Deep neural network algorithms for noise reduction and their application to cochlear implants

Short Title: Deep Neural Network Models for Cochlear Implants

*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*

August 25, 2022

*Correspondence: dagu@wisc.edu

Abstract

Despite excellent performance in quiet, cochlear implants (CIs) only partially restore normal levels of intelligibility in noisy settings. Recent developments in machine learning have resulted in deep neural network (DNN) models that achieve noteworthy performance in speech enhancement and separation tasks. However, there are no commercially available CI audio processors that utilize DNN models for noise reduction. We implemented two DNN models intended for applications in CIs: (1) a recurrent neural network (RNN), which is a lightweight template model, and (2) SepFormer, which is the current top-performing speech separation model in the literature. The models were trained with a custom training dataset (30 hours) that included four configurations: speech in non-speech noise and speech in 1-talker, 2-talker, and 4-talker speech babble backgrounds. The enhancement of the target speech (or the suppression of the noise) by the models was evaluated by commonly used acoustic evaluation metrics of quality and intelligibility, including (1) signal-to-distortion ratio, (2) "perceptual" evaluation of speech quality, and (3) short-time objective intelligibility. Both DNN models yielded significant improvements in all acoustic metrics tested. The two DNN models were also evaluated with thirteen CI users using two types of background noise: (1) CCITT noise (speech-shaped stationary noise) and (2) 2-talker babble. Significant improvements in speech intelligibility were observed when the noisy speech was processed by the models, compared to the unprocessed conditions. This work serves as a proof of concept for the application of DNN technology in CIs for improved listening experience and speech comprehension in noisy environments.

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

Introduction

Cochlear implant (CI) listeners struggle to understand speech in noisy environments, despite the fact that most listeners can achieve satisfactory speech intelligibility in quiet settings. This is especially true when the background noise is non-stationary and contains modulations (Cullington & Zeng, 2008; Fu et al., 1998). Some acoustic front-end processing algorithms have been shown to be capable of improving speech intelligibility in fluctuating noise. For example, some studies (Wouters & Vanden Berghe, 2001; Hersbach et al., 2012) as well as existing commercial solutions for CI (MED-EL, 2021; AdvancedBionics, 2012) have demonstrated that two-microphone directionality can result in a masking release benefit of up to 10 dB. However, these approaches require a listener to face the target, to help facilitate spatial separation between the target and background. In this work, we implemented and evaluated two single-microphone noise reduction algorithms that do not rely on spatial separation. Traditionally, most single-channel noise reduction algorithms designed for hearing-assistive devices are driven by signal-processing strategies based on time-averaged signal statistics. Classic examples include spectral subtraction (Boll, 1979), wiener filtering (Scalart & Filho, 1996), as well as commercial solutions such as ambient and transient noise reduction from MED-EL (2021), and ClearVoice from AdvancedBionics (2012). These models can improve speech intelligibility to a certain extent in statistically consistent backgrounds, such as a stationary noise (Loizou et al., 2005; Mauger et al., 2012; Dawson et al., 2011). For speech in a more complex, non-stationary background (e.g., a multi-talker babble), machine-learning techniques, such as deep neural networks (DNNs) and Gaussian mixture models (GMMs), have been shown to be successful in improving speech intelligibility for listeners with typical hearing (G. Kim et al., 2009), for listeners with hearing loss (Healy et al., 2013, 2015, 2019; Chen et al., 2016; Monaghan et al., 2017; Bramsløw et al., 2018), and for CI listeners (Hu & Loizou, 2010; Goehring et al., 2017; Lai et al., 2018). Recent improvements over these models were introduced by DNN-based regression models, where the training target was a continuous (i.e., soft) mask, instead of a binary mask (Madhu et al., 2013; Bentsen et al., 2018). However, all models mentioned above operate based on *a priori* knowledge of the target and/or background by using the same target speaker (Lai et al., 2018; Chen et al., 2016), background interference (Goehring et al., 2017), or both (G. Kim et al., 2009; Hu & Loizou, 2010; Healy et al., 2013, 2015, 2019; Goehring et al., 2017; Lai et al., 2018; Bramsløw et al., 2018; Bentsen et al., 2018) for the training and testing process.

