INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Low-Latency Single Channel Speech Dereverberation using U-Net Convolutional Neural Networks

Ahmet E. Bulut[1,2], *Kazuhito Koishida*[2]

[1]Center for Robust Speech Systems, University of Texas at Dallas, TX 75080
[2]Microsoft Corporation, One Microsoft Way, Redmond, WA 98052

ahmet.bulut@utdallas.edu, kazukoi@microsoft.com

Abstract

Speech signal reverberation due to reflections in a physical obstacle is one of the main difficulties in speech processing as well as the presence of non-stationary background noise. In this study we explore DNN-based single-channel speech dereverberation with state-of-the-art performance comparisons. We propose a CNN auto-encoder architecture with skip connections focusing on real-time and low-latency applications. The proposed system is evaluated with the REVERB challenge dataset that includes simulated and real reverberated speech samples. Our experimental results show that the proposed system has superior results on the challenge evaluation dataset as opposed to a baseline system that uses deep neural network (DNN) based weighted prediction error (WPE) algorithm. We also extend the comparison with state of the art systems in terms of most commonly used objective metrics and our system achieves better results in the most of objective metrics. Moreover a latency analysis of the proposed system is performed and trade-off between processing time and performance is examined.

Index Terms: Speech dereverberation, Speech enhancement, low-latency, U-Net, convolutional neural networks

1. Introduction

Background noise and reverberation are two of the main interferences known to considerably degrade the quality of signal that are gathered within naturalistic scenarios. Background noise might occasionally be present in a captured speech but particularly with the usage of distance microphones reverberation could be constant issue for the speech processing. Although it has been studied for many years, speech dereverberation remains to be a challenging problem especially for single channel and low-latency applications.

Earlier studies [1, 2, 3, 4] addressed the dereverberation and noise reduction problem with several algorithms. However since they individually used different evaluation data it was difficult to make fair comparison between the systems. The REVERB challenge [5] put together a common dataset for the evaluation of the dereverberation algorithms which were developed community-wide. A study based on the modification of the direct-to-reverberant ratio (DRR) was proposed in [6] and applied successfully to the single channel scenario. In another study [7] used a transformation in autocorrelation domain, called zero phase procedure, in order to detect and remove the non-periodic corruption. Wisdom et al. [8] applied the short-time fan-chirp transform (STFChT) to extend the length of the short-time Fourier transform (STFT) analysis window to achieve an overcomplete time-frequency representation. Whereas, with the clear success of DNN based methods in recent years, [9] and [10] proposed a learning based approach and used deep neural networks (DNN) to find a transformation from reverberant spectrogram to the corresponding clean spectrogram.

Weninger et al. [11] employed deep bidirectional Long Short-Term Memory (biLSTM) de-noising auto-encoders (DAE) to achieve blind feature space dereverberation. In a recent study [12] wide residual network (WRN) architecture was employed for reducing the reverberation/noise effect and achieved promising results. Inspired by the recent success of convolutional neural networks (CNN) based U-Net architecture in many image processing applications, Ernst et al. [13] presented two variations of such networks: one of them has encoder-decoder network with skip connections and a generative adversarial network (GAN) with U-Net as generator. By applying CNN based architecture they succeeded to preserve global and local information in the reconstruction successfully and achieved superior results against the competing methods.

Following the success of U-Net based architecture and demonstrated success of such architecture in our previous speech enhancement study [14], we adopt similar network structure for the dereverberation task as well. However along with the aim of fulfilling the low-latency requirement an architecture is designed to maintain the number of trainable parameters limited. To perform in-dept quality and processing time performance comparison we reimplement the study [15] which is based on inverse filter estimation method called weighted prediction error (WPE) algorithm with a DNN-based spectrum estimator to make the conventional WPE algorithm successfully work for very short observed data. This choice makes the study a very good candidate for bench marking our proposed low-latency dereverberation system. In this paper, we propose a simple and effective CNN based U-Net architecture which operates on T-F domain and try to have relatively low processing window in the temporal dimension and construct an architecture tailored for the corresponding input. We choose to concentrate on magnitude prediction by disregarding phase information. Our proposed network includes encoder and decoder layers which fulfil the downsampling and upsampling of the input data respectively. Log-spectral distance (LSD) metric is used as the loss function of the training. We analyse the objective speech quality of the systems and further investigate the processing time of baseline systems with respect to the proposed system at inference time. Moreover we investigate the performance of the proposed system under various latency conditions.

