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

Virtual acoustic channel expansion based on neural networks for weighted prediction error-based speech dereverberation

Joon-Young Yang and Joon-Hyuk Chang

Department of Electronics and Computer Engineering
Hanyang University, Seoul, Republic of Korea
dreadbird06@gmail.com, jchang@hanyang.ac.kr

Abstract

In this study, we propose a neural-network-based virtual acoustic channel expansion (VACE) framework for weighted prediction error (WPE)-based speech dereverberation. Specifically, for the situation in which only a single microphone observation is available, we aim to build a neural network capable of generating a virtual signal that can be exploited as the secondary input for the dual-channel WPE algorithm, thus making its dereverberation performance superior to the single-channel WPE. To implement the VACE-WPE, the neural network for the VACE is initialized and integrated to the pre-trained neural WPE algorithm. The entire system is then trained in a supervised manner to output a dereverberated signal that is close to the oracle early arriving speech. Experimental results show that the proposed VACE-WPE method outperforms the single-channel WPE in a real room impulse response shortening task.

Index Terms: speech dereverberation, weighted prediction error, neural network, multi-channel linear prediction

1. Introduction

Speech dereverberation aims to remove or suppress the late reverberation component in a speech signal captured by distant microphones in a reverberant enclosure. Among the various techniques employed, the weighted prediction error (WPE) [1] algorithm, which blindly estimates a reverberation filter in an iterative manner, has been widely used as the front-end to improve the robustness of speech-triggered applications [2, 3]. In general, when multi-channel speech signals are available, the multi-input multi-output (MIMO) WPE will be superior to its single-channel counterpart, as it can benefit from the underlying multi-channel linear prediction (MCLP) algorithm to exploit the different diffuse patterns of the late reverberation observed through multiple microphones. However, for small electronic devices, installing more than one microphone may not be feasible owing to the extra cost of expanding the microphone channels.

In this context, to utilize the MIMO WPE algorithm to dereverberate a single-channel observation effectively, we propose a neural-network-based virtual acoustic channel expansion (VACE) framework. The main concept of the proposed VACE technique is the generation of a virtual signal that can assist the dual-channel WPE to output an actual signal "dryer" than that of the single-channel WPE from a given observation. To this end, we first pre-train the constituent networks for the neural WPE [2] and VACE to build the dual-channel VACE-WPE system, and subsequently fine-tune the neural network for the VACE to dereverberate the observed single-channel signal. The proposed VACE-WPE method is compared with the single-channel and actual dual-channel WPE algorithms for a real room impulse response (RIR) shortening task, and the results are evaluated in terms of various objective speech quality metrics.

2. Related work

A series of studies conceptually relevant to the proposed framework can be found in [4, 5, 6]. In [4], the so-called *virtual microphone* technique was proposed, whereby multiple sets of the amplitude and phase of a virtual signal in the short-time Fourier transform (STFT) domain are generated through the complex logarithmic interpolation of dual-microphone observations. In [5], the amplitude interpolation method was improved by minimizing the β-divergence between the virtual and actual microphone signal amplitudes, which was further extended by exploiting a convolutional neural network as the amplitude estimator [6]. The virtual microphone technique was shown to be effective for providing informative auxiliaries to some types of beamformers for speech enhancement [4, 5, 6]. Nonetheless, in this work, we investigate a neural-network-based single-to-dual acoustic channel expansion for WPE-based speech dereverberation without any constraints regarding the microphone array geometry. Accordingly, we use the term *virtual acoustic channel* rather than *virtual microphone*, as the generated virtual signal is not constrained to a specific microphone arrangement.

3. System overview

3.1. Signal model

In this study, we only consider the scenario in which a speech signal is captured by a single microphone in a noiseless reverberant enclosure. However, as our aim is to generate a virtual secondary signal to compose a dual-channel input for the MIMO WPE algorithm, we present a signal model for the multi-microphone scenario as an extension of the single-microphone case. In the STFT domain, the signal model is represented as follows:

$$\mathbf{X}_{t,f} = \mathbf{X}_{t,f}^{(\text{early})} + \mathbf{X}_{t,f}^{(\text{late})}, \tag{1}$$

where $\mathbf{X}_{t,f}$ is the D-channel stack of the microphone observations, and $\mathbf{X}_{t,f}^{(\text{early})}$ and $\mathbf{X}_{t,f}^{(\text{late})}$ are the early arriving speech and the late reverberation, respectively. Herein, the former is assumed to be obtained upon convolution of the source speech with the RIR truncated up to the point 50 ms after the main peak, while the remaining part is responsible for the latter.

