Overcoming Limitations: Deep Neural Network Models Show Promise in Reducing Multi-Talker Noise Interference for Cochlear Implant Listeners

Short Title: Cochlear implants with deep neural network

*Agudemu Borjigin[a,b], Kostas Kokkinakis[c], Hari M. Bharadwaj[d], and Joshua S. Stohl[e]

[a] *Weldon School of Biomedical Engineering, Purdue University, West Lafayette 47907, IN, USA*

[b] *Waisman Center, University of Wisconsin Madison, Madison, WI 53705, USA*

[c] *Concha Labs, San Francisco, CA 94114, USA*

[d] *Department of Communication Science and Disorders, University of Pittsburgh, Pittsburgh, PA 15213, USA*

[e] *North American Research Laboratory, MED-EL Corporation, Durham, NC 27713, USA*

June 12, 2023

*Correspondence: dagu@wisc.edu

Abstract

Although cochlear implants (CIs) perform well in restoring speech intelligibility in quiet, they fail to provide the same level of benefit in noisy settings. Current noise-reduction solutions in hearing aids and CIs only remove predictable, stationary noise, and are ineffective against realistic, non-stationary noise such as multi-talker background. Recent developments in machine learning have resulted in deep neural network (DNN) models that achieve noteworthy performance in speech enhancement and separation tasks. However, little work has investigated the potential of DNN models in addressing the limitations of current signal processing algorithms in removing non-stationary noise interference. The research in this regard is even more scarce for CIs. The small body of work that did investigate the application of DNN models in CIs has been limited by the very small dataset used for model training and testing, simple model architectures that are not best suited for speech processing, and the lack of testing with CI listeners. In this work, we implemented two DNN models that are well suited for applications in speech audio processing: (1) a recurrent neural network (RNN), a lightweight template model, and (2) SepFormer, which is a transformer-based, state-of-the-art model for speech separation. The models were trained with a custom-created training dataset ($\sim$30 hours) that contained noisy speech in both non-speech environmental sounds, and more complex speech babble background. The models were then tested with thirteen CI listeners. Both RNN and SepFormer models significantly improved CI listener's speech intelligibility in noise without compromising the perceived quality of the speech. These models not only increased the intelligibility in stationary non-speech noise to the level that is comparable to those demonstrated by traditional signal processing strategies, but also, most importantly, introduced substantially ***more*** improvements in non-stationary ***speech-babble*** noise, where conventional signal processing strategies fall short with very little benefits. These results show the promise of using machine-learning based models as a complementary or even replacement algorithm for current existing signal processing strategies, to tackle more complex listening challenges involving speech-noise interference.

Key words: deep learning models, recurrent neural network, transformer, speech enhancement, speech intelligibility in non-stationary noise

1 Introduction

Cochlear implant (CI) users typically achieve satisfactory speech intelligibility in quiet environments such as audiology clinics, but struggle to converse in noisy social settings. There have been significant efforts towards the development of noise reduction algorithms in the field of hearing aids, and these algorithms are increasingly being translated into both CI research and commercial CI products. Noise reduction algorithms can be classified into single- vs multi-microphone strategies, as reviewed by Kokkinakis et al. (2012) and Henry et al. (2023). A single-microphone noise reduction strategy involves processing the audio signal from a single microphone to reduce background noise. This can be achieved by using many approaches such as spectral subtraction (Yang & Fu, 2005; Verschuur et al., 2013), subspace estimate (Loizou et al., 2005), gain adjustment based on short-time signal to noise ratio (SNR) estimate (Mauger, Arora, & Dawson, 2012; Mauger, Dawson, & Hersbach, 2012; Cohen, 2003; Dawson et al., 2011), and ideal binary mask estimate (Hu & Loizou, 2008b; Koning et al., 2015). These traditional single-microphone algorithms are heavily driven by signal-processing strategies based on time-averaged signal statistics and certain assumptions about the noise. For example, in spectral subtraction, the noise is assumed to be non-modulated and stationary (as well as broadband and additive) whereas the modulation is used for determining the presence of speech target signal. Because of these inherent assumptions, these algorithms tend to perform relatively well only if the background is statistically consistent/predictable, such as a stationary noise. Yang & Fu (2005) observed improvement with speech-shaped noise but only moderate improvement with 6-talker babble; CI listeners in Mauger, Arora, & Dawson (2012)'s study showed improvements of 19% on average in speech recognition in speech-weighted noise and small improvement of 7% in 20-talker babble, but no significant improvement in a 4-taller babble; Dawson et al. (2011) also observed greatest benefit in speech-weighted noise. A number of studies evaluated a commercial noise-reduction strategy—Advanced Bionic's ClearVoice (AdvancedBionics, 2012), which works by adjusting the short-term gains according to the estimate of instantaneous SNR. Dingemanse & Goedegebure (2015) did not see improvements in speech recognition with ClearVoice, even in non-speech noise. ClearVoice was only shown to increase the performance primarily in non-speech noise when combined with some parameter adjustments such as adjusting maximum comfort levels (Dingemanse & Goedegebure, 2018), or combined with other technologies such as remote microphone (Wolfe et al., 2015) and multi-microphone noise-reduction strategies (Geißler et al., 2015; Hersbach et al., 2012). Similar

commercial noise reduction strategies for CIs include Automatic Sound Management 3.0 from MED-EL (2021) and SmartSound from Cochlear (Mauger et al., 2014). Listening with CIs and other hearing assistive devices is especially challenging in non-stationary noise with modulations, such as multi-talker background interference (Cullington & Zeng, 2008; Fu et al., 1998). Therefore, the inherent inability of single-channel noise reduction solutions to effectively remove non-stationary noise interference remains to be the most critical limitation.

Compared to single-microphone algorithms, multi-microphone noise reduction strategies, also commonly referred to as beamforming, involve using multiple microphones to capture the audio signals from different directions and then processing the signals to reduce background noise. Beamforming was first introduced to commercial CI products by Cochlear Corporation in 2005, over a decade after it became a standard feature in hearing aids (Buechner et al., 2014). Beamforming consists of a front and rear microphone. A sound originating from behind arrives at the rear microphone before the front microphone. This external time delay is used by the signal processing to attenuate the sound from the back. Adaptive beamforming was later introduced to account for moving noise source that is not always at the back of the listener but could be slightly off to the side from behind (Bentler et al., 2006). Directionality is typically better with more microphones (Dillon, 2008). However, due to the minimum distance requirement between the microphones and the size of a CI processor, it is impractical to have more than two physical microphones on a single processor. Binaural beamformers were then developed to increase the number of microphones by utilizing all four physical microphones from two processors for bilateral CI listeners. Picou et al. (2014) showed that binaural directionality provided significantly better benefit in speech-in-noise hearing than monaural directional microphones regardless of the SNR level. Although multi-microphone noise reduction approaches have generally shown significant improvement in speech intelligibility in noise among CI listeners, they require additional hardware and can be more complex to implement. The effectiveness of multi-microphone noise reduction strategies is limited to the scenario where the target speech and noise are spatially separated and the listener faces the target all the time. These algorithms can sometimes be adversely affected by reverberation, which exists in most listening situations, and can degrade source localization (Dawson et al., 2011; Henry et al., 2023). The effectiveness of a beamformer is also affected by how the device is worn by the listener. For a behind-the ear processor, the optimal sitting position of the device on the ear results in the front and rear microphones positioned close to horizontally, which allows for a match between the real delay and the assumed external time delay

based on the physical distance between two microphone, enabling the optimal cancellation of noise. If the device is worn with a tilt angle further behind the ear, for example, the real time delay will be smaller than the external delay assumed by the strategy (Ricketts, 2001). Therefore, the ultimate solution to listening challenges in a wider range of realistic situations is a single-microphone noise reduction algorithm that not only suppresses the stationary noise, but also has the ability to effectively remove more common yet complex, non-stationary background noise such as a multi-talker babble.