The generalization of such models to unseen instances of noise beyond the training materials, is critical for ensuring the efficacy of the models in real-time and real-world applications that can be put to use in audio processors. It is, however, impractical to train the models with all speech-in-noise mixtures that a user would normally encounter in real-world listening situations. Indeed, in most studies, the model performance according to acoustic intelligibility metrics decreased significantly when unseen testing data were presented to the models (May & Dau, 2014; Chen & Wang, 2017; Goehring et al., 2017). Recent studies have demonstrated the potential of using recurrent neural network (RNN) models for better generalization by including recurrent connections, feedback, and gate elements (Weninger et al., 2015; Chen & Wang, 2017; Kolbæk et al., 2017; Graves et al., 2013). A classic architecture with these elements is the so-called long short-term memory (LSTM) structure that accumulates information from the past and hence enables the network to form a temporary memory (Hochmair et al., 2015; LeCun et al., 2015), which is essential for properly managing and learning speech context. RNN-LSTM based models have been shown to substantially improve speech-in-noise perception for listeners with hearing loss (Bramsløw et al., 2018; Healy et al., 2019; Keshavarzi et al., 2019, 2018) and CI users (Goehring et al., 2019).

Despite the wide adoption of RNN models in modern audio processing systems in many domains, the sequential nature of the RNN architecture often renders them difficult to train. This computational inefficiency is inherent in speech recognition and synthesis, as well as speech enhancement and speech segregation problems (Subakan et al., 2021). The impact from the limitation of RNN sequential processing becomes especially prominent with long speech sentences. Practically, this shortcoming of RNN can be avoided by using a mechanism known in the literature as a "transformer". The transformer is a fully attention-based mechanism that can effectively replace the recurrence structure present in common RNN models (Vaswani et al., 2017). This architecture allows the model to attend to the entire sequence all at once and establish connections between distinct elements, which ultimately leads to a more efficient learning of long-term dependencies. The transformer has gained competitive performance and considerable popularity in speech recognition (Karita et al., 2019), speech synthesis (Li et al., 2019), speech enhancement (J. Kim et al., 2020), and audio source separation (Subakan et al., 2021). The SepFormer model developed by Subakan et al. (2021) is currently the top-performing model in speech separation applications, according to Papers with Code website.

In this work, we applied the state-of-the-art SepFormer model towards speech enhancement. Although being state-of-the-art in terms of performance, this model is not amenable to real-time operation due to

its complex architecture. Therefore, while using SepFormer as a reference for the flagship benchmark model, we also applied a low-complexity RNN model to account for the constraints on processing time and computational power in audio processors. The effectiveness of these two proposed DNN models was verified through commonly used acoustic intelligibility metrics. We also investigated the clinical effectiveness of these two models by behaviorally testing CI listeners with challenging noise types and signal-to-noise ratio (SNR) levels. The goal of this work is to serve as a proof-of-concept towards the adoption of DNN technology in CI devices for better speech perception in complex everyday listening environments.