3.2. Neural WPE algorithm

In a vein similar to that described in Section 3.1, we describe the MIMO WPE algorithm, whose single-channel version can easily be obtained by setting the number of microphones, D, to 1. The classical WPE uses the MCLP technique to estimate the late reverberation component, $\mathbf{X}_{t,f}^{(\text{late})}$, and cancels it out

http://dx.doi.org/10.21437/Interspeech.2020-1553

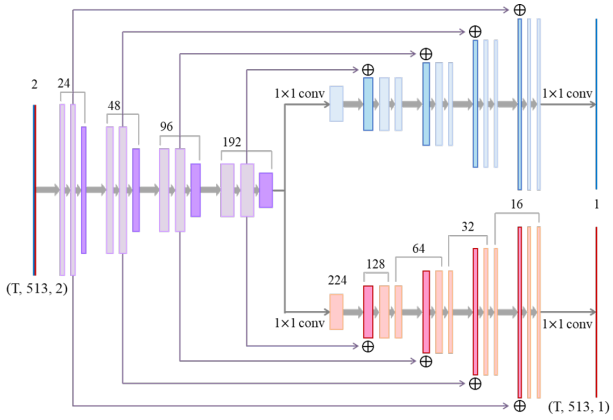

Figure 1: *Architecture of the VACE network.*

from the observation; the dereverberated output is assumed to be sampled from a zero-mean complex Gaussian distribution with time-varying variance [1]. The linear prediction (LP) filter coefficients are calculated to obtain the maximum likelihood estimate of the early arriving speech, $\mathbf{X}_{t,f}^{(\text{early})}$, following the iterative procedure described below:

$$\text{Step 1)} \quad \lambda_{t,f} = \frac{1}{D}\sum_{d}|Z_{d,t,f}|^2, \tag{2}$$

$$\text{Step 2)} \quad \mathbf{R}_f = \sum_{t}\frac{\tilde{\mathbf{X}}_{t-\Delta,f}\tilde{\mathbf{X}}_{t-\Delta,f}^{\mathrm{H}}}{\lambda_{t,f}} \in \mathbb{C}^{DK\times DK}, \tag{3}$$

$$\mathbf{P}_f = \sum_{t}\frac{\tilde{\mathbf{X}}_{t-\Delta,f}\mathbf{X}_{t,f}^{\mathrm{H}}}{\lambda_{t,f}} \in \mathbb{C}^{DK\times D}, \tag{4}$$

$$\mathbf{G}_f = \mathbf{R}_f^{-1}\mathbf{P}_f \in \mathbb{C}^{DK\times D}, \tag{5}$$

$$\text{Step 3)} \quad \mathbf{Z}_{t,f} = \mathbf{X}_{t,f} - \mathbf{G}_f^{\mathrm{H}}\tilde{\mathbf{X}}_{t-\Delta,f}, \tag{6}$$

where $\mathbf{Z}_{t,f}$ is the estimated early arriving speech, d is the microphone channel index, $\lambda_{t,f}$ is the average power spectral density (PSD) of $\mathbf{Z}_{t,f}$, K is the order of the LP filter, Δ is a delay for the LP, and $\tilde{\mathbf{X}}_{t-\Delta,f}$ and $\mathbf{G}_f$ are the stacked representations (from Δ-th to $(\Delta+K-1)$-th past time frames) of the observation and the filter coefficients, respectively.

Unlike the classical WPE, the neural WPE [2] employs a pre-trained neural network to estimate the PSD of the early arriving speech in a channel-independent manner, which allows for iteration-free calculation of the LP filter. Specifically, in this study, we train the neural network to estimate the log-scale power spectra (LPS) of $X_{d,t,f}^{(\text{early})}$ given the LPS of $X_{d,t,f}$ by minimizing the mean squared error (MSE) between the estimated and the oracle early arriving speech. Note that we only consider the offline processing scenario of a full utterance using the batch-mode WPE.

