Stimuli were presented to the participants diotically through Sennheiser® HD600 headphones using an RME® Fireface UCX soundcard in a sound-proof testing booth, while the experimenter used a monitoring microphone to rate the responses. A full testing session lasted about 1 hour, including the acclimatization phase and short breaks.

3.2.2. Speech intelligibility results

All participants were able to perform the task and completed the adaptive procedure for all conditions. Individual SRT results are shown in the left panel of Figure 3 for each participant and the average results (denoted M on the x-axis) are shown in the right panel. A repeated-measures analysis of variance revealed a significant main effect of condition [$F(4, 14) = 547.9$; $p < 0.001$] with no violation of sphericity. Bonferroni-corrected pairwise comparisons showed highly significant mean differences between all conditions except for the difference between the "Unprocessed" and "Single microphone" conditions (see right panel of Figure 3). Mean SRTs were [6 dB, 5.6 dB, 2.2 dB, -2.5 dB, -17.3 dB] for the five conditions [Unprocessed, Single microphone, 2 microphones, 6 microphones, IRM].

4. Discussion

Three algorithms based on the DP-RNN [4] and FaSNet [16] were developed to jointly alleviate noise and reverberation effects on speech perception with CIs. We aimed to test these promising end-to-end algorithms to assess their performance in situations with noise and reverberation. Such challenging, but realistic, situations may be prone to time-domain distortions and difficult to master for any algorithm. The algorithms were implemented to fulfill latency requirements for real-time processing in hearing devices and had access to different numbers of microphones placed either uni- or bilaterally in a behind-the-ear hearing device. We discuss performance for the signal-based prediction measures and the speech reception thresholds from a listening experiment under CI-simulated conditions.

The three algorithms improved objective measure scores over the unprocessed condition across a range of SNRs. We note from Figure 2 and Table 1, that the single-microphone approach improved predicted intelligibility scores (eSTOI, NCM) and signal distortion (SI-SDR) only for positive SNRs. In contrast, the algorithms using multiple microphones showed predicted improvements even for negative SNRs down to -16 dB. Objective metrics indicated better signal to distortion ratios and predicted intelligibility for the multi-microphone approaches over the single-microphone one, especially in the noisiest cases.

For the fifteen participants that took part in this study, statistical analysis showed significant improvements of up to 7 dB in SRTs between the multi-microphone approach over both the unprocessed condition and the single-microphone approach. We also note on Figure 3 that some participants with low SRT scores ($<$ 6 dB) in the unprocessed condition obtained higher SRTs for the single-microphone algorithm condition, meaning speech intelligibility decreased after processing. On the other hand, participants with higher SRTs for the unprocessed condition ($>$ 8 dB) seemed to benefit from the processing with the single-microphone algorithm. These potential differences across participants could arise from some participants being able to tolerate more distortions or artefacts introduced by the speech enhancement methods than others. This could also be related to variable performance of the single-microphone algorithm across SNRs as indicated by the objective measures.

While the purpose of this study was not to validate the prediction accuracy of intelligibility metrics, it is of interest whether objective measures and results from the listening study agreed. We note that predicted intelligibility (eSTOI and NCM) presented in Figure 2 correctly reflected the ranking between conditions obtained with the SRTs from the listening study. However, the objective measures eSTOI and NCM produced quite different results quantitatively compared to the findings from the listening study. This emphasises the need for more accurate, possibly CI-specific, intelligibility prediction measures for noisy and reverberant situations.

4.1. Conclusion and future possibilities

Two DNN algorithms were employed to perform joint denoising and dereverberation. We find a clear superiority of multi-microphone approaches in a listening study using cochlear implant simulations. Interestingly, even with only 2 microphones placed unilaterally, there was a significant improvement by 3.8 dB in SRT, which is a promising finding as many CI listeners only have a single device available to them. Bilateral processing may yield larger benefits still (7 dB) but also increases latency and power consumption due to streaming across the ears. Despite predicted benefits by the objective measures, the single-microphone algorithm did not improve SRTs, which may be related to the joint task employed here. This motivates further investigation in conjunction with bilateral sound processing with DNNs. These findings need to be confirmed with CI listeners and future work should also investigate interaction effects on auditory awareness of the acoustic environment.

5. References

[1] 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.

[2] J. Badajoz-Davila, J. M. Buchholz, and R. Van-Hoesel, "Effect of noise and reverberation on speech intelligibility for cochlear implant recipients in realistic sound environments," *The Journal of the Acoustical Society of America*, vol. 147, no. 5, pp. 3538–3549, 2020.

