DNN-BASED ENHANCEMENT OF NOISY AND REVERBERANT SPEECH

Yan Zhao[⋆] *DeLiang Wang*[⋆†] *Ivo Merks*[∗] *Tao Zhang*[∗]

[⋆] Department of Computer Science & Engineering, The Ohio State University, Columbus, OH 43210 USA
[†] Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210 USA
[∗] Starkey Hearing Technologies, Eden Prairie, MN 55344 USA

zhao.836@osu.edu, dwang@cse.ohio-state.edu, {ivo_merks, tao_zhang}@starkey.com

ABSTRACT

In the real world, speech is usually distorted by both reverberation and background noise. In such conditions, speech intelligibility is degraded substantially, especially for hearing-impaired (HI) listeners. As a consequence, it is essential to enhance speech in the noisy and reverberant environment. Recently, deep neural networks have been introduced to learn a spectral mapping to enhance corrupted speech, and shown significant improvements in objective metrics and automatic speech recognition score. However, listening tests have not yet shown any speech intelligibility benefit. In this paper, we propose to enhance the noisy and reverberant speech by learning a mapping to reverberant target speech rather than anechoic target speech. A preliminary listening test was conducted, and the results show that the proposed algorithm is able to improve speech intelligibility of HI listeners in some conditions. Moreover, we develop a masking-based method for denoising and compare it with the spectral mapping method. Evaluation results show that the masking-based method outperforms the mapping-based method.

Index Terms— speech intelligibility test, speech denoising, spectral mapping, ideal ratio mask, deep neural networks

1. INTRODUCTION

In daily environments, room reverberation and background noise both distort the speech signal. Such distortions severely degrade the performance of automatic speech recognition (ASR) and speaker identification (SID), as well as the ability of listeners to understand speech. While normal-hearing (NH) listeners are able to tolerate such distortions to a large extent, hearing-impaired (HI) listeners show poor performance [1]. Even though a lot of effort has been made to combat reverberation and noise, and substantial performance improvements on the ASR [2, 3, 4] and the SID [5] tasks have been obtained, no monaural algorithm has been able to improve speech intelligibility of HI listeners in the noisy and reverberant conditions. Therefore, denoising and dereverberation remain a major challenge.

Our objective is to improve speech intelligibility of HI subjects. It is well documented that without background noise, both NH and HI listeners show considerable tolerance to room reverberation. In other words, with no noise, human speech recognition is impaired only when the reverberation time (T_{60}) is long. For HI listeners, T_{60} needs to be at least 1 s before the intelligibility score drops to below 50% [6, 7, 8]; this is the case even for cochlear implantees [9]. For NH listeners, the recognition rates drop to below 50% with 2 s or longer T_{60} [6, 10]. In real-world environments, the reverberation time is typically less than 1 s; that is to say, removing noise alone should potentially provide speech intelligibility improvements. Furthermore, not all reverberation is harmful to speech intelligibility. Indeed early reflections can benefit speech intelligibility [11, 12].

Han et al. [13] proposed a spectral mapping approach to perform dereverberation, which has been extended to enhance the noisy and reverberant speech [4]. The idea is to utilize deep neural networks (DNNs) to learn a mapping function from the magnitude spectrum of noisy and reverberant speech to that of corresponding noise-free and anechoic speech. However, informal listening clearly indicates that the method does not improve speech intelligibility. One possible reason is the different nature of reverberation and noise. In general, room reverberation corresponds to a convolution process of a direct sound with a room impulse response (RIR) [14], while background noise is usually considered an additive signal to clean speech. Learning a mapping function to deconvolve and denoise simultaneously may be too difficult for a standard DNN.

The above observations motivate us to pursue a different mapping function to remove noise only. The rationale for just performing denoising is that even HI listeners can tolerate a significant amount of reverberation without intelligibility degradation. The results of a preliminary listening test demonstrate that this new mapping function can lead to intelligibility benefits for HI listeners. In addition, we develop a masking-based method to denoise noisy and reverberant speech. Evaluation results show that our masking-based method outperforms spectral mapping in terms of predicted intelligibility score.

The paper is organized as follows. The next section will discuss the relation to previous work. Section 3 and section 4 describe the proposed algorithm and objective evaluation results. In section 5, the results of a preliminary listening test are shown. Section 6 presents our masking-based method and comparisons with the spectral mapping method. Concluding remarks are presented in the last section.