Network Architectures

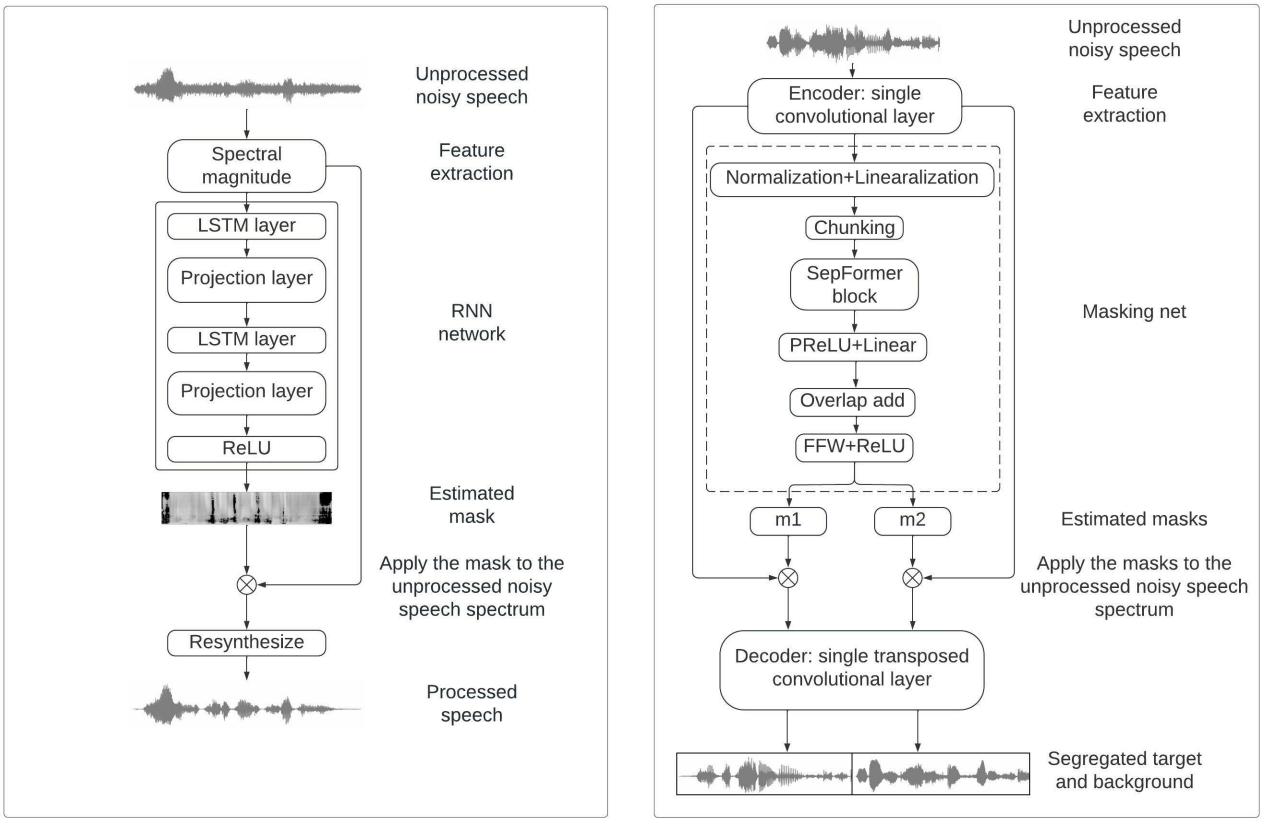
RNN

The schematic of the single-channel, RNN-based speech enhancement algorithm is illustrated in Figure 1a. A clean target speech and either babble-speech 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. 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 of) the "de-noised" spectrum. The RNN network consists of two LSTM layers, with each followed by a projection layer. 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 with all training samples (i.e., epoch). The training was terminated after 100 epochs to avoid overfitting, where the model performance with validation dataset stabilized with no further significant improvements.

SepFormer

While the simple, light-weight 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 limits of current DNN technology in speech enhancement and source separation domains. The model architecture is shown in Figure 1b. A single-layer convolutional network was used as an encoder to learn the STFT-like representation of the input noisy signal. 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 STFT-like representations back into separate sources within the mixture. The extracted STFT-like 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 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.



(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

Acoustic intelligibility models

The DNN models were first evaluated quantitatively using three commonly used acoustic evaluation metrics: source-to-distortion ratio (SDR) (Vincent et al., 2006), short-time objective intelligibility (STOI) (Taal et al., 2011, 2010), and "perceptual" evaluation of speech quality (PESQ) (Rix et al., 2001; Hu & Loizou, 2008). These accurate and reliable acoustic evaluation methods provided helpful information regarding 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 according to each metric. The 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 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 (2008) 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.

Behavioral Listening Experiments

Subjects

A total of thirteen adults fitted with MED-EL CIs (MED-EL GmbH, Innsbruck, Austria) participated in the study. They were between the ages of 20 and 72 (7 male and 6 female). Their mean age at testing was 58.6 years (SD = 14.7 years). The average duration of CI use was 6.5 years (SD = 5 years). Demographic information is provided 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.

Test setup

All participants were tested using their everyday program. Before the testing session, their audio processor was programmed with the recipient's daily 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 and 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.

Procedure

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 (TTB) at SNRs of +5 and +10 dB. Each processing condition (unprocessed, RNN, SepFormer) was evaluated with a list of twenty sentences for each combination of masker type and 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 prevented the automatic scoring of their responses. The total testing time for all experimental conditions tested, was approximately 2.5 hours including multiple breaks.

Results

Acoustic evaluations

The models produced significant improvements across all acoustic evaluation metrics. The acoustic evaluation scores for the RNN model obtained from the 340 test samples are shown in Figures 2a-2c. As shown in Figure (2a), processing with the RNN model improved SDR scores over the unprocessed condition (dashed lines) across all SNRs, as well as across both masker types (green: 2-talker babble, blue: CCITT, non-speech noise). These improvements in SDR introduced by the RNN model (i.e., the elevation from the dashed lines to solid lines) were all statistically significant (p-values less than 0.0001). The improvements in speech quality and intelligibility introduced by the RNN model can also be evidenced by the other two metrics: PESQ (see Figure 2b) and STOI (see Figure 2c). These improvements were also statistically significant, with a p-value that is less than 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., SDR and PESQ). This is probably because of ceiling effects; the SNR tested was high overall (starting from 1 dB SNR) and speech intelligibility was not a significant issue in these models of normal hearing (Tang et al., 2017). For the SepFormer model, the acoustic evaluation scores for the unprocessed noisy mixtures (dashed lines) remained the same since the test materials did not change, as shown in Figure 2d-2f. However, the scores for the processed audio signals by the SepFormer model (solid lines) had even more separation from the dashed lines, indicating better performance by SepFormer than RNN. The superior performance of SepFormer over RNN is especially evident in the 1 dB SNR condition across all three evaluation metrics.