This paper is organized as follows: We first describe the overview of the system and the details of the proposed network architecture in Section 2. And in Section 3 experimental details are explained as well as the information about the dataset and

The work was completed when AEB was an intern at Microsoft Corporation.

http://dx.doi.org/10.21437/Interspeech.2020-2421

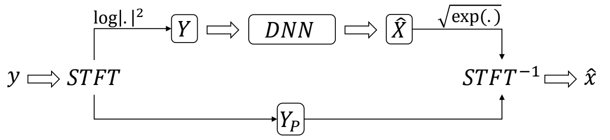

Figure 1: *System overview. y is the input reverberated signal and Y is LPS of the signal and $\hat{X}$ is estimated LPS which combined with reverberated phase Y_P to get estimated signal $\hat{x}$.*

baseline methods. We present the results and some analysis in Section 4 and Section 5 concludes the paper.

2. Proposed System

2.1. System Overview

The mapping function from reverberated to clean feature space can be learned by means of a DNN which is trained on a dataset of parallel reverberated and clean speech files. As an input to the network we use log-power spectrogram (LPS), Y which can be obtained after applying a short-time Fourier transform (STFT) to the input noisy waveform, y to get magnitude Y_M and phase Y_P and then calculated as $Y = \log(|Y_M|^2)$.

Within the scope of this study, at inference time, we only forward propagate the Y features through the network and reconstruct the enhanced signal $\hat{x}$ by applying the inverse STFT with estimated signal magnitude, $\hat{X}_M$ and reverberated signal phase, Y_P as shown in Figure 1.

2.2. Network Architecture

The proposed network architecture is illustrated in Figure 2. To encode the input features we deploy 12 2-dimensional convolution (conv2d) layers named as (e1-e12) and to decode we apply 8 2-dimensional sub-pixel convolution (subconv2d) layers named as (d1-d8). The sub-pixel convolution layers are successfully applied to speech super-resolution and speech enhancement tasks [16, 14]. The main idea is to compute more feature channels on the convolution layer and resize them into the target upsample dimension. To each layer we apply leaky rectified linear unit (LReLU) activation function followed by batch normalization. Downsampling is applied only to spectral dimension firstly and then it is applied to temporal dimension. The detailed stride size, kernel size, and number of channel for each layer are described on Table 1. For the decoder layers (d1-d11) we apply skip-connections with corresponding encoder layers in reverse order which are (e11-e1), respectively. Moreover we apply dropout to the first 4 layers (d1-d4) with a probability rate of 0.5.

For the training loss, we experiment three types of functions, namely L_1, L_2 norms and log-spectral distance (LSD). Our overall testing shows that LSD yields slightly better results for dereverb tasks. In general terms, LSD measures the distance between two spectrograms in decibels, and it is defined as follows:

$$loss_{\mathrm{LSD}} = \frac{1}{T}\sum_{i=1}^{T}\sqrt{\frac{1}{S}\sum_{j=1}^{S}[X(i,j) - \hat{X}(i,j)]^2} \qquad (1)$$

where X and $\hat{X}$ are the clean and estimated LPS, respectively and T is the number of frames and S is the number of spectral bins.