2. RELATION TO PRIOR WORK

Despite many speech enhancement algorithms have been proposed to deal with noisy and reverberant speech [2, 3, 4, 15], they do not address hearing impairment. On the other hand, a DNN-based speech segregation system with demonstrated intelligibility improvements for HI listeners [16] does not deal with room reverberation. The aim of our proposed algorithms is to improve the speech intelligibility of HI listeners in both reverberant and noisy environments. Different

This research was supported in part by an AFOSR grant (FA9550-12-1-0130), an NIDCD grant (R01 DC012048), a contract from Starkey, and the Ohio Supercomputer Center. Part of this research was conducted when D. L. Wang was a visiting scholar at Starkey.

from [4], we design a new mapping function which utilizes clean reverberant speech as the desired signal. Furthermore, with the new target signal, we develop a masking-based method to enhance corrupted speech.

3. ALGORITHM DESCRIPTION

Features of a noisy and reverberant signal are extracted by short time Fourier transform (STFT). Given a time domain signal $s(t)$ sampled at 16 kHz, we divide the signal using a 20 ms frame window with a 10 ms window shift. For each time frame, a 320-point fast Fourier transform (FFT) is applied resulting in 161 frequency bins. Only magnitude information is considered. Therefore, at time frame m, we obtain a 161-dimensional feature vector $\boldsymbol{y}(m)$ of noisy and reverberant speech. In order to take advantage of temporal information, we employ a context window to incorporate features of adjacent time frames. Hence, the feature vector $\boldsymbol{F}$ at time frame m is constructed as

$$\boldsymbol{F}(m) = [\boldsymbol{y}(m-c), \cdots, \boldsymbol{y}(m), \cdots, \boldsymbol{y}(m+c)] \quad (1)$$

where c is the context window size. Although a bigger window encodes more context information, our experiments suggest that performance gain becomes slight when the window size increases beyond a certain point. Considering the computational cost, we set $c = 5$. Consequently, the feature dimension for the DNN input is $11 \times 161 = 1771$. Since the magnitude spectrum has a large dynamic range, a log operation is applied to compress the values. Before DNN training, the features are normalized to zero mean and unit variance.

To learn a mapping function, the log magnitude spectrum of clean reverberant speech is treated as the desired output of the DNN-based enhancement system. For the purpose of a bounded training target, the target log magnitude spectrum is normalized to the range of [0,1].

The loss function is mean square error (MSE). Since the DNN is trained to learn a mapping function f from the log magnitude spectrum of noisy and reverberant speech to that of corresponding clean reverberant speech, the loss function is defined as follows,

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{F}; \Theta) = \|\boldsymbol{x} - f(\boldsymbol{F})\|^2 \quad (2)$$

where $\boldsymbol{F}$ denotes the normalized feature vector; $\boldsymbol{x}$ denotes the normalized log magnitude spectrum of clean reverberant speech; Θ denotes the parameters of the mapping function f, which is learned during the training phase.

Fig. 1 shows a diagram of the proposed DNN-based spectral mapping algorithm. The DNN architecture includes 4 hidden layers with 1024 units in each layer. This setup is a trade-off between performance and computational cost. The activation function for the hidden layers is the rectified linear function (ReLU) [17]. For the output layer, the sigmoidal activation function is used. For training the neural network, adaptive gradient descent [18] is utilized as the optimization method.

The time domain signal is resynthesized by using the phase of noisy and reverberant speech. Although post-processing on corrupted phase can slightly improve objective speech intelligibility and quality metrics [4], we do not find speech intelligibility benefits in informal listening tests. Thus no post-processing is applied in any of our experiments.