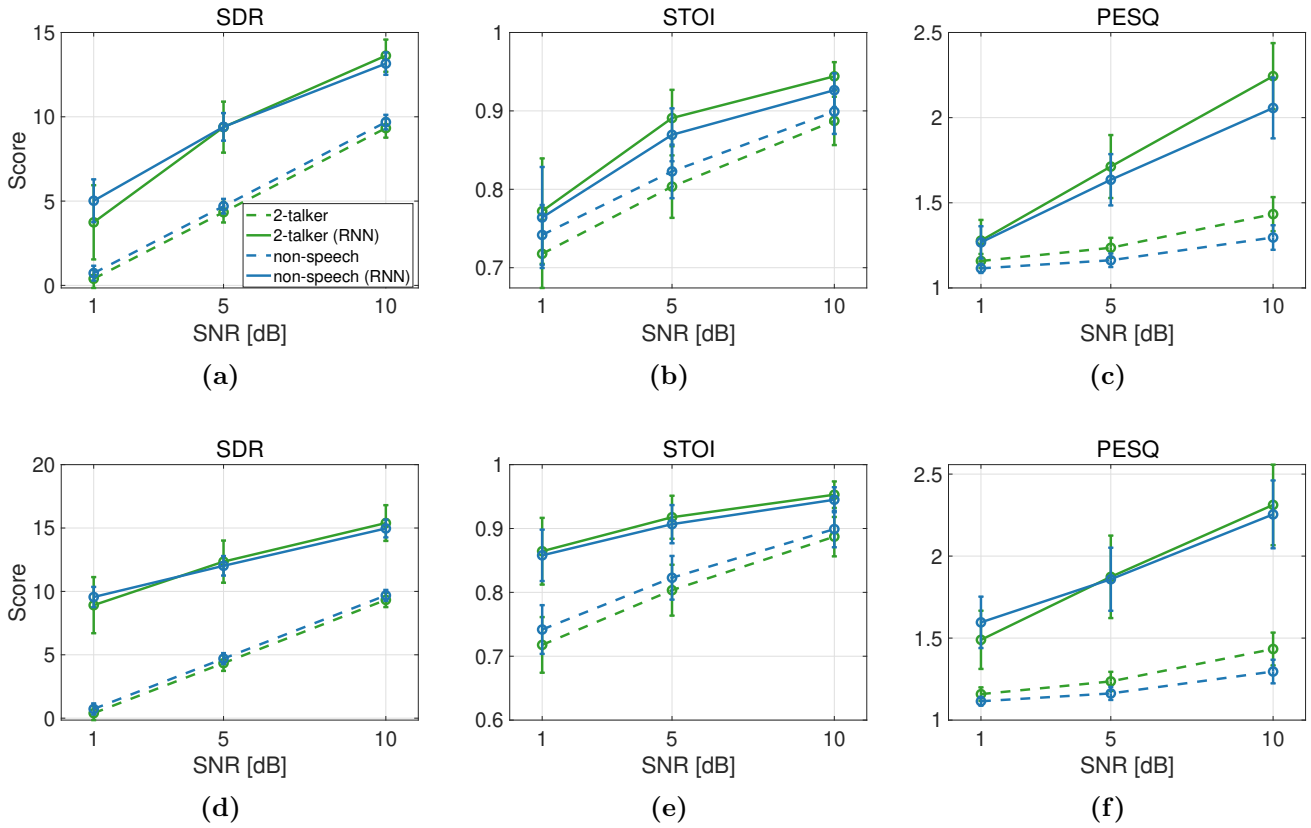


Figure 2. (a)-(c): acoustic evaluation scores for signals processed with the RNN architecture; (d)-(f): acoustic evaluation scores for signals processed with the SepFormer architecture.

Behavioral testing with CI listeners

The speech intelligibility scores were measured behaviorally as percent correct for the three testing conditions, including "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 acoustic evaluation metrics, CI listeners were tested with both 2-talker babble and CCITT masker, as well as both 5 and 10-dB SNR conditions. 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), involving the processing condition, SNR, and the type of masker as within-subject factors, was carried out. The test results indicated that there were no statistically significant three-way, but significant two-way interactions (masker type and conditions: $p = 0.03$). Since there was a significant main effect of SNR ($p < .0001$), we evaluated the data under each SNR separately. Masker type and condition were preserved due to the interaction observed between them.