Table 1: *Detailed configuration of the proposed network architecture for each layer.*

Layer	Kernel No.	Kernel Size	Stride	Output Shape
Input	-	-	-	(16, 256, 1)
e1	64	(5, 7)	(1, 2)	(16, 128, 64)
e2	128	(3, 5)	(1, 2)	(16, 64, 128)
e3	128	(3, 3)	(1, 2)	(16, 32, 128)
e4	128	(3, 3)	(1, 2)	(16, 16, 128)
e5	128	(3, 3)	(1, 2)	(16, 8, 128)
e6	128	(3, 3)	(1, 2)	(16, 4, 128)
e7	128	(3, 3)	(1, 2)	(16, 2, 128)
e8	128	(3, 1)	(1, 2)	(16, 1, 128)
e9	256	(3, 1)	(2, 1)	(8, 1, 256)
e10	256	(3, 1)	(2, 1)	(4, 1, 256)
e11	256	(3, 1)	(2, 1)	(2, 1, 256)
e12	256	(1, 1)	(2, 1)	(1, 1, 256)
d1 + e11	256	(1, 1)	(1, 1)	(2, 1, 512)
d2 + e10	256	(1, 1)	(1, 1)	(4, 1, 512)
d3 + e9	256	(1, 1)	(3, 1)	(8, 1, 512)
d4 + e8	128	(1, 1)	(3, 1)	(16, 1, 256)
d5 + e7	128	(1, 1)	(3, 1)	(16, 2, 256)
d6 + e6	128	(1, 1)	(3, 1)	(16, 4, 256)
d7 + e5	128	(1, 1)	(3, 3)	(16, 8, 256)
d8 + e4	128	(1, 1)	(3, 3)	(16, 16, 256)
d9 + e3	128	(1, 1)	(3, 3)	(16, 32, 256)
d10 + e2	128	(1, 1)	(3, 3)	(16, 64, 256)
d11 + e1	64	(1, 1)	(3, 3)	(16, 128, 128)
d12	1	(1, 1)	(3, 5)	(16, 256, 1)

3. Experiments

3.1. Dataset

The dataset that we use in this study is provided by the REVERB challenge [5] organizers. It consists of simulated and real data. The simulated test data is created by mixing the speech data from WSJCAM0 corpus [17] with three Room Impulse Renspnse (RIR) whose reverberation time (RT_{60}) is measured as $0.25s$, $0.5s$, and $0.7s$ at two source-microphone distances: far (200 cm) and near (50 cm). Moreover a stationary of noise added to the mixtures with a SNR of $20dB$. The simulated training data is created by using 24 RIRs whose RT_{60} ranges from $0.2s$ to $0.8s$. The real data is from MC-WSJ-AV corpus [18] which is captured in a reverberant meeting room with a RT_{60} of $0.7s$ at two source-microphone distances: far (250 cm) and near (100 cm).

3.2. Preprocessing and Training Setup

The spectral representation is obtained by applying 512-point STFT with a Hanning window of size 512 and a hop size of 256 to the audio files that are sampled at 16kHz. Only the 257-point STFT magnitudes are considered by removing the symmetric half. We remove the last STFT point as well which yields a power-of-2 input dimension. In order to achieve a fixed dimension for the processing of both train and test sets, we use 16 frames of clips which in turn creates an input dimension of 16