Machine learning, especially deep neural network models, has transformed many fields, including audio processing. Although machine learning has yet to make an impact in hearing research and clinics (Lesica et al., 2021), it appears to have the potential to overcome the limitation of current noise reduction solutions in CI and hearing aids in reducing non-stationary noise (Henry et al., 2023). Research in machine learning-based noise reduction algorithms have only begun to surface in a larger scale in the last decade with the advent of more powerful computing resources and the availability of larger datasets. Early work took a small step forward by developing machine learning-based models for suppressing relatively non-stationary, but non-speech environmental noise and demonstrated improvement in speech intelligibility among CI listeners (Hu & Loizou, 2010) and objective quality measures (Gopalakrishna et al., 2012). More recent studies have shown improved speech intelligibility in more non-stationary background noise such as a multi-talker babble. For example, Lai et al. (2017) showed improved speech intelligibility in 2-talker babbles among normal hearing (NH) listeners with CI simulations with a deep denoising autoencoder (DDAE), and further improved the performance among Mandarin-speaking CI listeners when a noise classifier was added to the DDAE model (Lai et al., 2018). Goehring et al. (2017) demonstrated improved speech intelligibility in a 20-talker babble among CI listeners when a neural network model was used to process the noisy speech. However, these models mentioned above operate based on *a priori* knowledge of the target and/or background by using the same materials from training during testing. For example, Hu & Loizou (2010), Goehring et al. (2017) and Lai et al. (2018) simply split the same speech corpus into training and testing dataset and used essentially the same noise types to create noisy speech. The model performance is expected to decrease significantly if truly *unseen* testing data are presented to the models. The larger and more diverse dataset is key to improving model generalizability. Another important aspect of generalizability is model architecture. Recent studies have demonstrated the potential of using recurrent neural network (RNN) models for better generalization by including a classic architecture called long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997). LSTM accumulates information

from the past and hence enables the network to form a temporary memory, which is essential for properly managing and learning speech context. Many studies have demonstrated the success of RNN-LSTM based models in speech recognition, enhancement, and separation applications (Graves et al., 2013; Weninger et al., 2015; Chen & Wang, 2017; Kolbæk et al., 2017). However, the application of such models in noise reduction for hearing aids and cochlear implants has been limited except for a few recent studies including Healy et al. (2019) and Goehring et al. (2019).

This work aims to further explore the potential of machine learning models in reducing non-stationary, multi-talker background noise. However, we also aim to develop models with greater generalizability by leveraging large training dataset as well as more advanced model architectures, such as RNN. Despite the wide adoption of RNN in modern audio processing systems and in many other domains, RNN architecture suffers from "vanishing gradient" or "short-term memory", which renders them ineffective for long speech instances. Therefore, in addition to RNN, we adopted an architecture known as "Transformer", which can process input signal all at once through parallel processing, which ultimately leads to a more efficient learning of long-term dependencies. Transformer has gained competitive performance and considerable popularity in speech recognition (Karita et al., 2019), speech synthesis (Li et al., 2019), speech enhancement (Kim et al., 2020), and audio source separation (Subakan et al., 2021), as well as other applications such as ChatGpt. We adopted the SepFormer from Subakan et al. (2021), a Transformer-based, top-performing model in speech separation applications at the time of this study, according to Papers with Code website. This state-of-the-art SepFormer model was used here as a reference for the flagship benchmark model to explore the maximum benefit/upper bound of DNN-based noise reduction for CIs, while the RNN model served as a relatively low-complexity, but still an advanced model to explore the real-time feasibility of a DNN-based model by accounting for the constraints on processing time and computational power in a CI audio processor. In addition to the choice of more powerful model architectures and training the models with a large custom-created dataset, the effectiveness of the models was evaluated not only with objective intelligibility metrics, but also with CI listeners. Most previous work was done without subject testing, especially with CI listeners. Most studies have also given priority to intelligibility. However, it is also important to evaluate the perceived quality. In this work, we evaluated both intelligibility and quality to also investigate the impact (or cost) of the processing algorithms on the subjective quality of the processed speech.

Table 1. Demographic information for the subjects who participated in this study. Age_T: age at the time of testing (years); Age_HL: age of onset of hearing loss (years); Dur_CI: duration of CI use (months); #_El: number of active electrodes; Score: IEEE score in quiet (percent correct). Strategy: 1 - Fine structure processing with *parallel* stimulation in the four apical channels; 2 - Fine structure processing with *sequential* stimulation in the four apical channels; 3- Fine structure processing with *sequential* stimulation in a variable number of apical channels.

Subject	**Gender**	**Age_T**	**Age_HL**	**Dur_CI**	**#_El**	**Strategy**	**Score**
ME128	M	70	30	95	10	FS4-p[1]	79
ME149	F	57	27	53	8	FS4[2]	92
ME153	M	71	34	45	12	FS4	91
ME159	F	60	17	64	12	FSP[3]	92
ME185	F	65	8	119	12	FS4-p	76
ME191	M	64	42	40	11	FS4-p	93
ME196	M	65	11	24	11	FS4	92
ME202	M	55	0	40	12	FS4	56
ME203	F	63	5	41	12	FS4-p	73
ME204	M	63	41	33	12	FS4-p	48
ME206	F	72	27	66	12	FS4-p	92
ME207	M	20	0	229	12	FS4	86
SSD100	F	37	24	158	12	FS4-p	68

2 Methods

2.1 Participants

A total of thirteen adults (7 males) fitted with MED-EL CIs (MED-EL GmbH, Innsbruck, Austria) participated in the study. They were between the ages of 20 and 72 (mean = 58.6 years, SD = 14.7 years). The average duration of CI use was 6.5 years (SD = 5 years). Demographic information is detailed in Table 1, including each subject's default clinical sound coding strategy. This study was approved by the Western Institutional Review Board (Protocol 20100066). All subjects gave informed written consent prior to testing. Only research subjects whose participation in the study would cause financial hardship received financial compensation for their participation.

2.2 Deep Neural Network Architectures

2.2.1 RNN

The schematic of the single-channel, RNN-based speech enhancement algorithm is illustrated in Figure 1a. A clean target speech and either speech babble or non-speech noise were mixed to create unprocessed

noisy speech. The features used as input to the RNN model were the spectral magnitudes of the short-time Fourier transformation (STFT) of the mixtures (size of Fast Fourier Transform = 512). The spectral magnitudes were extracted using Hamming-windowed frames with a window size of 32 samples and a hop size of 16 samples applied to signals sampled at 16 kHz. "Add-one" log (i.e., adding 1 to the value before taking log) was applied to the spectral magnitude to reduce the influence of very small values (values that are smaller than 1). The predicted mask (i.e., the model outcome) was a continuous "soft" mask instead of a binary mask (where the values of the mask are either one or zero). Previous work suggests that "soft" continuous masks result in better speech quality and intelligibility than binary masks under various noise conditions (Madhu et al., 2013). The original unprocessed noisy mixtures were multiplied with the predicted mask derived from the RNN model in order to generate the "de-noised" spectrum of the mixture. The "de-noised" estimate would ideally consist of the enhanced target speech only. This "de-noised" spectrum was compared with the spectrum of the clean target speech to compute the mean square error (MSE) loss for training optimization. The processed speech estimate was then recovered by resynthesizing (i.e., taking the inverse STFT) the "de-noised" spectrum. The RNN network consists of an input layer (size = 512) and two hidden LSTM layers (256 units), with each LSTM followed by a projection layer (128 units). A rectified linear unit (ReLU) is used as an activation function. A PyTorch-powered speech toolkit—SpeechBrain—was used to implement, train, and test the RNN model. The Adam optimizer was used for minimizing the MSE loss during the training process (Kingma & Ba, 2017), with learning rate set to 0.0001. The model performance was evaluated and monitored with a validation dataset at the end of each learning cycle (i.e., epoch, containing all training samples). The training was terminated after 100 epochs to avoid overfitting, where the model performance with validation dataset stabilized with no further significant improvements.