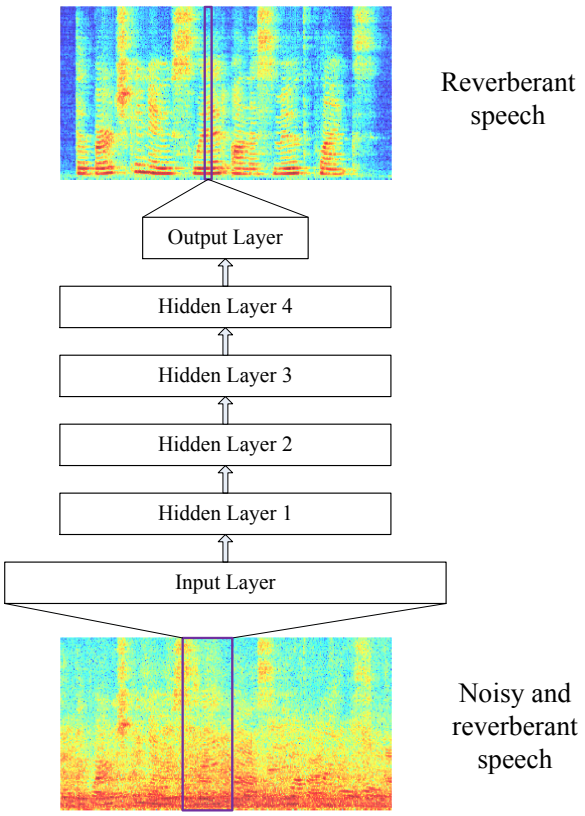

Fig. 1. DNN-based spectral mapping

4. EVALUATION RESULTS

We evaluate the proposed algorithm on the IEEE corpus [19] spoken by a female speaker. There are 72 phonetically balanced lists in the corpus, each with 10 sentences. Sentences from the first 30 lists are selected to generate training data. Sentences from list 70-72 are used to generate validation data. The proposed algorithm is tested on sentences from list 51-60.

Two reverberant rooms are simulated: room 1 with size $10\ m \times 7\ m \times 3\ m$ and room 2 with size $5\ m \times 6\ m \times 3\ m$. Room 1 is used to generate RIRs for training and validation sets, and room 2 for testing. Different RIRs are generated with the positions of receiver and speaker randomly chosen while fixing their distance to 4 m. Three values of T_{60} are considered, namely, 0.3 s, 0.6 s and 0.9 s. For training and validation sets, we generate two RIRs for each T_{60}; for test set, we generate one RIR corresponding to one T_{60}. All RIRs are generated by using an RIR generator [20] which employs the image model [21]. Consequently, there are $300\times3(T_{60}\text{s})\times2$(RIRs) = 1800 reverberant utterances in the training set, $30\times3(T_{60}\text{s})\times2$(RIRs) = 180 reverberant utterances in the validation set, and $100\times3(T_{60}\text{s})\times1$(RIR) = 300 reverberant utterances in the test set.

Babble noise and speech shaped noise (SSN) are used in our study, with babble noise being nonstationary and SSN stationary. Both noises last about 4 min. We divide the noise into two parts: the first 3 min is used for training and validation and the remaining noise is used for testing. Thus there is no noise overlap between train-

ing/validation data and test data. The noisy and reverberant speech is constructed by

$$y(t) = x(t) + \alpha n(t) \quad (3)$$

where $y(t)$, $x(t)$ and $n(t)$ denote noisy and reverberant speech, reverberant speech, and noise signal, respectively; α is a parameter used to adjust signal-to-noise ratio (SNR). Note that in SNR calculation, reverberant speech ($x(t)$) is treated as signal [22]. To add noise to reverberant speech, we randomly select a segment from noise signal and add it to reverberant target at a specified SNR. Noisy and reverberant speech is mixed at three SNRs, -5 dB, 0 dB and 5 dB. Finally, we get 1800×3(SNRs)×2(noises) = 10800 utterances for training, 180×3(SNRs)×2(noises) = 1080 utterances for validation, and 300×3(SNRs)×2(noises) = 1800 utterances for testing.

	STOI			PESQ		
	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
unprocessed	0.413	0.563	0.706	1.247	1.651	2.059
STFT → STFT	0.546	0.665	0.752	1.688	2.077	2.417

Table 1. Average STOI and PESQ scores after enhancement by proposed spectral mapping at each SNR for babble noise.

	STOI			PESQ		
	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
unprocessed	0.436	0.588	0.730	1.285	1.619	1.986
STFT → STFT	0.596	0.706	0.776	1.782	2.140	2.436

Table 2. Average STOI and PESQ scores after enhancement by proposed spectral mapping at each SNR for SSN.

Short-time objective intelligibility (STOI) [23] and perceptual evaluation of speech quality (PESQ) [24] are employed to evaluate speech intelligibility and quality, respectively. These are standard objective metrics where the value range for STOI is between 0 and 1, roughly corresponding to recognition rate, and the value range for PESQ is between -0.5 and 4.5. Since we only remove noise from noisy and reverberant speech, clean reverberant speech is used as the reference signal in the evaluation.