3.3. Neural network for the VACE

3.3.1. Network architecture

Inspired by recent studies on phase-aware speech enhancement [7, 8, 9], we opt to use the real and imaginary (RI) components of the STFT coefficients as the input and output representations of the VACE network (VACENet). Our choice for the network architecture involves a convolutional encoder–decoder structure similar to the U-Net [10] but with a few modifications as listed below:

- we use convolutions with the stride of 2 for downsampling instead of max-pooling (MaxPool).
- gated linear units (GLU) [11] are used instead of simple convolutions, except for those for downsampling and upsampling.
- we use 1×1 convolutions in the bottleneck of the network.
- for the expansive path, we use separate decoder streams for estimating the RI components, as advised in [7].

The rest of the structure is the same as that of the U-Net, including the number of downsampling and upsampling operations and positions of the concatenations between the encoder and the decoder feature maps. A sketch of the architecture is depicted in Fig. 1, where the rectangles denote the feature maps, whose height and width represent their relative size and depth, respectively. Each of the wide arrows represents a 2D convolution (Conv2D) with a kernel size of 3, and $\oplus$ denotes the concatenation of the feature maps along the depth axis.

3.3.2. Loss function

In a spirit similar to that of the multi-metrics learning approach [8], we define the loss function for training the VACENet as a combination of the losses computed in various domains of the signal representations. To be more specific, the loss function comprises a frequency-domain loss and a time-domain loss, each of which is defined as follows:

$$\begin{aligned} L_1^{\text{freq}}(A,B) = {}& \text{MSE}(A^r,B^r) + \text{MSE}(A^i,B^i) \\ & + \alpha\cdot\text{MSE}(\ln|A|,\ln|B|), \end{aligned} \tag{7}$$

$$L_1^{\text{time}}(a,b) = \text{MAE}(a,b), \tag{8}$$

$$L_1(A,B) = L_1^{\text{freq}}(A,B) + \beta\cdot L_1^{\text{time}}(a,b), \tag{9}$$

where A and B are the STFT-domain representations, $\ln|A|$ and $\ln|B|$ are the log-scale magnitudes, a and b are the time-domain signals obtained by taking the inverse STFT of A and B, respectively, the superscripts r and i denote the RI components, respectively, α and β are scaling factors to control the scales between the losses defined in different domains of the signal representations, and $\text{MAE}(\cdot,\cdot)$ computes the mean absolute error between the inputs.

3.3.3. Pre-training

Before the VACENet is trained to generate a valid virtual signal that can facilitate dereverberation of the observed single-channel speech via the MIMO WPE, it is necessary to set an appropriate initial point of the network to avoid divergence. In this study, we simply choose to pre-train the network by performing a self-regression task, under the assumption that the actual dual-channel speech recordings may not deviate significantly from each other. Given the single-channel observation, X_1, the output of the VACENet, X_v, can be represented as follows:

$$X_v = \mathfrak{F}(X_1;\Theta), \tag{10}$$

where $\mathfrak{F}(\cdot;\Theta)$ denotes the VACENet parameterized by Θ. In the pre-training stage, the neural network is trained to minimize $L_1(X_v,X_1)$.

3.3.4. Fine-tuning

After the initialization is completed, the VACENet is integrated to the pre-trained neural WPE algorithm, as depicted in