2.2.2 SepFormer

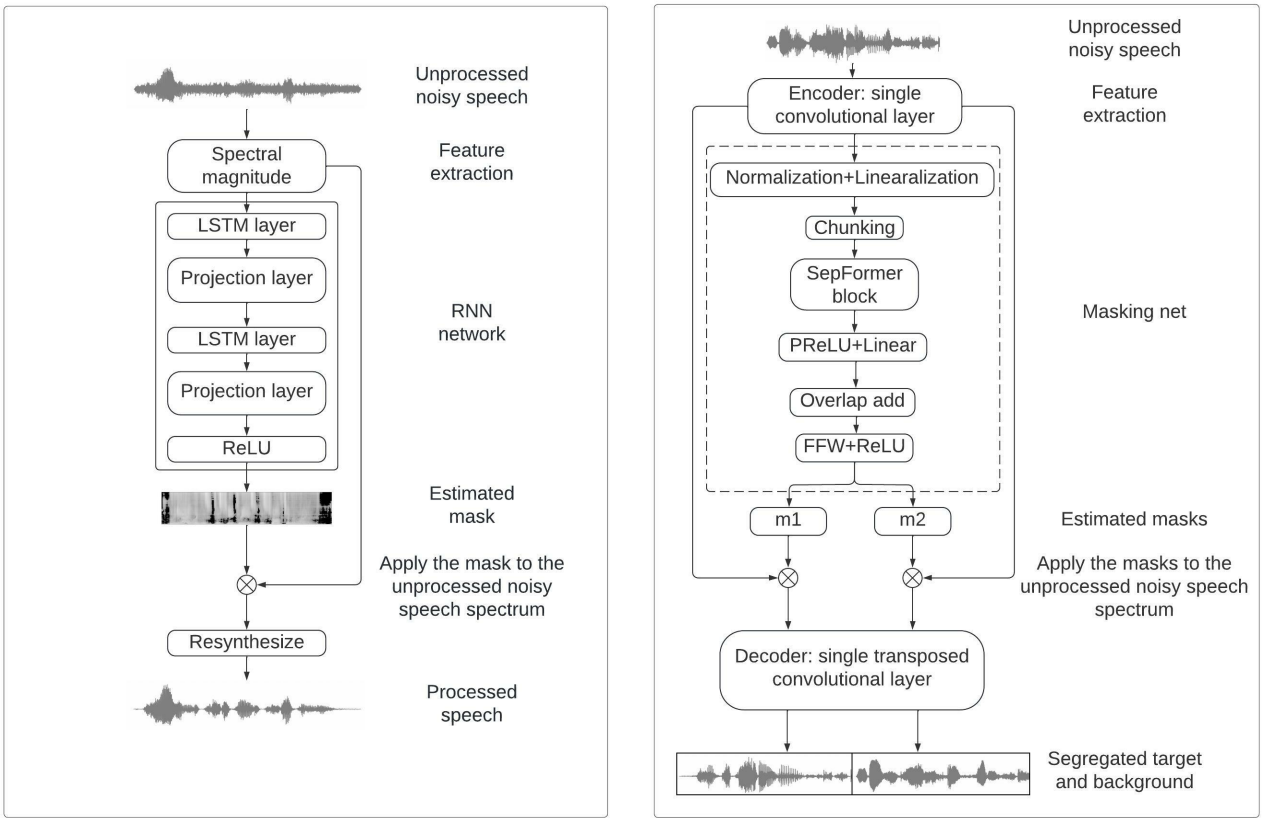
While the generic RNN model serves as a proof of concept for DNN algorithms that are suitable for small devices such as CIs, we also implemented the current state-of-the-art model for speech separation applications—SepFormer, to explore the ceiling limit of current DNN technology in noise reduction for CIs (especially non-stationary multi-talker babble noise). The model architecture is shown in Figure 1b. Instead of directly feeding STFT features of the noisy mixture to the network as in the RNN model, a single-layer convolutional network was used as an encoder to learn the 2-dimensional features of the input

noisy signal (256 convolutional filters with a kernel size of 16 samples and a stride factor of 8 samples). Similarly, at the end of the process, a transposed convolution layer with the same stride and kernel size as in the encoder was used to turn the source-separated features back into separate sources within the mixture. The extracted features of the noisy mixture go into the masking network, which estimates the masks for the foreground (i.e., target speech) and background. These masks were also continuous or soft-decision masks that provided continuous gains from 0 to 1 as in the RNN model. In the masking network, the features were first normalized and processed by a linear layer. They were then buffered into chunks of size 250 along the time axis with an overlap factor of 50%. Next, they were fed into the core of the masking net—SepFormer block. This block consists of two transformer structures that learned both short and long-term dependencies. More details about this model can be found in Subakan et al. (2021). The output of the SepFormer block was then processed by a parametric rectified linear unit (PReLU) and linear layer. The overlap-add scheme, described in Luo & Mesgarani (2019), was used to sum up the chunks. This summed representation was passed through two feed-forward layers and a ReLU activation function to finally generate the masks for both the foreground and background sources. The training procedure and infrastructure were the same as for the RNN model.

2.3 Dataset

2.3.1 Training and validation dataset

Clean target speech audios were from LibriSpeech ASR corpus, an open-source, large-scale (∼1000 hours, 2484 speakers) corpus of read English speech. The speech audios were extracted from read audiobooks from the LibriVox initiative, and every recording has been carefully segmented and aligned with one another. This study used the training subset that contains 100-hours of clean speech. The non-speech materials were from WHAM!, an open-source, large-scale (∼ 82 hours) dataset of noise audios. The noise audios were recorded from different urban spots across the San Francisco Bay Area, mainly comprising restaurants, cafes, bars, and parks. Each recording was processed to eliminate any parts that include intelligible speech. Four types of noisy speech were created for training: target in non-speech noise (WHAM! noise), and target in 1, 2, 4-talker babble. Recordings from LibriSpeech were mixed to create the noisy speech in multi-talker babble. All speakers were made sure to be distinct from one another and speaking different content. Each type of noisy speech was mixed at SNRs from 1 to 10 dB in 1-dB



(a) RNN **(b)** SepFormer

Figure 1. Schematic diagrams of the DNN architecture and signal processing frameworks used in this study: (a) RNN and (b) SepFormer

steps, with equal representation. The loudness of target was kept constant at ITU level of 29. A total of 5590 mixtures were created for each noise type, which totals a training dataset of $\sim$ 30 hours. Validation dataset contains 410 mixtures for each type of noisy speech. To speech up the process, the training and validation were conducted on graphics processing unit (GPU, Google Colaboratory).

2.3.2 Testing dataset

Clean target speech materials were extracted from IEEE corpus, which is often used in auditory research and clinics. This corpus contains recordings of 33 different talkers. Most speakers read the full set of 720 IEEE sentences. Three hundred and forty sentences from the IEEE corpus were mixed with CCITT noise (speech shaped stationary noise according to ITU-T Rec. G.227) or 2-talker babble at SNRs of +1, +5 and +10 dB. Two-talker babbles were produced by mixing the IEEE recordings from 120 sentences that were not chosen as target sentences. Male recordings were used for target while female recordings

were used to generate 2-talker babbles to create the noisy mixture. Note that IEEE corpus was never used during training or validating the models. The non-speech noise used for training and validation was relatively sparse environmental sounds, whereas the CCITT used for testing was spectral-temporally dense stationary noise. This testing dataset was used for both objective evaluations (section 2.4.1) and subject testing (section 2.4.2). The remaining 240 IEEE sentences were used for subject testing in quiet. Note that the testing was conducted on central processing unit (CPU, MacBook Pro i7 2.2GHz).

2.4 Procedure

2.4.1 Objective evaluations

The DNN models were first evaluated quantitatively using three commonly used acoustic evaluation metrics: scale-invariant source-to-distortion ratio (SI-SDR) (Roux et al., 2018), short-time objective intelligibility (STOI) (Taal et al., 2011, 2010), and "perceptual" evaluation of speech quality (PESQ) (Rix et al., 2001; Hu & Loizou, 2008a). These objective evaluation methods helped inform the overall expected benefit before conducting behavioral listening test with CI users. All three evaluation metrics compare the clean reference speech and the same speech recovered from the noisy mixture and quantify the agreement between the two, which allows for an estimate of the improvement in quality and intelligibility due to model processing. The traditional SDR metric decomposes the estimated source into four components representing respectively the true source, spatial distortions, interference, and artifacts. The final SDR score is computed by calculating the ratio of the source energy to the sum of all other projection energies (i.e., spatial distortions, interference, and artifacts) as described in Vincent et al. (2006). The SI-SDR with slight modifications as described in Roux et al. (2018) has been shown to be more robust and is now the standard for DNN noise reduction evaluation. The STOI metric was initially designed to predict the intelligibility of speech processed by enhancement algorithms. Recently, Falk et al. (2015) demonstrated that STOI outperformed all other measures for predicting intelligibility demonstrated by CI listeners. The STOI first applies time-frequency analysis to both clean reference and processed speech. An intermediate intelligibility measure is obtained by estimating the linear correlation coefficient between clean and processed time-frequency units. The final STOI score is the average of all intermediate intelligibility estimates from all time-frequency units. The PESQ score ranges between –0.5 and 4.5. It was calculated by comparing the reference signal with the processed signal by deploying a perceptual model of the

human auditory system. The PESQ is computed as a linear combination of average disturbance value and average asymmetric disturbance value. The parameters for the linear combination can be further modified towards predicting different aspects of speech quality. More details can be found in Rix et al. (2001), Hu & Loizou (2008a) and Kokkinakis & Loizou (2011). In general, the PESQ has been shown to be capable of reliably predicting the quality of processed speech. Kokkinakis & Stohl (2021) showed that parameter optimization for reverberation suppression algorithms based on the PESQ metric resulted in better performance than the STOI metric. Therefore, in the present context, the PESQ was chosen to detect and quantify the overall effects of DNN processing on the signal quality. Note that these objective evaluation metrics were developed based on data from NH listeners. However, these metrics were conducted to mainly inform rather than replace the subject testing. Even the models adapted for CI listening can never fully approximate the actual testing with CI listeners.