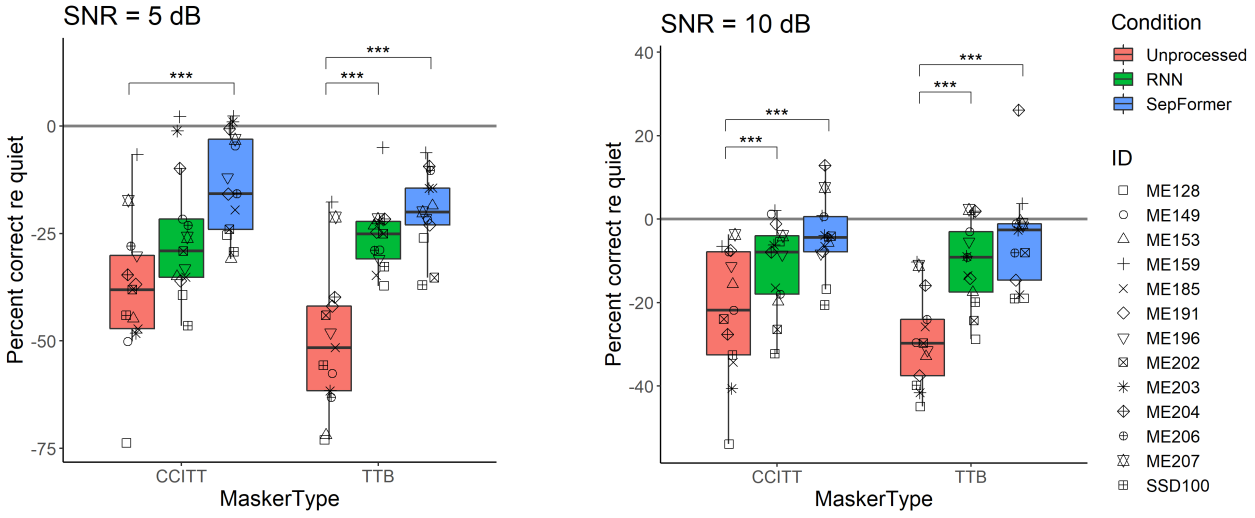
For the 5 dB SNR conditions, a repeated measures two-way ANOVA indicated a statistically significant effect of the processing conditions ($F[2, 24] = 38.0, p = 3.4e - 08$). There was also a main effect of masker type ($F[1, 12] = 15.0, p = 2.0e - 03$), and a significant interaction between processing condition and masker type ($F[2, 24] = 4.0, p = 3.2e - 02$). 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 in CCITT noise revealed that the SepFormer scores were significantly greater than the scores obtained in the unprocessed condition ($p = 1.8e - 2$, 32.1% increase). For scores obtained with the two-talker babble masker, both the RNN ($p = 1.3e - 2$, 37.1% increase) and SepFormer ($p = 1.5e - 3$, 41% 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 ($p = 4.9e - 1$) or the two-talker babble ($p = 1.0$). Very interestingly, both models introduced remarkably more performance increase in speech noise than non-speech noise (an increase of 37.1% vs. 14.5% by RNN; 41% vs. 32.1% by SepFormer).

As expected, when the background noise level was reduced such that the SNR increased from 5 to 10 dB, performance increased overall. Without any model processing, the median scores in the unprocessed conditions increased by 28.3% and 31.1% in non-speech and speech noise, respectively. Both models further improved the scores for both masker types: RNN introduced an increase of 3.4% and 18% in non-speech and speech noise, respectively. As shown previously, SepFormer still outperformed RNN and further increased the scores by 13.4% and 22.9% non-speech and speech noise, respectively. Even though the room for improvements from model processing were limited in these relatively easier conditions, the aforementioned trend of greater speech enhancement in speech noise over non-speech stationary noise persists in both models. In the 10 dB SNR conditions, a repeated measures two-way ANOVA indicated a statistically significant effect of the processing condition ($F[2, 24] = 31.6, p = 1.9e{-}07$). Based on post-hoc comparisons (with Bonferroni corrections),the scores from unprocessed conditions were significantly lower than those obtained in the RNN condition ($p = 2.9e - 02$) and in the SepFormer condition ($p = 9.5e - 04$). Neither the effect of masker type ($F[1, 12] = 1.9, p = 1.9e - 1$), nor the interaction between masker type and condition were significant ($F[2, 24] = 2.1, p = 1.4e - 1$).

Instead of showing absolute scores, Figure 3 depicts scores for each masker type separately, with reference to listening scores obtained in quiet without interference (therefore mostly negative values below the zero line). Figure 3a shows the data collected in the 5 dB SNR condition, while Figure 3b shows

the data obtained in the 10 dB SNR condition. At first glance, in both 5 and 10 dB SNR conditions, 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. The aforementioned statistical significance still remains in these referenced scores. Note that We showed the 10-dB SNR data separated by masker type to make it match the 5-dB SNR figure. Note also that there was no significant effect of masker type, and no interaction with condition. Therefore, statistical differences represent the pattern of results when scores from the two masker types were averaged.

In addition, as shown in Figure 4, 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. Considering the large individual variability that is typical of CI population, these results demonstrate great promise towards clinical application in the near future for better noise reduction in more complex listening environments.