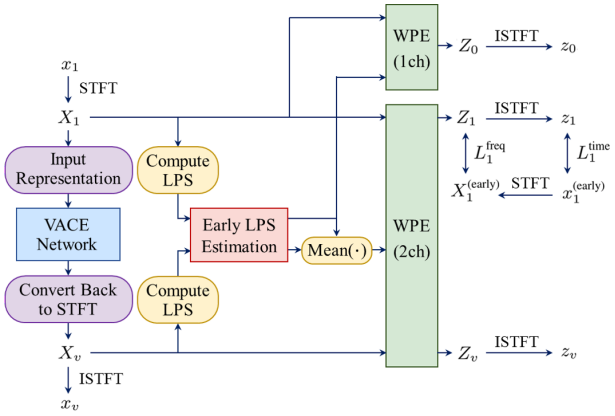

Figure 2: *Block diagram of the proposed VACE-WPE system.*

Fig. 2. The virtual signal, X_v, is directly introduced to the dual-channel WPE with the observed signal, X_1, both of which are also passed through the early speech LPS estimation model (LPSNet). Note that the parameters of the LPSNet are not subject to training and are hence frozen in the fine-tuning stage. The outputs of the LPSNet are converted back to a linear scale, averaged for each time-frequency unit, and used as the PSD estimate required in (2). Finally, the entire system is trained to output the dereverberated signal close to the oracle early arriving speech via minimization of the following loss function:

$$L = L_1(Z_1, X_1^{(\text{early})}), \tag{11}$$

where Z_1 denotes the output signal of the dual-channel VACE-WPE corresponding to the actual input signal, X_1. Notice that we also depicted the single-channel neural WPE method in Fig. 2, which is the target algorithm to be compared with the proposed dual-channel VACE-WPE method.

4. Experimental setup

4.1. Dataset

All the experiments were conducted on the TIMIT [12] corpus, where we excluded the common-transcript utterances from the whole dataset in advance. For the training, we also removed the utterances with durations of less than 2.8 s and cut out a small portion for validation purposes. Consequently, we obtained a training set of 3,023 utterances from 462 speakers, a validation set of 458 utterances, and a test set comprising 1,344 utterances from 168 speakers.

To prepare the reverberated speech for training the VACE-WPE system, we used the simulated RIR dataset of [13], which is widely used for data augmentation in Kaldi's speech and speaker recognition recipes [14]. Among the small, medium, and large room RIRs, we excluded the small room RIRs, and selected 16,200 medium room and 5,400 large room RIRs for training; 1,800 medium room and 600 large room RIRs were selected for validation. For the evaluation, we used real RIRs taken from the REVERB Challenge 2014 [15] dataset. As described in [15], the dataset contains eight different RIRs for each of the small, medium, and large room environments, whose reverberation times (T60) are about 0.25, 0.5, and 0.7 s, respectively. Three different test speech datasets were prepared for each of the room conditions with different reverberation levels, and only the first of the eight microphone channels was used for data preparation.

4.2. Model specifications

The speech signals sampled at 16 kHz were converted to the STFT domain using the 64 ms Hann window with a hop size of 16 ms. Accordingly, the 513-dimensional LPS and the stack of 513-dimensional RI components were used as the input features for the LPSNet and VACENet, respectively.

For the LPSNet architecture, we adopted the dilated convolutional network proposed in [16]. This consists of a series of Conv2D and MaxPool operations, followed by a stack of dilated 1D convolution (Conv1D) blocks, and finally a fully-connected output layer. We slightly modified the original model by reducing the kernel size of the Conv2D from (9, 9) to (5, 5) and the number of feature maps from (32, 64) to (24, 48), while the number of dilated Conv1D blocks was increased from 2 to 4. The input LPS features were normalized using a trainable batch normalization (BN) [17] layer. For more details regarding the network architecture, please refer to [16].

For the VACENet, we also applied the BN at the input layer separately to each attribute of the RI components, whereas each of the output RI components were denormalized using the global mean and variance statistics precomputed from a number of RI components of the reverberated speech signals. Meanwhile, in a complete VACE-WPE system, we set the LP parameters of the batch-mode WPE to $(\Delta, K) = (3, 20)$.

4.3. Training

We used an on-the-fly data generator for the mini-batch composition. Specifically, a speech utterance was randomly selected from the entire training set and cropped to a 2.8-s-long excerpt, convolved with a randomly chosen RIR, and then converted to STFT coefficients; four of such excerpts were gathered to compose a single mini-batch. A single training epoch was defined as the iterations over 6,000 mini-batches.