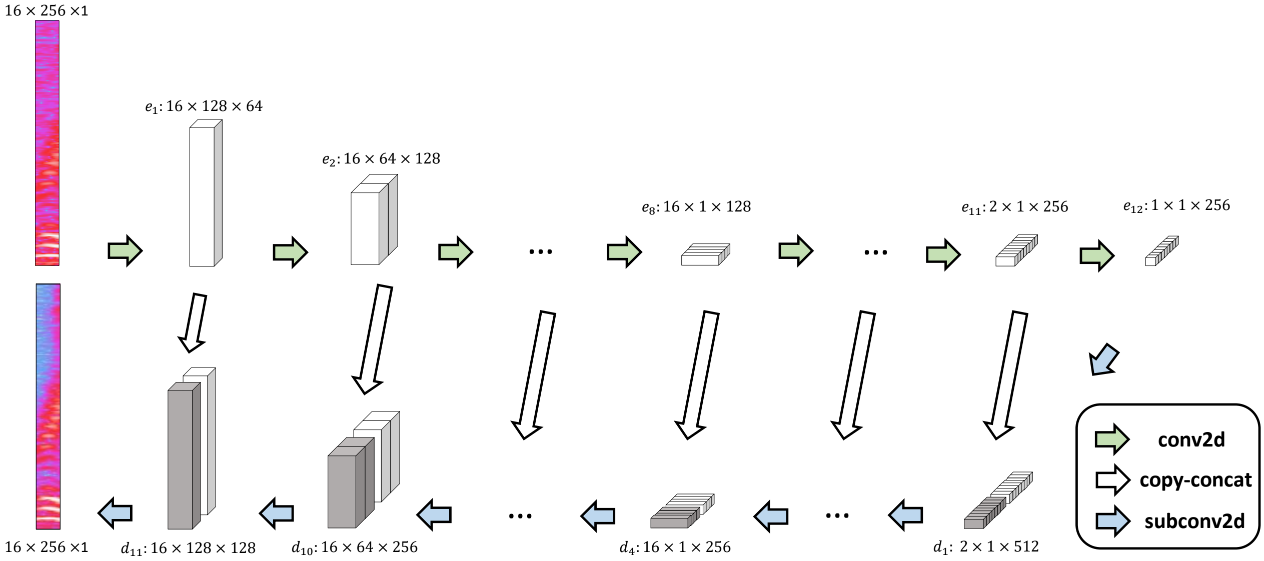

Figure 2: *Proposed network architecture.*

× 256 × 1 processing window. We choose relatively small, 16 frame-window (0.256 sec), on the temporal domain in order to meet our low-latency constraint. The input to the network are normalized to have zero mean and unit variance.

The network is trained with the Adam optimizer [19] with a batch size of 64 and learning rate of 0.0001 for 50 epochs. The decay rates of optimizer are $\beta_1 = 0.5$ and $\beta_2 = 0.9$. The weights of the network are initialized from the normal distribution with zero mean and 0.02 standard deviation [14].

3.3. Baseline Methods and Evaluation Metric

We compare our proposed approach with a system that utilizes DNN-based spectrum estimation to construct linear inverse filters by using WPE [15]. The baseline system that we refer as DNN-WPE, is trained with the dataset described in Section 3.1 by following the corresponding experimental configuration that they proposed.

In conventional WPE the filter that is used to construct desired signal from the reverberated signal, can be effectively estimated in the maximum likelihood sense. It assumes that the desired signal has a zero-mean complex Gaussian distribution with a time varying variance called power spectral density (PSD) which is an unknown parameter to be estimated. However it is addressed that [15] if the duration of the data limited the error in PSD estimation degrades significantly. To overcome the problem, rather than relying on the iterative optimization procedure to estimate the PSD, a DNN is utilized hence successful signal estimation is achieved.

The spectral representation is obtained within a Hanning window of size 512 and a hop size of 128 to the audio files that are sampled at 16kHz.

As for the DNN architecture, unidirectional Long Short Term Memory (LSTM) is used for the first layer which is followed by two fully-connected layers with ReLU activations. The number of memory cells in the LSTM is 500, and the number of nodes in the fully-connected layers is 2048. The network is trained by standard stochastic gradient decent (SGD) algorithm using the MMSE cost function. The input and output features of the network were log amplitude spectra and the estimated output values are used to reconstruct the dereverberated signal after the inference. The processing window length is 11 frames which is generated by combining the features of 5 left and 5 right context frames of the current frame.

Several metrics are used for the objective evaluation of the dereverberation systems. For the scope of this study we use; cepstral distance (CD) [20], log likelihood ratio (LLR) [20], frequency-weighted segmental SNR [20], Speech -to-reverberation modulation energy ratio (SRMR) [21] metrics. For real evaluation data, only the non-intrusive SRMR metric is used because only SRMR metric does not require ground truth of the target speech. For CD and LLR lower values indicate better performance whereas for FWSegSNR and SRMR higher values are expected ideally. In order to evaluate latency and processing time of the systems, the length of the processing window and real-time factor (RTF) are used respectively. We define the latency (L) as the summation of the shift length of the processing window (W) and the duration of time needed to process it. RTF is the most common speed performance metric for speech processing applications and defined as the ratio of processing time of the input over the actual duration of the input. Any application is considered real-time if its RTF is less than 1. The computer that is used for all latency and RTF calculations has an Intel Xeon Gold 6130 CPU @ 2.10GHz and a GeForce 2080 RTX Ti, 24GB GPU.