(a) 5 dB condition **(b)** 10 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 thick horizontal lines. (a) SNR = 5 dB (b) SNR = 10 dB;

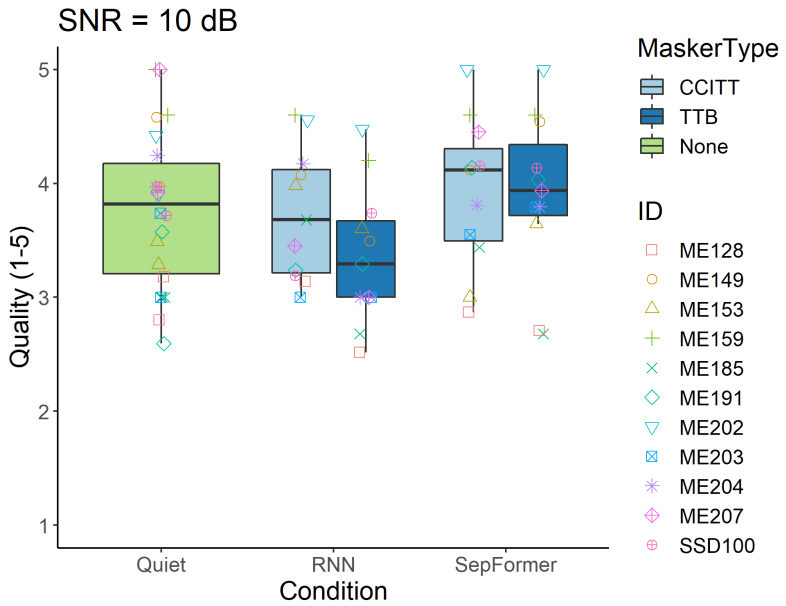


Figure 4. Subjective evaluation scores. According to the mean opinion score (MOS) scale which assigns a numerical measure of the human-judged overall quality: 1 corresponds to "bad", while 5 represents "excellent". Two blocks of testing were conducted for the "Quiet" condition since there were two types of maskers for each processing condition.

Improvement with models vs. demographic factors

A linear, mixed effects model with the change in score relative to the unprocessed noisy condition as the dependent variable, masker type, model (RNN/SepFormer), and demographic factors as independent variables, and ID as a random factor was performed. The demographic information is listed in Table 1 (gender, age at testing, duration of CI use, age at onset of hearing loss, number of active electrodes, and coding strategy). A stepwise regression analysis was performed on the full linear model, and the output of the stepwise analysis was a linear model that only included masker type and DNN model, with ID as a random factor (demographic factors always increased the Akaiki criterion of the linear model.) This suggests that, for this sample, the potential benefit of a DNN-based speech enhancement algorithm was not related to any of the selected demographic factors.

Discussion

Both the RNN and SepFormer models introduced significant improvements in speech quality and intelligibility. The improvements were demonstrated across all acoustic evaluation metrics, including SDR,

PESQ, and STOI. The improvement in the evaluation scores from the unprocessed to processed noisy mixtures was statistically significant across all SNR test conditions, in both masker types, across all three metrics, and for both models. The observed statistical significance is visually apparent for the two quality metrics assessed herein: SDR and PESQ. The observed differences are not as apparent for the intelligibility metric—STOI. This is perhaps because the room for improvement in normal-hearing model intelligibility is small in the relatively high-SNR conditions we tested, with the worst condition being 1 dB SNR. In general, the SepFormer model, as the current state-of-the-art, outperformed the simpler RNN model. This advantage is especially prominent in lower-SNR conditions. For instance, in the 1 dB SNR condition, the SepFormer introduced an improvement of 8.52 in SDR score as compared to the improvement of 3.35 from the RNN model. Even in STOI scores from the intelligibility metric, where the benefits from both models were not as great as in the other two evaluation metrics of quality, the SepFormer model still outperformed the simpler RNN structure.

The improvements in speech quality and intelligibility shown across the acoustic evaluation metrics were largely mirrored in the behavioral testing conducted with CI listeners. The participants' intelligibility scores were significantly improved from unprocessed conditions after processing with the both DNN models, closing the distance between the scores with and without the presence of background noise. Given that SepFormer is the current state-of-the-art model for speech enhancement, while RNN is a simple two-layer template model, it is not surprising that the SepFormer outperformed the RNN in every scenario. Both models still demonstrated improvements in speech intelligibility scores, even in the 10 dB SNR conditions, where "de-noising" or enhancement of the target speech stimulus was a relatively straightforward task. Both models introduced more improvements in speech noise than non-speech noise, which is contrary to the benefit demonstrated by traditional signal processing strategies that are designed for removing statistically predictable and relatively stationary noises. These traditional signal processing algorithms fall short in more complex non-stationary speech backgrounds (Boll, 1979; Scalart & Filho, 1996; Dawson et al., 2011; Loizou et al., 2005; Mauger et al., 2012). The better performance in speech-in-noise demonstrated by the RNN and SepFormer models is in agreement with the results from previous studies on machine-earning based noise-reduction models, such as DNNs and Gaussian mixture models (GMMs) (G. Kim et al., 2009; Bramsløw et al., 2018; Chen et al., 2016; Healy et al., 2019, 2015, 2013; Monaghan et al., 2017; Goehring et al., 2017; Hu & Loizou, 2010; Lai et al., 2018). This result shows the promise of using machine-learning based models as a complementary algorithm to current existing