2.4.2 Behavioral Listening Experiments

Test setup. All participants were tested using their everyday clinical program. The stimulus was delivered to the audio processor through the direct audio input (DAI) cable, which attenuates the microphone inputs by approximately 30 dB while passing the direct input signal without attenuation. The DAI also bypasses the front-end directionality and wind-noise reduction features. The test stimuli were presented at an input level corresponding to 65 dB SPL (root mean square (RMS) level). None of the participating subjects used MED-EL's channel-specific, ambient and transient noise reduction algorithms in their daily maps at the time of testing. At the beginning of the testing session, the audiologist provided instructions regarding the study procedures, and then connected the recipient's processor to the audio port of a Windows-based touchscreen tablet (Microsoft Surface Pro) through the DAI cable. The proprietary psychophysical software suite, PsyWorks v.6.1 (MED-EL GmbH, Innsbruck, Austria), was used to present the speech stimuli from the tablet to the audio processor. The calibration was performed using a built-in feature within the PsyWorks software and was adjusted according to each recipient's audio processor.

Intelligibility and quality measurements. Each processing condition (unprocessed, RNN, SepFormer) was evaluated with a list of twenty sentences from the test dataset for each combination of masker type and SNR (+10 and +5 dB SNR). Each subject performed a total of 13 tests (2 masker types x 3 conditions (2 models and 1 unprocessed) x 2 SNRs + 1 quiet). The testing was carried out in a self-administered manner. The subjects used the tablet and PsyWorks to present the speech materials to their

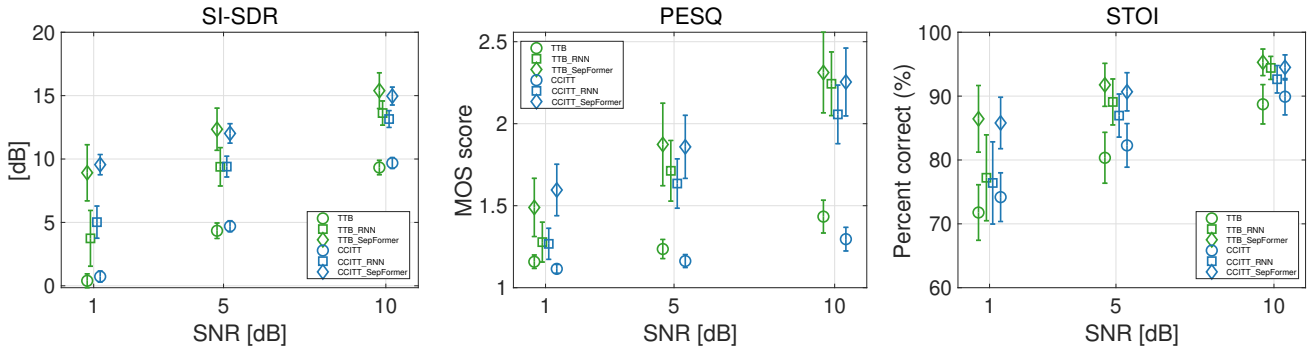
own audio processors. Subjects were assigned a unique presentation order using a "Latin square" design and were blinded to the processing condition. The subjects either vocalized their responses through a microphone located in front of them or typed them, according to their preference. The responses were captured in real-time by an automatic speech-to-text module (Google API) in the case of spoken responses. Spoken responses could be edited by typing before submission, and the PsyWorks software automatically scored words as correctly or incorrectly identified. Words containing additions, substitutions, or omissions were scored as incorrect. The percent correct scores for each condition were calculated by dividing the number of correct words by the total number of words. After each list, the percent correct was displayed and stored electronically. All participants were native English speakers, and none of the participants that elected to speak their responses had speech difficulties that jeopardized the scoring of their responses. The total testing time for all experimental conditions tested, was approximately 2.5 hours including multiple breaks.

3 Results

3.1 Objective evaluations

The models produced significant improvements across all objective evaluation metrics. The objective evaluation scores for both RNN and SepFormer models with 340 test samples are shown in Figures 2a-2c. As shown in Figure 2a, RNN processing improved SI-SDR scores over the unprocessed condition across all SNRs, for both masker types. These improvements in SI-SDR introduced by the RNN model are all statistically significant ($p < 0.0001$). The improvements in speech quality and intelligibility introduced by the RNN model can also be evidenced by the other two metrics: PESQ (Figure 2b) and STOI (Figure 2c). These improvements are also statistically significant ($p < 0.0001$). Although statistically significant, the improvement in the speech intelligibility metric (i.e., STOI) was not as prominent as in the two speech quality metrics (i.e., SI-SDR and PESQ). This is probably due to the ceiling effects: the SNR tested was high overall (starting from 1 dB SNR) and speech intelligibility was not a significant issue in these test conditions (Tang et al., 2017). For the SepFormer model, the objective evaluation scores for the processed audio signals had even more separation from the unprocessed baseline, indicating better performance by SepFormer than RNN. The superior performance of SepFormer over RNN is especially evident in the

noisier, 1-dB SNR condition across all three evaluation metrics.



(a) **(b)** **(c)**

Figure 2. (a)-(c): objective evaluation scores for unprocessed baseline and signals processed with DNN models. Higher scores are better. Statistical significance was found for all cases between the processed and unprocessed conditions. MOS: mean opinion scores.

3.2 Behavioral testing with CI listeners

3.2.1 Speech-in-noise intelligibility

The speech intelligibility scores were measured behaviorally as percent correct for the following three processing conditions: "unprocessed", "processed by RNN", and "processed by SepFormer". The DNN models introduced significant improvements in speech intelligibility scores for all CI listeners tested. As with the objective evaluation metrics, CI listeners were tested with both 2-talker babble and CCITT masker. The speech intelligibility performance was measured at both 5 and 10-dB SNR test conditions, as well as . The percent correct scores are normally distributed according to the Shapiro-Wilk test for normality. Therefore, we conducted the following statistical analyses. A three-way analysis of variance (ANOVA, with repeated measures) was carried out, involving the processing condition, SNR, and the type of masker as within-subject factors. The test results indicated that there were no statistically significant three-way interactions, but significant two-way interactions (masker type and conditions: $p = 0.03$). Since there was a significant main effect of SNR ($p < 0.0001$), we evaluated the data under each SNR separately. Masker type and processing condition were preserved due to the interaction observed between them.

10-dB SNR. Instead of showing absolute percent scores, Figure 3 depicts scores referenced to listening scores obtained in quiet without interference (therefore mostly negative values below the zero line). At first

glance, across both speech and non-speech masker types, both the RNN and SepFormer models restored speech intelligibility to performance levels approaching those of quiet settings. RNN increased the percent correct scores from the unprocessed condition by 11.1% and 18.1% on average in non-speech and speech noise, respectively. SepFormer outperformed RNN and further increased the scores from the RNN by 6.7% and 5.8% on average in non-speech and speech noise, respectively. A repeated measures two-way ANOVA indicated a statistically significant effect of the processing condition ($F[2, 24] = 31.6, p < 0.0001$). Based on post-hoc comparisons (with Bonferroni corrections), the scores from unprocessed conditions were significantly lower than those obtained in the RNN condition ($p < 0.0001$) and in the SepFormer condition ($p < 0.0001$). Neither the effect of masker type ($F[1, 12] = 1.9, p = 0.19$), nor the interaction between masker type and condition were significant ($F[2, 24] = 2.1, p = 0.14$). However, both RNN and SepFormer models appeared to have introduced greater speech enhancement in *speech noise* over non-speech stationary noise on average. The effect of masker type did not reach statistical significance probably due to the fact that 10-dB SNR is a relatively easier listening condition, where the room for improvements for the models might be have be limited. Therefore, the models were also evaluated in more challenging 5-dB SNR listening conditions as below.

5-dB SNR. As expected, when the background noise level was increased to 5-dB SNR, performance decreased overall. Without any model processing, the mean scores in the unprocessed conditions decreased by 16.3% and 21% on average in non-speech and speech noise, respectively. Similar to the 10-dB SNR condition, the scores were referenced to performance in quiet (Figure 3b). A repeated measures two-way ANOVA indicated a statistically significant effect of the processing conditions ($F[2, 24] = 38.0, p < 0.0001$). There was also a main effect of masker type ($F[1, 12] = 15.0, p = 0.002$), and a significant interaction between processing condition and masker type ($F[2, 24] = 4.0, p = 0.032$). Given the significant interaction between processing condition and masker type, post-hoc analyses were performed separately for CCITT noise and two-talker babble. Post-hoc comparisons (with Bonferroni corrections) between scores of different processing conditions in CCITT noise revealed that the SepFormer scores were significantly greater than the scores obtained in the unprocessed condition ($p = 0.018$, 24.8% increase). For scores obtained with the two-talker babble masker, both the RNN ($p = 0.013$, 23.9% increase) and SepFormer ($p = 0.0015$, 30% increase) models produced significantly better scores than the unprocessed condition. Scores obtained in the RNN and SepFormer conditions were not significantly different from one another for the CCITT noise ($p = 0.49$) or the two-talker babble ($p = 1.0$). Very interestingly, both models

introduced more performance increase in *speech noise* than non-speech noise (an increase of 23.9% vs. 12.8% by RNN; 30% vs. 24.8% by SepFormer).