4. Results and Discussion

We first analyze the objective speech quality and processing time for the baseline and proposed system on the simulated evaluation dataset as shown in Table. 2. For the proposed method, the results of different size of overlap are presented. That is to say, Proposed$\{1, 2, 4, 8, 16\}$ stands for the systems that using $\{1, 2, 4, 8, 16\}$ frames shifted-processing. We also include some results from the state-of-the-art papers that use the same training and evaluation of REVERB challenge dataset. The baseline DNN-WPE system is worse than the proposed systems

Table 2: *Performance comparison of the systems on simulated evaluation set. Values are averaged out for room types (1, 2, 3) and microphone positions (near, far).* ***W:*** *Window Length (ms),* ***L:*** *Latency (ms),* ***RTF:*** *Real-time Factor. Proposed{1, 2, 4, 8, 16} stands for the systems that using {1, 2, 4, 8, 16} frames shifted-processing.*

	CD	**LLR**	**FWSegSNR**	**SRMR**	**W/L (ms)**	**RTF**
Unprocessed	3.97	0.58	3.62	3.68	-	-
Cauchi et al. [6]	3.55	0.59	6.09	4.29	-	-
Gonzalez et al. [7]	4.38	0.43	4.39	5.09	-	-
Wisdom et al. [8]	3.57	0.57	7.07	4.55	-	-
Xiao et al. [9]	2.50	0.50	7.55	**5.77**	110/-	-
Ribas et al. [12]	3.59	0.47	4.80	3.59	2000/-	-
Ernst et al. (aU-Net+GAN) [13]	2.50	0.41	10.79	4.88	2048/-	-
DNN-WPE	2.92	0.39	7.53	4.68	**88**/89	**0.01**
Proposed16	2.50	0.33	11.62	5.09	256/271	0.06
Proposed8	**2.47**	**0.32**	**11.72**	5.11	256/143	0.12
Proposed4	2.49	0.33	11.66	5.11	256/79	0.23
Proposed2	2.52	0.33	11.53	5.08	256/47	0.47
Proposed1	2.59	0.34	11.25	5.05	256/**31**	0.93

Table 3: *Performance comparison of the systems on real evaluation dataset. Values are averaged out for microphone positions (near, far).*

	SRMR
Unprocessed	3.18
Cauchi et al. [6]	4.82
Gonzalez et al. [7]	4.70
Wisdom et al. [8]	4.89
Xiao et al. [9]	4.36
Ribas et al. [12]	3.24
Ernst et al. (aU-Net+GAN) [13]	**5.58**
DNN-WPE	4.97
Proposed16	5.55
Proposed8	5.37
Proposed4	5.32
Proposed2	5.26
Proposed1	5.06

in terms of speech quality measures but has the smallest processing window and the best RTF score which makes it a good candidate for the both offline and online systems. The first system, Proposed16 that uses 16 frames (256 ms) block-by-block processing and at the following rows we keep reducing the shift size and lower the latency concurrently while observing RTF of the systems. Although the proposed network always processes 16 frames (256 ms) of the full input window and produces the corresponding prediction, only the last shift-size portion of the prediction is actually used for the output. Note that, in the initial few frames where there is not enough input data available to reach the input buffer size, it is filled by repeating the very first shift-size portion of the input. It can be observed that our proposed system is able to operate at real-time with a very low latency duration as little as 31 ms and with a tolerable degradation on speech quality. As it can be seen clearly, the "Proposed8" system has the best speech quality measures in terms of CD, LLR, and FWSegSNR however system [9] has the best result in terms of SRMR. The system [13] has comparable results in terms of speech quality measures but has 8 times as large processing window as of the proposed systems.