signal processing strategies, especially for tackling more complex listening challenges, where existing solutions generally fail (Loizou et al., 2005; Mauger et al., 2012; Dawson et al., 2011). The model processing also introduced statistically negligible distortions in perceived speech quality as shown in the subjective evaluation scores. The fact that there were no significant contributions from any demographic disparity among participants suggests that the algorithms could broadly benefit the CI population regardless of their hearing etiology, duration of deafness, experience with CI listening, and processor settings.

These results are consistent with previous studies on the application of machine learning algorithms to "de-noising" strategies for CIs (Hu & Loizou, 2010; Goehring et al., 2017, 2019; Lai et al., 2018). Our study further confirms the promise of using DNN-based algorithms to improve speech-in-noise hearing outcomes in CIs. While Hu & Loizou (2010) and Lai et al. (2018) used the same speaker for training and testing, our study further investigated the models' capability of generalization by using different speech and non-speech materials for testing vs. training. We also used a much larger and more diverse training dataset. The training dataset was approximately 30 hours in total duration with four distinct target-masker configurations. Within each configuration, ten different SNR conditions in a 1-dB step were implemented with equal representation. The models were also trained more extensively with 100 learning cycles or epochs, whereas some other studies trained their models with a much less number of learning cycles. For instance, Goehring et al. (2019) trained their model for only one epoch. Although the successful training with such a small number of learning cycles helps demonstrate the model's promise of being continuously trained "on the go" with smaller resources, we wanted to explore the full capacity of the models for noise reduction through a more thorough training process.

In addition to the basic RNN model, we also implemented and tested the current best-performing model in the field of speech separation—SepFormer, to explore the current limit to which a DNN-based strategy could suppress the noise for CI devices. While, not surprisingly, the SepFormer model outperformed RNN in every test, it 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 it not suitable for a real-time applications such as a CI audio processor. 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 (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 existing CIs. The RNN model, on the other hand, was 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 during inference, which makes them well-suited for real-time speech enhancement in assisted hearing devices.

Conclusions

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 everyday listening environments, 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. Future machine learning models could consider simulating more aspects of the "cocktail-party listening", such as spatial-hearing sensitivity, for even better de-noising performance. Machine learning technologies have already widely infiltrated many aspects of our daily life. Along with other recent research efforts, 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. Future plans include the deployment of the DNN models tested herein in a research processor to be tested against existing commercial solutions for noise reduction, as well as to be further evaluated in a wider variety of listening conditions.

Table 1. Demographic information for the subjects who participated in this study

Subject ID	**Gender**	**Age at time of testing (yrs)**	**Age at onset of hearing loss (yrs)**	**Duration of CI use (months)**	**Active electrodes**	**Clinical strategy**	**IEEE score in quiet (pcnt correct)**
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

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.

[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.

References

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

Bentsen, T., May, T., Kressner, A. A., & Dau, T. (2018, May). The benefit of combining a deep neural network architecture with ideal ratio mask estimation in computational speech segregation to improve speech intelligibility. *PLOS ONE*, *13*(5), e0196924. Retrieved 2022-03-17, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196924 (Publisher: Public Library of Science) doi: 10.1371/journal.pone.0196924

Boll, S. (1979, April). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *27*(2), 113–120. (Conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing) doi: 10.1109/TASSP.1979.1163209

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

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

Chen, J., Wang, Y., Yoho, S. E., Wang, D., & Healy, E. W. (2016, May). Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises. *The Journal of the Acoustical Society of America*, *139*(5), 2604–2612. Retrieved 2022-03-20, from https://asa.scitation.org/doi/full/10.1121/1.4948445 (Publisher: Acoustical Society of America) doi: 10.1121/1.4948445

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), 382–390. Retrieved 2022-03-20, from https://journals.lww.com/ear-hearing/fulltext/2011/05000/Clinical_Evaluation_of_Signal_to_Noise_Ratio_Based.11.aspx?casa_token=-0kDlsdi-ScAAAAA:5Wm499dl3lCixxtBYmN5LHOcXL4qrKIFGYlYLBdctDIdf67m0oPoSq5Fh5Pj9ja068BbUzZWyLPYmkoAZJuom5c doi: 10.1097/AUD.0b013e318201c200