[3] O. Hazrati and P. C. Loizou, "The combined effects of reverberation and noise on speech intelligibility by cochlear implant listeners," *International journal of audiology*, vol. 51, no. 6, pp. 437–443, 2012.

[4] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 46–50.

[5] R. P. Carlyon and T. Goehring, "Cochlear implant research and development in the twenty-first century: a critical update," *Journal of the Association for Research in Otolaryngology*, vol. 22, no. 5, pp. 481–508, 2021.

[6] T. Goehring, F. Bolner, J. J. Monaghan, B. Van Dijk, A. Zarowski, and S. Bleeck, "Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users," *Hearing research*, vol. 344, pp. 183–194, 2017.

[7] T. Goehring, M. Keshavarzi, R. P. Carlyon, and B. C. Moore, "Using recurrent neural networks to improve the perception of speech in non-stationary noise by people with cochlear implants," *The Journal of the Acoustical Society of America*, vol. 146, no. 1, pp. 705–718, 2019.

[8] 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.

[9] 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.

[10] 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.

[11] A. A. Hersbach, D. B. Grayden, J. B. Fallon, and H. J. McDermott, "A beamformer post-filter for cochlear implant noise reduction," *The Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. 2412–2420, 2013.

[12] R. M. Baumgärtel, H. Hu, M. Krawczyk-Becker, D. Marquardt, T. Herzke, G. Coleman, K. Adiloğlu, K. Bomke, K. Plotz, T. Gerkmann *et al.*, "Comparing binaural pre-processing strategies ii: Speech intelligibility of bilateral cochlear implant users," *Trends in hearing*, vol. 19, p. 2331216515617917, 2015.

[13] K. Kokkinakis, O. Hazrati, and P. C. Loizou, "A channel-selection criterion for suppressing reverberation in cochlear implants," *The Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3221–3232, 2011.

[14] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing," in *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE, 2019, pp. 260–267.

[15] X. Xiao, C. Xu, Z. Zhang, S. Zhao, S. Sun, S. Watanabe, L. Wang, L. Xie, D. L. Jones, E. S. Chng *et al.*, "A study of learning based beamforming methods for speech recognition," in *CHiME 2016 workshop*, 2016, pp. 26–31.

[16] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6394–6398.

[17] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, R. Viveros Munoz *et al.*, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *INTERSPEECH*, vol. 2. International Speech Communication Association (ISCA), 2021, pp. 686–690.

[18] S. Graetzer, M. A. Akeroyd, J. Barker, T. J. Cox, J. F. Culling, G. Naylor, E. Porter, and R. Viveros-Muñoz, "Dataset of british english speech recordings for psychoacoustics and speech processing research: The clarity speech corpus," *Data in Brief*, vol. 41, p. 107951, 2022.

[19] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "Wham!: Extending speech separation to noisy environments," in *Proc. Interspeech*, Sep. 2019.

[20] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 36–40.

[21] J. Bench, Å. Kowal, and J. Bamford, "The bkb (bamford-kowal-bench) sentence lists for partially-hearing children," *British journal of audiology*, vol. 13, no. 3, pp. 108–112, 1979.

[22] J. Le 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)*. IEEE, 2019, pp. 626–630.

[23] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," in *Proc. Interspeech*, 2020.

[24] J. A. Grange, J. F. Culling, N. S. Harris, and S. Bergfeld, "Cochlear implant simulator with independent representation of the full spiral ganglion," *The Journal of the Acoustical Society of America*, vol. 142, no. 5, pp. EL484–EL489, 2017.

[25] T. Goehring, A. W. Archer-Boyd, J. G. Arenberg, and R. P. Carlyon, "The effect of increased channel interaction on speech perception with cochlear implants," *Scientific Reports*, vol. 11, no. 1, pp. 1–9, 2021.

[26] A. J. Oxenham and H. A. Kreft, "Speech perception in tones and noise via cochlear implants reveals influence of spectral resolution on temporal processing," *Trends in Hearing*, vol. 18, p. 2331216514553783, 2014.

[27] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[29] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–3405, 2009.

[30] F. Chen and P. C. Loizou, "Predicting the intelligibility of vocoded and wideband mandarin chinese," *The Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3281–3290, 2011.

[31] A. MacLeod and Q. Summerfield, "A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use," *British journal of audiology*, vol. 24, no. 1, pp. 29–43, 1990.