Table 1 and Table 2 list the average STOI and PESQ values of unprocessed and processed signals. In the tables, we denote the spectral mapping algorithm by "**STFT→STFT**". At a specified SNR level, we average the evaluation values across the three values of T_{60}. By comparing STOI and PESQ values of unprocessed noisy-reverberant speech with those of enhanced signals, we see a clear improvement in predicted intelligibility and quality at all three SNRs and with two noises. Smaller improvements are obtained at higher SNRs since, under such conditions, speech intelligibility and quality of unprocessed signals are better, and less room exists for further improvements.

5. PRELIMINARY LISTENING TEST

Although improvements on STOI indicate potential improvements on actual speech intelligibility, it is important to validate whether the proposed approach can improve speech intelligibility of HI listeners in noisy and reverberant environments. To answer this question, we conducted a preliminary intelligibility test.

5.1. Test Methodology

To avoid repeated sentences heard by one subject, test stimuli consist of 32 lists (list 39-70) from the IEEE corpus. Out of the remaining 40 lists, 32 lists (list 1-32) are used to train the DNN, and 2 lists (list 71-72) are used as the validation set and also for a practice session prior to the listening test.

We generate the stimuli at two T_{60} values of 0.6 s and 0.9 s, and two SNR levels (either 0 dB and 5 dB, or 5 dB and 10 dB). These T_{60} and SNR values are chosen so that in the more difficult SNR case, the intelligibility score of unprocessed signal is below 50% for HI listeners, and in the less difficult case, the score is higher than 50%. In this way, we can evaluate the algorithm's ability to improve speech intelligibility at different levels of difficulty. So there are 2(T_{60}s)×2(SNRs)×2(noises)×2(processed, unprocessed) = 16 conditions to test. It should be pointed out that shorter noises are used to prepare the data, with babble noise being about 5 s and SSN about 6 s. In addition, random segments are taken from the whole duration of each noise in order to generate training, validation, and test sets. These settings are different from those described in section 4, but are consistent with the evaluation methodology in [4].

Each sentence in the IEEE corpus has 5 keywords. Scoring is based on the number of keywords correctly identified by the subject. The subjects are allowed to guess and report a subset of the words in a sentence, and repeat the words verbally to the experimenter. The 16 conditions are divided into two sessions: session I with babble noise and session II with SSN. The two sessions are alternated between consecutive subjects. For each condition, we present 20 sentences, i.e., 2 lists of the IEEE corpus. Before data collection, a practice session including 20 sentences is administered to familiarize a listener with the test flow and the kind of signals to listen to. The second session will also be proceeded with the practice session if it is not conducted right after the first one.

Noise	T_{60} (s)	**SNR** (dB)	**HI-1** Δ in %	**HI-2** Δ in %	**HI-3** Δ in %	**HI-4** Δ in %	**Ave.**
Babble	0.6	0	+22.0 (4.0)				+22.0
		5	+28.0 (29.0)		-33.0 (88.0)	+15.0 (64.0)	+3.3
		10			+12.0 (80.0)	+19.0 (57.0)	+15.5
	0.9	0	+9.0 (0.0)	+18.0 (3.0)			+13.5
		5	+18.0 (15.0)	+3.0 (26.0)	+4.0 (88.0)	-10.0 (84.0)	+3.8
		10			+21.0 (38.0)	+14.0 (37.0)	+17.5
SSN	0.6	0	+6.0 (73.0)		-19.0 (78.0)	-1.0 (66.0)	-4.7
		5	-18.0 (80.0)		-1.0 (52.0)	+23.0 (29.0)	+1.3
	0.9	0	+14.0 (32.0)	+22.0 (10.0)	+17.0 (74.0)	-11.0 (79.0)	+10.5
		5	-1.0 (60.0)	+3.0 (24.0)	-1.0 (44.0)	-1.0 (20.0)	+0.0

Table 3. Speech intelligibility results of 4 HI listeners. The scores of unprocessed conditions are listed inside the parentheses.