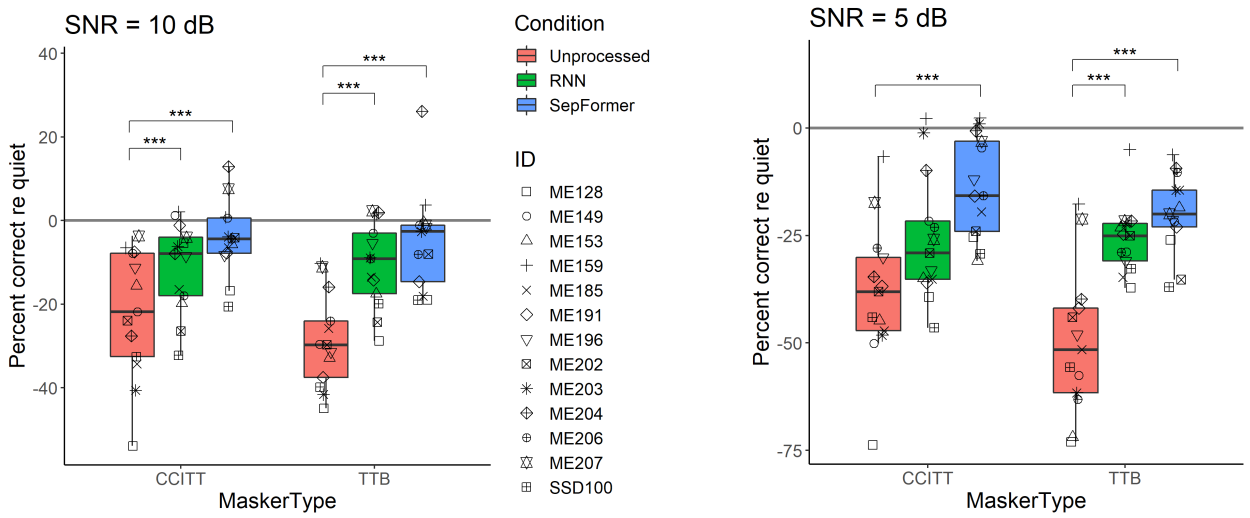
Comparison with other studies. Figure 4 shows the comparison of percent correct improvements introduced by the models in current study and those from several previous studies. In non-speech stationary noise (Figure 4a), models from current studies achieved comparable performance to traditional, signal processing algorithms (Cohen, 2003; Yang & Fu, 2005; Mauger, Arora, & Dawson, 2012). The benefit from those traditional signal processing algorithms drops in speech babble noise (Yang & Fu (2005) – 6-talker babble; Mauger, Arora, & Dawson (2012) – 4-talker babble). However, both RNN and SepFormer models introduced even *more* improvements in a 2-talker babble noise, indicating significant advantage of deep-learning based noise reduction algorithms in reducing non-stationary noise interference.

3.2.2 Subjective evaluation of the quality

In addition, as shown in Figure 5, participants rated the quality of the processed speech by both RNN and SepFormer models as approximately similar to that of the unprocessed signals before mixing with noise (i.e., in quiet). This suggests that model processing did not significantly distort the speech quality while suppressing the background interference. The benefits from model processing that have been reported so far are also consistently present across almost all individuals tested. The demographic information across individuals listed in Table 1 did not correlate with the benefits shown by the models. Considering the large individual variability that is typical of CI population, these results demonstrate great promise of these DNN models towards clinical application in the near future for better noise reduction in more complex listening environments.

4 Discussion

In this work, we implemented, trained, and tested the RNN and SepFormer models. The models were evaluated with a total of 13 CI listeners. Compared to widely adopted classification and regression models such as convolutional neural network (CNN) models, RNN does not require an input of fixed dimensions such as an image file with standardized size, which makes it more suitable for processing running audios. RNN also keeps a memory of prior information and makes predictions based on both previous and current information (i.e., through LSTM). This feature is important for input signals such as speech, of which seg-

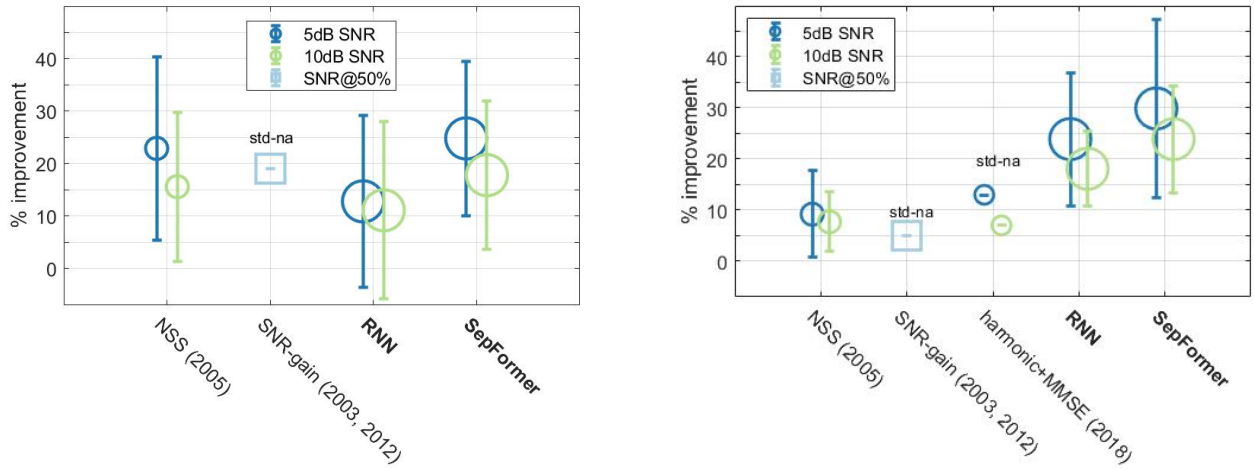


(a) 10 dB condition **(b)** 5 dB condition

Figure 3. Individual sentence recognition performance (referenced to performance in quiet) plotted as a function of each processing condition for both speech (two-talker babble, or TTB) and non-speech (speech-shaped noise, or CCITT) maskers. The boxes depict the values between the 25th and 75th percentiles, and the whiskers represent minimum and maximum values. Medians are shown as horizontal lines. (a) SNR = 10 dB (b) SNR = 5 dB;

ments of information across time interconnect with each other to form the contextual meaning. SepFormer is a transformer-based model that takes the step further by extending RNN's sequential processing to parallel processing, which best suits the need for processing longer sequences of signal. These two models were then trained with a large dataset containing various target-noise configurations, totaling 22360 instances for each training iteration. More specifically, we custom-created the large training and validation dataset from the commonly used speech and non-speech corpus in machine-listening field—LibriSpeech and WHAM!, respectively. The IEEE corpus, which is more common in hearing research, and CCITT noise was used for creating the testing dataset. The models were also trained more extensively with 100 learning cycles, to explore the full capacity of the models in noise-reduction performance.

Both RNN and SepFormer models significantly improved CI listener's speech intelligibility in noise (almost to the levels in quiet in certain conditions). This improvement in speech intelligibility in stationary non-speech noise is comparable to those demonstrated by traditional signal processing strategies. More importantly, both RNN and SepFormer introduced substantially ***more*** improvements in non-stationary ***speech-babble*** noise, where conventional signal processing strategies are limited. This highlight from both models (i.e., comparable improvement in stationary noise as conventional strategies, but more in babble-



(a) non-speech, stationary noise **(b)** speech babble noise

Figure 4. Comparison of model performance between current study and previous studies tested in similar conditions. The size of the circle is proportional to the number of CI participants tested in the study. The center of the circles align with population mean and the error bar is 1 standard deviation. Note that values from some studies are visual approximation from the figures, which explains the lack of error bar (indicated by text annotation: "std-na"). Except for "harmonic+MMSE (2018)", all models listed are traditional signal processing algorithms that are not machine-learning based.

speech noise), is more prominent in 5-dB SNR noise conditions, but still present in 10-dB SNR conditions, where the room for improvement is relatively small. It is also noteworthy that the improvements in intelligibility did not come at the price of compromising the perceived quality of the speech among CI listeners. The fact that there were no significant contributions from any demographic disparity across participants suggests that the algorithms have the potential to broadly benefit the general CI population regardless of their hearing etiology, duration of deafness, experience with CI listening, and processor settings. These results show the promise of using machine-learning based models as a complementary or even replacement algorithm for current existing signal processing strategies, to tackle more complex listening challenges involving speech-noise interference.