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

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

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 reverberation. *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

Healy, E. W., Yoho, S. E., Chen, J., Wang, Y., & Wang, D. (2015, September). An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type. *The Journal of the Acoustical Society of America*, *138*(3), 1660–1669. Retrieved 2022-03-20, from https://asa.scitation.org/doi/full/10.1121/1.4929493 (Publisher: Acoustical Society of America) doi: 10.1121/1.4929493

Healy, E. W., Yoho, S. E., Wang, Y., & Wang, D. (2013, October). An algorithm to improve speech recognition in noise for hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *134*(4), 3029–3038. Retrieved 2022-03-20, from https://asa.scitation.org/doi/full/10.1121/1.4820893 (Publisher: Acoustical Society of America) doi: 10.1121/1.4820893

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 2022-03-17, from https://journals.lww.com/ear-hearing/FullText/2012/07000/Combining_Directional_Microphone_and.12.aspx?casa_token=OGC6EKKsm6AAAAAA:_cQ0afdE_71RjPLpCBBH5YsVmBQ-lDgkhcWSgll7xfSburVOJ_NhVCEZ8q8Gd6i-TaIBVcOKtro7BMFQ54jx0Ks doi: 10.1097/AUD.0b013e31824b9e21

Hochmair, I., Hochmair, E., Nopp, P., Waller, M., & Jolly, C. (2015, April). Deep electrode insertion and sound coding in cochlear implants. *Hearing Research*, *322*, 14–23. Retrieved 2021-08-23, from https://

www.sciencedirect.com/science/article/pii/S0378595514001701 doi: 10.1016/j.heares.2014.10.006

Hu, Y., & Loizou, P. C. (2008, 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. (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

Keshavarzi, M., Goehring, T., Turner, R. E., & Moore, B. C. J. (2019, March). Comparison of effects on subjective intelligibility and quality of speech in babble for two algorithms: A deep recurrent neural network and spectral subtraction. *The Journal of the Acoustical Society of America*, *145*(3), 1493–1503. Retrieved 2022-03-21, from https://asa.scitation.org/doi/full/10.1121/1.5094765 (Publisher: Acoustical Society of America) doi: 10.1121/1.5094765

Keshavarzi, M., Goehring, T., Zakis, J., Turner, R. E., & Moore, B. C. J. (2018, January). Use of a Deep Recurrent Neural Network to Reduce Wind Noise: Effects on Judged Speech Intelligibility and Sound Quality. *Trends in Hearing*, *22*, 2331216518770964. Retrieved 2022-03-21, from https://doi.org/10.1177/2331216518770964 (Publisher: SAGE Publications Inc) doi: 10.1177/2331216518770964

Kim, G., Lu, Y., Hu, Y., & Loizou, P. C. (2009, September). An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, *126*(3), 1486–1494. Retrieved 2022-03-20, from https://asa.scitation.org/doi/full/10.1121/1.3184603 (Publisher: Acoustical Society of America) doi: 10.1121/1.3184603

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., & 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

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

LeCun, Y., Bengio, Y., & Hinton, G. (2015, May). Deep learning. *Nature*, *521*(7553), 436–444. Retrieved 2022-03-21, from https://www.nature.com/articles/nature14539 (Number: 7553 Publisher: Nature Publishing Group) doi: 10.1038/nature14539

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 2022-03-20, 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., 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 2022-03-20, from https://asa.scitation.org/doi/full/10.1121/1.3665990 (Publisher: Acoustical Society of America) doi: 10.1121/1.3665990

May, T., & Dau, T. (2014, December). Requirements for the evaluation of computational speech segregation systems. *The Journal of the Acoustical Society of America*, *136*(6), EL398–EL404. Retrieved 2022-03-21, from https://asa.scitation.org/doi/full/10.1121/1.4901133 (Publisher: Acoustical Society of America) doi: 10.1121/1.4901133

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

Monaghan, J. J. M., Goehring, T., Yang, X., Bolner, F., Wang, S., Wright, M. C. M., & Bleeck, S. (2017, March). Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *141*(3), 1985–1998. Retrieved 2022-03-20, from https://asa.scitation.org/doi/full/10.1121/1.4977197 (Publisher: Acoustical Society of America) doi: 10.1121/1.4977197

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

Scalart, P., & Filho, J. (1996, May). Speech enhancement based on a priori signal to noise estimation. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings* (Vol. 2, pp. 629–632 vol. 2). (ISSN: 1520-6149) doi: 10.1109/ICASSP.1996.543199

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. Retrieved 2022-03-22, from https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

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

Wouters, J., & Vanden Berghe, J. (2001, October). Speech Recognition in Noise for Cochlear Implantees with a Two-Microphone Monaural Adaptive Noise Reduction System. *Ear and Hearing*, *22*(5), 420–430.

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