5.2. Results

Four HI listeners with symmetric hearing loss were recruited at the Starkey Headquarters in Eden Prairie MN to participate in the listening test. HI-1 and HI-3 have mild hearing loss and the other two have moderate hearing loss. The changes of percent intelligibility scores are listed in Table 3. During the listening test, for HI-1 and HI-2, session I was tested using 0 dB and 5 dB SNRs. After this test, the SNRs were increased by 5 dB. Part of HI-2's results was discarded because his hearing aids were not taken off during the test. These are why some entries of Table 3 are left blank.

There are conditions where the speech intelligibility of processed signals is lower than that of unprocessed ones. However, more conditions result in improvement (i.e., positive numbers in Table 3). By taking average across all 4 HI listeners, in almost all

conditions we obtain some speech intelligibility improvements. The improvement is consistent for the babble noise at the SNR of 10 dB for both T_{60} values. Although the listening test is pilot in nature, the results show the promise of the proposed spectral mapping algorithm for improving the speech intelligibility of HI listeners in noisy and reverberant environments.

It is worth noting that the intelligibility benefits for HI listeners come from a new training target of the DNN mapping function, i.e., clean reverberant speech. With the mapping function in [4], we were unable to obtain any speech intelligibility improvement.

6. MASKING-BASED ALGORITHM

A recent study [25] on training targets for supervised speech separation shows that time-frequency (T-F) masking performs better than spectral mapping. Therefore, we propose a masking-based algorithm to enhance the noisy and reverberant speech and compare with the spectral mapping method described in section 3.

We define the ideal ratio mask (IRM) as follows [25],

$$IRM(t,f) = \sqrt{\frac{S^2(t,f)}{S^2(t,f)+N^2(t,f)}} \quad (4)$$

where $S^2(t,f)$ and $N^2(t,f)$ denote the energy of reverberant speech and additive noise in each T-F unit, respectively. It is worth noting that clean reverberant speech is again used as the target signal in our definition.

In our masking-based algorithm, instead of predicting the log magnitude spectrum of clean reverberant speech, the IRM is used as the training target. In the test phase, the estimated IRM is applied to the magnitude spectrogram of noisy and reverberant speech, and the enhanced signal is resynthesized by using the phase of unprocessed signal.

Although DNNs have a capacity to learn abstract features from raw inputs, well-designed features may still be helpful. Thus, in addition to using log magnitude spectrum as the input feature, we employ as the input a set of complementary features [25, 26], i.e., a combination of amplitude modulation spectrogram (AMS, 15 dimensions), relative spectral transform and perceptual linear prediction (RASTA-PLP, 13 dimensions), mel-frequency cepstral coefficients (MFCC, 31 dimensions), Gammatone filterbank power spectra (GF, 64 dimensions), and their delta (123 dimensions) and double delta (123 dimensions) components. Therefore, the feature dimension for each time frame is 369. Before training the DNN, the features are normalized to zero mean and unit variance.

Since spectral mapping is performed in the spectrogram domain, the IRM is computed in the spectrogram domain too. To simplify the descriptions, we use following notations,

- **STFT → IRM**: estimate the IRM by using the log spectral magnitude of noisy and reverberant speech as the input feature
- **CF → IRM**: estimate the IRM by using the complementary feature of noisy and reverberant speech as the input feature

We conduct the experiments on the same dataset described in section 4. The average STOI and PESQ values are listed in Table 4 and Table 5 for babble noise and SSN, respectively. Comparing with the results in Tables 1 and 2, our masking-based algorithm improves over the mapping-based algorithm in both STOI and PESQ. Since we focus on enhancing speech intelligibility, the STOI improvements for the babble noise conditions are separately shown in Fig. 2, highlighting the significant improvements obtained by the masking-based method. These results are in agreement with the denoising results in [25]. The main performance gain results from the adoption of the masking training target. The use of the complementary feature set provides some additional improvements compared with the log magnitude feature.

	STOI			PESQ		
	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
unprocessed	0.413	0.563	0.706	1.247	1.651	2.059
STFT → IRM	0.563	0.699	0.805	1.748	2.162	2.568
CF → IRM	0.569	0.711	0.813	1.791	2.208	2.601

Table 4. Average STOI and PESQ scores for STFT→IRM and CF→IRM at each SNR for babble noise.