Given that SepFormer was the top-performing, state-of-the-art model for speech separation and enhancement at the time of the study while RNN is a basic two-layer template model, it is not surprising that the SepFormer outperformed RNN in every scenario. It showed us the current upper limit to which a DNN-based strategy could suppress the noise for CI devices. However, SepFormer is a very complicated model, containing over 26 million parameters. The processing time of such a computationally heavy model turned out to be almost 5 times the duration of the incoming signal on average, which renders

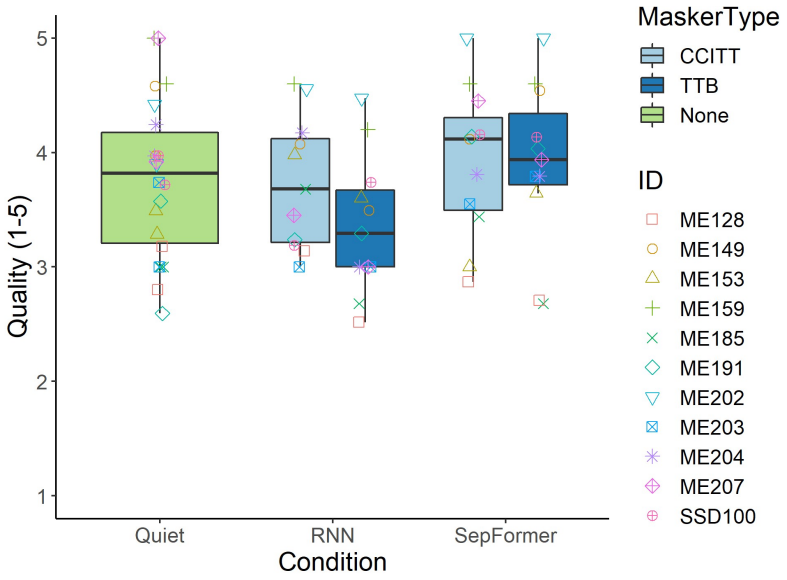


Figure 5. Subjective evaluation scores, according to the mean opinion score (MOS) scale. The score assigns a numerical measure of the human-judged overall quality: 1 corresponds to "bad", while 5 represents "excellent". Two blocks of testing were conducted for the "Quiet" condition equalize the number of samples since there were two types of maskers for each processing condition. Note that the quality was evaluated for two processing conditions in 10-dB SNR. Due to time restriction, the quality was not evaluated for 5-dB or unprocessed conditions.

it not suitable for a real-time applications such as a CI audio processor. Note that the inference and calculation of the processing time was conducted on CPU (MacBook Pro i7 2.2GHz) instead of GPU, considering hearing aids and CIs mostly adopt CPUs. The time constraint for the processing delay in a real-time device such as CI should be below about 10-20 ms to avoid disturbance in speech production and audio-visual integration (Stone & Moore, 1999; Goehring et al., 2018, 2019; Bramsløw et al., 2018). This timing delay should be even lower for individuals with single sided deafness who are fitted with CI on the deaf side while receiving acoustic input on the other side (Zirn et al., 2015). Another limitation comes from the high demand for computational power and memory. Thus, at this time, it is unrealistic to run these highly complex state-of-the-art models even in high-end audio digital signal processors used in the CI processors. The RNN model, on the other hand, is a much computationally lighter model and the processing only took around 3% of the input signal duration on average. This implies that RNNs can in theory retain low-latency inference, which makes them better-suited DNN model for real-time speech enhancement in CIs.

Future plans include optimizing and lightening the DNN models and testing them real-time in research processors with CI listeners. The training and testing should also be expended to languages other than

English. Finally, the deployment of the DNN models in research processors should be tested against existing commercial solutions for noise reduction, as well as be further evaluated in a wide variety of realistic listening environments outside research labs.

5 Conclusion

Improving speech comprehension in noisy environments is one of the most challenging problems in CI research and the development of future CI technologies. The results from this study demonstrate the promise of using DNN-based technologies for noise reduction in more complex and unpredictable, but common, multi-talker speech-babble noise interference, where current signal processing strategies fail to produce satisfactory performance. Both models evaluated in this study are single-channel models and therefore do not rely on *a priori* assumptions about the location of the sound sources. Machine learning technologies have already widely infiltrated many aspects of our daily life. Along with recent initial attempts of bringing machine learning based solutions into the field of hearing research and technology, our study further showcases the potential benefit of introducing machine learning algorithms into existing CI technologies in order to enhance speech perception in complex noisy listening settings.

Funding Acknowledgements: This research was supported by funding from the National Institutes of Health [Grant R01DC015989]; and MED-EL corporation.

Acknowledgements: We thank our research subjects for their participation in this study. We also thank Jenna Felder for helping collect data from CI participants.

Declaration of Conflicting Interests: The Authors declare that there is no conflict of interest.

Data Accessibility: All data will be made available upon reasonable requests.

References

AdvancedBionics. (2012). *CLEARVOICE, clinical results.* Retrieved from https://cochlearimplanthelp.files.wordpress.com/2012/06/ab_clearvoice_data_etc.pdf

Bentler, R., Palmer, C., & Mueller, G. H. (2006, March). Evaluation of a Second-Order Directional Microphone Hearing Aid: I. Speech Perception Outcomes. *Journal of the American Academy of Audiology*, *17*(3), 179–189. Retrieved 2023-04-07, from http://www.thieme-connect.de/DOI/DOI?10.3766/jaaa.17.3.4 (Publisher: American Academy of Audiology) doi: 10.3766/jaaa.17.3.4

Bramsløw, L., Naithani, G., Hafez, A., Barker, T., Pontoppidan, N. H., & Virtanen, T. (2018, July). Improving competing voices segregation for hearing impaired listeners using a low-latency deep neural network algorithm. *The Journal of the Acoustical Society of America*, *144*(1), 172–185. Retrieved 2022-03-20, from https://asa.scitation.org/doi/full/10.1121/1.5045322 (Publisher: Acoustical Society of America) doi: 10.1121/1.5045322

Buechner, A., Dyballa, K.-H., Hehrmann, P., Fredelake, S., & Lenarz, T. (2014, April). Advanced Beamformers for Cochlear Implant Users: Acute Measurement of Speech Perception in Challenging Listening Conditions. *PLoS ONE*, *9*(4), e95542. Retrieved 2023-04-10, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3995709/ doi: 10.1371/journal.pone.0095542

Chen, J., & Wang, D. (2017, June). Long short-term memory for speaker generalization in supervised speech separation. *The Journal of the Acoustical Society of America*, *141*(6), 4705–4714. Retrieved 2022-03-21, from https://asa.scitation.org/doi/full/10.1121/1.4986931 (Publisher: Acoustical Society of America) doi: 10.1121/1.4986931

Cohen, I. (2003, September). Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, *11*(5), 466–475. (Conference Name: IEEE Transactions on Speech and Audio Processing) doi: 10.1109/TSA.2003.811544

Cullington, H. E., & Zeng, F.-G. (2008, January). Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects. *The Journal of the Acoustical Society of America*, *123*(1), 450–461. Retrieved 2022-03-17, from https://

asa.scitation.org/doi/full/10.1121/1.2805617 (Publisher: Acoustical Society of America) doi: 10.1121/1.2805617

Dawson, P. W., Mauger, S. J., & Hersbach, A. A. (2011, June). Clinical Evaluation of Signal-to-Noise Ratio–Based Noise Reduction in Nucleus® Cochlear Implant Recipients. *Ear and Hearing*, *32*(3), 382. Retrieved 2023-04-05, from https://journals.lww.com/ear-hearing/fulltext/2011/05000/Clinical_Evaluation_of_Signal_to_Noise_Ratio_Based.11.aspx doi: 10.1097/AUD.0b013e318201c200

Dillon, H. (2008). Hearing aids. Hodder Arnold. Retrieved 2023-04-07, from https://dspace.nal.gov.au/xmlui/handle/123456789/773 (Accepted: 2017-12-06T22:37:22Z)

Dingemanse, J. G., & Goedegebure, A. (2015, June). Application of Noise Reduction Algorithm ClearVoice in Cochlear Implant Processing: Effects on Noise Tolerance and Speech Intelligibility in Noise in Relation to Spectral Resolution. *Ear and Hearing*, *36*(3), 357. Retrieved 2023-04-05, from https://journals.lww.com/ear-hearing/fulltext/2015/05000/Application_of_Noise_Reduction_Algorithm.7.aspx doi: 10.1097/AUD.0000000000000125