By applying shifted-processing window we compromise the RTF value as compared to the baseline system but we achieve relatively good performance in terms of speech quality and intelligibility with the proposed simple but effective DNN architecture. Moreover, low latency and acceptable RTF value make the proposed system a good candidate for the low-latency required systems. We have to note that for the DNN-WPE and the proposed systems both operate on the spectral domain and the latency calculation includes the sum of the duration of spectral featurization and reconstruction.

In Table 3, you can see speech quality performance comparison of the baseline and proposed systems as well as the state-of-the-art systems on real evaluation data. Since the ground truth is not available for the real data only SRMR metrics are reported. We have to note that all the proposed systems outperform the baseline DNN-WPE system and all of state-of-the-art systems except for the system [13], while the gap against [13] is very small. Proposed16 outperforms the first four state-of-the-art systems and comparable to Ernst's. Whereas, given the significant difference of processing windows between these systems, this comparable performance result is tolerable in the expense of decreased latency. Moreover, we have to note that the proposed system has higher degradation with smaller latency on the real data compared to the simulated data.

5. Conclusion

In this paper we propose a simple but effective U-Net CNN architecture specifically for the dereverberation systems working under low-latency condition. In accordance with this purpose we try to keep the proposed architecture moderate and choose to operate on spectral domain for faster processing. We achieve superior results as opposed to the a baseline system using DNN assisted WPE. And it has been shown that the proposed system has a real-time operation under extreme low-latency conditions while maintaining performance quality of the system better in the most and comparable for the rest of the performance metrics to the baseline and state-of-the-art dereverberation systems.

6. References

[1] S. Doclo and M. Moonen, "Combined frequency-domain dereverberation and noise reduction technique for multi-microphone speech enhancement," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2001, pp. 31–34.

[2] E. A. Habets, S. Gannot, I. Cohen, and P. C. Sommen, "Joint dereverberation and residual echo suppression of speech signals in noisy environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1433–1451, 2008.

[3] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 231–246, 2009.

[4] H. W. Lollmann and P. Vary, "A blind speech enhancement algorithm for the suppression of late reverberation and noise," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 3989–3992.

[5] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[6] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukic, T. Gerkmann, S. Doclo, and S. Goetze, "Joint dereverberation and noise reduction using beamforming and a single-channel speech enhancement scheme," in *Proc. REVERB challenge workshop*, vol. 1, 2014, pp. 1–8.

[7] D. R. González, S. C. Arias, and J. R. C. de Lara, "Single channel speech enhancement based on zero phase transformation in reverberated environments," *the Proceedings of REVERB Challenge*, 2014.

[8] S. Wisdom, T. Powers, L. Atlas, and J. Pitton, "Enhancement and recognition of reverberant and noisy speech by extending its coherence," *arXiv preprint arXiv:1509.00533*, 2015.

[9] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "The ntu-adsc systems for reverberation challenge 2014," in *Proc. REVERB challenge workshop*, 2014, p. o2.

[10] K. Han, Y. Wang, D. 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, no. 6, pp. 982–992, 2015.

[11] F. Weninger, S. Watanabe, Y. Tachioka, and B. Schuller, "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4623–4627.

[12] D. Ribas, J. Llombart, A. Miguel, and L. Vicente, "Deep speech enhancement for reverberated and noisy signals using wide residual networks," *arXiv preprint arXiv:1901.00660*, 2019.

[13] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 390–394.

[14] A. E. Bulut and K. Koishida, "Low-latency single channel speech enhancement using u-net convolutional neural networks," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[15] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online wpe dereverberation." in *Interspeech*, 2017, pp. 384–388.

[16] S. E. Eskimez, K. Koishida, and Z. Duan, "Adversarial training for speech super-resolution," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 347–358, 2019.

[17] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1995, pp. 81–84.

[18] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multichannel wall street journal audio visual corpus (mc-wsj-av): Specification and initial experiments," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* IEEE, 2005, pp. 357–362.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

[21] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.