	STOI			PESQ		
	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
unprocessed	0.436	0.588	0.730	1.285	1.619	1.986
STFT → IRM	0.613	0.739	0.828	1.854	2.225	2.589
CF → IRM	0.634	0.752	0.834	1.922	2.282	2.619

Table 5. Average STOI and PESQ scores for STFT→IRM and CF→IRM at each SNR for SSN.

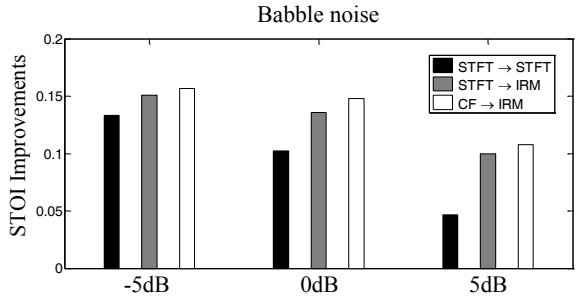

Fig. 2. STOI improvements at each SNR after removing babble noise

With better STOI and PESQ scores and informal listening impressions of the masking-based method, we predict further speech intelligibility improvements over those in Table 3 for HI listeners in noisy and reverberant environments.

7. CONCLUSION

Our objective was to improve the speech intelligibility of HI listeners in the noisy and reverberant environment. We have proposed to use spectral mapping to enhance the noisy and reverberant speech by removing noise only. A preliminary listening test has been conducted and the results have demonstrated the effectiveness of the proposed method. We have also developed a masking-based method by using reverberant target speech as the desired signal. Systematic evaluation using objective metrics indicates further improvements on both speech intelligibility and quality with the masking-based method. In future work, more formal intelligibility listening tests will be conducted to validate if the masking-based denoising approach can provide HI listeners with additional speech intelligibility improvement in noisy and reverberant environments.

8. REFERENCES

[1] E. L. J. George, S. T. Goverts, J. M. Festen, and T. Houtgast, "Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners," *Journal of Speech, Language, and Hearing Research*, vol. 53, pp. 1429–1439, 2010.

[2] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Advances in Neural Information Processing Systems*, 2000, pp. 758–764.

[3] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 1759–1763.

[4] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 982–992, 2015.

[5] X. Zhao, Y. Wang, and D. L. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 836–845, 2014.

[6] S. A. Gelfand and I. Hochberg, "Binaural and monaural speech discrimination under reverberation," *International Journal of Audiology*, vol. 15, pp. 72–84, 1976.

[7] K. S. Helfer and L. A. Wilber, "Hearing loss, aging, and speech perception in reverberation and noise," *Journal of Speech, Language, and Hearing Research*, vol. 33, pp. 149–155, 1990.

[8] A. K. Nábělek and P. K. Robinson, "Monaural and binaural speech perception in reverberation for listeners of various ages," *Journal of the Acoustical Society of America*, vol. 71, pp. 1242–1248, 1982.

[9] Y. Hu and K. Kokkinakis, "Effects of early and late reflections on intelligibility of reverberated speech by cochlear implant listeners," *Journal of the Acoustical Society of America*, vol. 135, pp. EL 22–28, 2014.

[10] N. Roman and J. Woodruff, "Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold," *Journal of the Acoustical Society of America*, vol. 133, pp. 1707–1717, 2013.

[11] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *Journal of the Acoustical Society of America*, vol. 113, pp. 3233–3244, 2003.

[12] N. Roman and J. Woodruff, "Intelligibility of reverberant noisy speech with ideal binary masking," *Journal of the Acoustical Society of America*, vol. 130, pp. 2153–2161, 2011.

[13] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 4628–4632.

[14] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, pp. 114–126, 2012.

[15] L. Wang, K. Odani, and A. Kai, "Dereverberation and denoising based on generalized spectral subtraction by multi-channel lms algorithm using a small-scale microphone array," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, pp. 1–11, 2012.

[16] E. W. Healy, S. E. Yoho, Y. Wang, and D. L. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 134, pp. 3029–3038, 2013.

[17] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[18] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.

[19] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio Electroacoust*, vol. 17, pp. 225–246, 1969.

[20] E. Habets, "Room impulse response generator," Available at `https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator`.

[21] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[22] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 625–638, 2009.

[23] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 2125–2136, 2011.

[24] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, vol. 2, pp. 749–752.

[25] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1849–1858, 2014.

[26] Y. Wang, K. Han, and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 270–279, 2013.