Dingemanse, J. G., & Goedegebure, A. (2018, March). Optimising the effect of noise reduction algorithm ClearVoice in cochlear implant users by increasing the maximum comfort levels. *International Journal of Audiology*, *57*(3), 230–235. Retrieved 2023-04-05, from https://doi.org/10.1080/14992027.2017.1390267 (Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14992027.2017.1390267) doi: 10.1080/14992027.2017.1390267

Falk, T. H., Parsa, V., Santos, J. F., Arehart, K., Hazrati, O., Huber, R., ... Scollie, S. (2015, March). Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and limitations of existing tools. *IEEE Signal Processing Magazine*, *32*(2), 114–124. (Conference Name: IEEE Signal Processing Magazine) doi: 10.1109/MSP.2014.2358871

Fu, Q.-J., Shannon, R. V., & Wang, X. (1998, December). Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing. *The Journal of the Acoustical Society of America*, *104*(6), 3586–3596. Retrieved 2022-03-17, from https://asa.scitation.org/doi/abs/10.1121/1.423941 (Publisher: Acoustical Society of America) doi: 10.1121/1.423941

Geißler, G., Arweiler, I., Hehrmann, P., Lenarz, T., Hamacher, V., & Büchner, A. (2015, March). Speech reception threshold benefits in cochlear implant users with an adaptive beamformer in real life situations. *Cochlear Implants International*, *16*(2), 69–76. Retrieved 2023-04-05, from https://doi.org/10.1179/1754762814Y.0000000088 (Publisher: Taylor & Francis _eprint: https://doi.org/10.1179/1754762814Y.0000000088) doi: 10.1179/1754762814Y.0000000088

Goehring, T., Bolner, F., Monaghan, J. J. M., van Dijk, B., Zarowski, A., & Bleeck, S. (2017, February). Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. *Hearing Research*, *344*, 183–194. Retrieved 2022-03-17, from https://www.sciencedirect.com/science/article/pii/S0378595516304348 doi: 10.1016/j.heares.2016.11.012

Goehring, T., Chapman, J. L., Bleeck, S., & Monaghan*, J. J. M. (2018, January). Tolerable delay for speech production and perception: effects of hearing ability and experience with hearing aids. *International Journal of Audiology*, *57*(1), 61–68. Retrieved 2022-03-22, from https://doi.org/10.1080/14992027.2017.1367848 (Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14992027.2017.1367848) doi: 10.1080/14992027.2017.1367848

Goehring, T., Keshavarzi, M., Carlyon, R. P., & Moore, B. C. J. (2019, July). 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*, *146*(1), 705–718. Retrieved 2022-03-15, from https://asa.scitation.org/doi/full/10.1121/1.5119226 (Publisher: Acoustical Society of America) doi: 10.1121/1.5119226

Gopalakrishna, V., Kehtarnavaz, N., Mirzahasanloo, T. S., & Loizou, P. C. (2012, June). Real-Time Automatic Tuning of Noise Suppression Algorithms for Cochlear Implant Applications. *IEEE Transactions on Biomedical Engineering*, *59*(6), 1691–1700. (Conference Name: IEEE Transactions on Biomedical Engineering) doi: 10.1109/TBME.2012.2191968

Graves, A., Mohamed, A.-r., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6645–6649). (ISSN: 2379-190X) doi: 10.1109/ICASSP.2013.6638947

Healy, E. W., Delfarah, M., Johnson, E. M., & Wang, D. (2019, March). A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and rever-

beration. *The Journal of the Acoustical Society of America*, *145*(3), 1378–1388. Retrieved 2022-03-20, from https://asa.scitation.org/doi/full/10.1121/1.5093547 (Publisher: Acoustical Society of America) doi: 10.1121/1.5093547

Henry, F., Glavin, M., & Jones, E. (2023). Noise Reduction in Cochlear Implant Signal Processing: A Review and Recent Developments. *IEEE Reviews in Biomedical Engineering*, *16*, 319–331. (Conference Name: IEEE Reviews in Biomedical Engineering) doi: 10.1109/RBME.2021.3095428

Hersbach, A. A., Arora, K., Mauger, S. J., & Dawson, P. W. (2012, August). Combining Directional Microphone and Single-Channel Noise Reduction Algorithms: A Clinical Evaluation in Difficult Listening Conditions With Cochlear Implant Users. *Ear and Hearing*, *33*(4), e13. Retrieved 2023-04-05, from https://journals.lww.com/ear-hearing/FullText/2012/07000/Combining_Directional_Microphone_and.12.aspx doi: 10.1097/AUD.0b013e31824b9e21

Hochreiter, S., & Schmidhuber, J. (1997, November). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. (Conference Name: Neural Computation) doi: 10.1162/neco.1997.9.8.1735

Hu, Y., & Loizou, P. C. (2008a, January). Evaluation of Objective Quality Measures for Speech Enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(1), 229–238. (Conference Name: IEEE Transactions on Audio, Speech, and Language Processing) doi: 10.1109/TASL.2007.911054

Hu, Y., & Loizou, P. C. (2008b, July). A new sound coding strategy for suppressing noise in cochlear implants. *The Journal of the Acoustical Society of America*, *124*(1), 498–509. Retrieved 2023-04-05, from https://asa.scitation.org/doi/full/10.1121/1.2924131 (Publisher: Acoustical Society of America) doi: 10.1121/1.2924131

Hu, Y., & Loizou, P. C. (2010, June). Environment-specific noise suppression for improved speech intelligibility by cochlear implant users. *The Journal of the Acoustical Society of America*, *127*(6), 3689–3695. Retrieved 2022-03-20, from https://asa.scitation.org/doi/full/10.1121/1.3365256 (Publisher: Acoustical Society of America) doi: 10.1121/1.3365256

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., ... Zhang, W. (2019, December). A Comparative Study on Transformer vs RNN in Speech Applications. In *2019 IEEE Automatic*

Speech Recognition and Understanding Workshop (ASRU) (pp. 449–456). doi: 10.1109/ASRU46091.2019.9003750

Kim, J., El-Khamy, M., & Lee, J. (2020, May). T-GSA: Transformer with Gaussian-Weighted Self-Attention for Speech Enhancement. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6649–6653). (ISSN: 2379-190X) doi: 10.1109/ICASSP40776.2020.9053591

Kingma, D. P., & Ba, J. (2017, January). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*. Retrieved 2022-03-25, from http://arxiv.org/abs/1412.6980 (arXiv: 1412.6980)

Kokkinakis, K., Azimi, B., Hu, Y., & Friedland, D. R. (2012, June). Single and Multiple Microphone Noise Reduction Strategies in Cochlear Implants. *Trends in Amplification*, *16*(2), 102–116. Retrieved 2023-03-27, from https://doi.org/10.1177/1084713812456906 (Publisher: SAGE Publications) doi: 10.1177/1084713812456906

Kokkinakis, K., & Loizou, P. C. (2011, May). Evaluation of objective measures for quality assessment of reverberant speech. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2420–2423). (ISSN: 2379-190X) doi: 10.1109/ICASSP.2011.5946972

Kokkinakis, K., & Stohl, J. S. (2021, August). Optimized gain functions in ideal time-frequency masks and their application to dereverberation for cochlear implants. *JASA Express Letters*, *1*(8), 084401. Retrieved 2022-07-25, from https://asa.scitation.org/doi/10.1121/10.0005740 (Publisher: Acoustical Society of America) doi: 10.1121/10.0005740

Kolbæk, M., Yu, D., Tan, Z.-H., & Jensen, J. (2017, October). Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks. , *25*(10), 1901–1913. (Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing) doi: 10.1109/TASLP.2017.2726762

Koning, R., Madhu, N., & Wouters, J. (2015, January). Ideal Time–Frequency Masking Algorithms Lead to Different Speech Intelligibility and Quality in Normal-Hearing and Cochlear Implant Listeners. *IEEE Transactions on Biomedical Engineering*, *62*(1), 331–341. (Conference Name: IEEE Transactions on Biomedical Engineering) doi: 10.1109/TBME.2014.2351854

Lai, Y.-H., Chen, F., Wang, S.-S., Lu, X., Tsao, Y., & Lee, C.-H. (2017, July). A Deep Denoising Autoencoder Approach to Improving the Intelligibility of Vocoded Speech in Cochlear Implant Simulation. *IEEE Transactions on Biomedical Engineering*, *64*(7), 1568–1578. (Conference Name: IEEE Transactions on Biomedical Engineering) doi: 10.1109/TBME.2016.2613960

Lai, Y.-H., Tsao, Y., Lu, X., Chen, F., Su, Y.-T., Chen, K.-C., ... Lee, C.-H. (2018, August). Deep Learning–Based Noise Reduction Approach to Improve Speech Intelligibility for Cochlear Implant Recipients. *Ear and Hearing*, *39*(4), 795–809. Retrieved 2022-03-15, from https://journals.lww.com/ear-hearing/fulltext/2018/07000/deep_learning_based_noise_reduction_approach_to.18.aspx?casa_token=mCO32iLZAyEAAAAA:h0F2Zb-fgVA1vujMPNehxLFHCXs4toevdOHq-XE1KGBzrn3V-AMp-Cdq_24zrRSD-KpGY8OuWN25k_wR0J2na9Cf doi: 10.1097/AUD.0000000000000537

Lesica, N. A., Mehta, N., Manjaly, J. G., Deng, L., Wilson, B. S., & Zeng, F.-G. (2021, October). Harnessing the power of artificial intelligence to transform hearing healthcare and research. *Nature Machine Intelligence*, *3*(10), 840–849. Retrieved 2021-11-01, from https://www.nature.com/articles/s42256-021-00394-z (Bandiera_abtest: a Cg_type: Nature Research Journals Number: 10 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Biomedical engineering;Diagnosis Subject_term_id: biomedical-engineering;diagnosis) doi: 10.1038/s42256-021-00394-z

Li, N., Liu, S., Liu, Y., Zhao, S., & Liu, M. (2019, July). Neural Speech Synthesis with Transformer Network. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 6706–6713. Retrieved 2022-03-22, from https://ojs.aaai.org/index.php/AAAI/article/view/4642 (Number: 01) doi: 10.1609/aaai.v33i01.33016706

Loizou, P. C., Lobo, A., & Hu, Y. (2005, November). Subspace algorithms for noise reduction in cochlear implants. *The Journal of the Acoustical Society of America*, *118*(5), 2791–2793. Retrieved 2023-04-04, from https://asa.scitation.org/doi/full/10.1121/1.2065847 (Publisher: Acoustical Society of America) doi: 10.1121/1.2065847

Luo, Y., & Mesgarani, N. (2019, August). Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(8), 1256–1266. (Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing) doi: 10.1109/TASLP.2019.2915167

Madhu, N., Spriet, A., Jansen, S., Koning, R., & Wouters, J. (2013, January). The Potential for Speech Intelligibility Improvement Using the Ideal Binary Mask and the Ideal Wiener Filter in Single Channel Noise Reduction Systems: Application to Auditory Prostheses. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(1), 63–72. (Conference Name: IEEE Transactions on Audio, Speech, and Language Processing) doi: 10.1109/TASL.2012.2213248

Mauger, S. J., Arora, K., & Dawson, P. W. (2012, November). Cochlear implant optimized noise reduction. *Journal of Neural Engineering*, *9*(6), 065007. Retrieved 2023-04-05, from https://dx.doi.org/10.1088/1741-2560/9/6/065007 (Publisher: IOP Publishing) doi: 10.1088/1741-2560/9/6/065007

Mauger, S. J., Dawson, P. W., & Hersbach, A. A. (2012, January). Perceptually optimized gain function for cochlear implant signal-to-noise ratio based noise reduction. *The Journal of the Acoustical Society of America*, *131*(1), 327–336. Retrieved 2023-04-05, from https://asa.scitation.org/doi/full/10.1121/1.3665990 (Publisher: Acoustical Society of America) doi: 10.1121/1.3665990

Mauger, S. J., Warren, C. D., Knight, M. R., Goorevich, M., & Nel, E. (2014, August). Clinical evaluation of the Nucleus® 6 cochlear implant system: Performance improvements with SmartSound iQ. *International Journal of Audiology*, *53*(8), 564–576. Retrieved 2023-06-03, from https://doi.org/10.3109/14992027.2014.895431 (Publisher: Taylor & Francis _eprint: https://doi.org/10.3109/14992027.2014.895431) doi: 10.3109/14992027.2014.895431

MED-EL. (2021). *Automatic Sound Management 3.0 with the SONNET 2 and RONDO 3 Audio Processors.* Retrieved from https://www.medel.pro/online-resources/white-papers

Picou, E. M., Aspell, E., & Ricketts, T. A. (2014, June). Potential Benefits and Limitations of Three Types of Directional Processing in Hearing Aids. *Ear and Hearing*, *35*(3), 339. Retrieved 2023-04-07, from https://journals.lww.com/ear-hearing/FullText/2014/05000/Potential_Benefits_and_Limitations_of_Three_Types.6.aspx doi: 10.1097/AUD.0000000000000004

Ricketts, T. A. (2001, December). Directional Hearing Aids. *Trends in Amplification*, *5*(4), 139–176. Retrieved 2023-04-07, from https://doi.org/10.1177/108471380100500401 (Publisher: SAGE Publications) doi: 10.1177/108471380100500401

Rix, A., Beerends, J., Hollier, M., & Hekstra, A. (2001, May). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)* (Vol. 2, pp. 749–752 vol.2). (ISSN: 1520-6149) doi: 10.1109/ICASSP.2001.941023

Roux, J. L., Wisdom, S., Erdogan, H., & Hershey, J. R. (2018, November). *SDR - half-baked or well done?* arXiv. Retrieved 2023-06-05, from http://arxiv.org/abs/1811.02508 (arXiv:1811.02508 [cs, eess]) doi: 10.48550/arXiv.1811.02508

Stone, M. A., & Moore, B. C. J. (1999, June). Tolerable Hearing Aid Delays. I. Estimation of Limits Imposed by the Auditory Path Alone Using Simulated Hearing Losses. *Ear and Hearing*, *20*(3), 182–192. Retrieved 2022-03-22, from https://journals.lww.com/ear-hearing/Fulltext/1999/06000/Tolerable_Hearing_Aid_Delays__I__Estimation_of.2.aspx?casa_token=yreyDi0_9qIAAAAA:Iihb31R3n_Z_NBPd9MPB7lodfbL4rQI4-7XEIdi5b_R2hGOmL6QUTiXGOxy6ME9BXoI4nu66HB82hfM1blELt7k

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2021, March). Attention is All You Need in Speech Separation. *arXiv:2010.13154 [cs, eess]*. Retrieved 2021-03-18, from http://arxiv.org/abs/2010.13154 (arXiv: 2010.13154)

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010, March). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4214–4217). (ISSN: 2379-190X) doi: 10.1109/ICASSP.2010.5495701

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011, September). An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 2125–2136. (Conference Name: IEEE Transactions on Audio, Speech, and Language Processing) doi: 10.1109/TASL.2011.2114881

Tang, Y., Arnold, C., & Cox, T. (2017, December). A Study on the Relationship between the Intelligibility and Quality of Algorithmically-Modified Speech for Normal Hearing Listeners. *Journal of Otorhinolaryngology, Hearing and Balance Medicine*, *1*, 5. doi: 10.3390/ohbm1010005

Verschuur, C., Lutman, M., & Abdul, N. H. W. (2013, July). Evaluation of a non-linear spectral subtraction noise suppression scheme in cochlear implant users. *Cochlear Implants International*. Retrieved 2023-03-28, from https://www.tandfonline.com/doi/abs/10.1179/cim.2006.7.4.188 (Publisher: Taylor & Francis) doi: 10.1179/cim.2006.7.4.188

Vincent, E., Gribonval, R., & Fevotte, C. (2006, July). Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(4), 1462–1469. (Conference Name: IEEE Transactions on Audio, Speech, and Language Processing) doi: 10.1109/TSA.2005.858005

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., & Schuller, B. (2015). Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR. In E. Vincent, A. Yeredor, Z. Koldovský, & P. Tichavský (Eds.), *Latent Variable Analysis and Signal Separation* (pp. 91–99). Cham: Springer International Publishing. doi: 10.1007/978-3-319-22482-4_11

Wolfe, J., Neumann, S., Marsh, M., Schafer, E., Lianos, L., Gilden, J., ... Jones, M. (2015, August). Benefits of Adaptive Signal Processing in a Commercially Available Cochlear Implant Sound Processor. *Otology & Neurotology*, *36*(7), 1181–1190. Retrieved 2022-11-17, from https://journals.lww.com/otology-neurotology/fulltext/2015/08000/Benefits_of_Adaptive_Signal_Processing_in_a.9.aspx doi: 10.1097/MAO.0000000000000781

Yang, L.-P., & Fu, Q.-J. (2005, March). Spectral subtraction-based speech enhancement for cochlear implant patients in background noise. *The Journal of the Acoustical Society of America*, *117*(3), 1001–1004. Retrieved 2023-03-28, from https://asa.scitation.org/doi/10.1121/1.1852873 (Publisher: Acoustical Society of America) doi: 10.1121/1.1852873

Zirn, S., Arndt, S., Aschendorff, A., & Wesarg, T. (2015, October). Interaural stimulation timing in single sided deaf cochlear implant users. *Hearing Research*, *328*, 148–156. Retrieved 2022-07-27, from https://www.sciencedirect.com/science/article/pii/S037859551500177X doi: 10.1016/j.heares.2015.08.010