We used $\alpha = 0.3$ and $\beta = 20$ in (7) and (9), whose values were determined by monitoring the first few iterations of the training. All the networks were trained using the Adam optimizer [18]. The initial learning rates for both the LPSNet and the VACENet were set to 10^{-4} and $5 \cdot 10^{-5}$ in the pre-training and fine-tuning stages, respectively, and annealed by half whenever the validation loss did not improve for two consecutive epochs. In addition, dropout [19] and gradient clipping [20] were critical for regularizing and stabilizing the training, where we set the dropout rate to 0.3 and the global norm threshold to 3.0. The weights of the networks were also $l2$-regularized with the scale of 10^{-5}.

Finally, the LP order was set to 10 instead of 20 during the training, which we empirically found to be more effective to train the VACE-WPE system.

4.4. Evaluation

The proposed system was compared with both the single-channel WPE and the dual-channel WPE employing the actual second-channel speech signal. Herein, the actual second-channel signal was obtained by convolving the source speech with the test set RIRs taken from the fifth of the eight microphone channels, which is located at the side opposite to the first microphone [15]. For the single-channel WPE, we set the LP order to 60. The dereverberation performance was evaluated in terms of the perceptual evaluation of speech quality (PESQ) [21], cepstrum distance (CD), log-likelihood ratio (LLR) [22], and non-intrusive signal-to-reverberation modulation energy ratio (SRMR) [23]. Note that because the objective of the WPE

Table 1: *Performance of the dereverberation systems under the different room environments.*

Model	Small				Medium				Large			
	PESQ	CD	LLR	SRMR	PESQ	CD	LLR	SRMR	PESQ	CD	LLR	SRMR
Unprocessed (x_1)	3.34	0.56	0.03	3.27	2.05	1.84	0.17	2.63	1.75	2.37	0.25	2.39
Single-channel WPE (z_0)	3.86	1.54	0.09	3.39	3.07	1.63	0.12	3.13	2.71	1.71	0.14	2.99
VACE-WPE (z_1)	3.93	1.22	0.06	3.39	3.19	1.41	0.11	3.27	2.87	1.53	0.13	3.30
VACE-WPE (z_v)	1.11	4.65	0.93	2.36	1.20	4.08	0.73	2.31	1.20	4.16	0.72	2.11
Dual-channel WPE (actual)	3.86	1.05	0.05	3.38	3.31	1.20	0.08	3.28	2.96	1.32	0.10	3.11

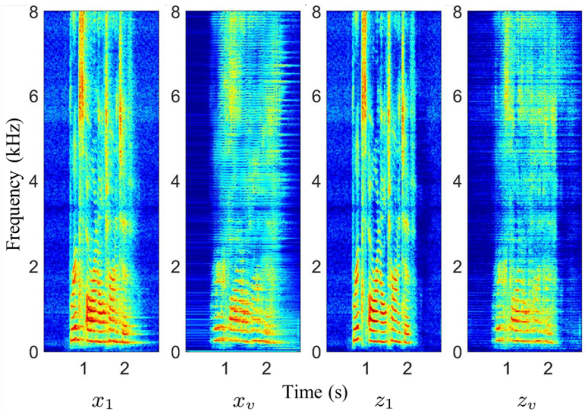

Figure 3: *Sample spectrograms in the large room condition.*

algorithm is to obtain the early arriving speech, we used $x_1^{(\text{early})}$ as the reference signal for the computation of the first three metrics.

5. Results and analysis

Table 1 summarizes the performance of the various dereverberation systems. In the second row, the output signal of the single-channel neural WPE is denoted as z_0. In the third and the fourth rows, the actual and virtual signal outputs of the VACE-WPE are refered to as z_1 and z_v, respectively, whereas the last row represents the first-channel output of the dual-channel neural WPE employing the actual second-channel signal as the input. Herein, we first analyze the results in the medium and large room environments, followed by the results in the small room.

Firstly, in the medium and large room conditions, comparing the first three rows indicates that both the WPE algorithms have certainly improved the quality of the unprocessed signal (x_1). Moreover, the proposed VACE-WPE (z_1) method outperformed the single-channel WPE (z_0) in terms of all of the evaluation metrics, which implies that the VACENet is capable of generating a virtual signal that is effective as the secondary input for the dual-channel WPE. Still, the dual-channel WPE fed with the actual dual-channel input signals was superior to the VACE-WPE with respect to most of the metrics. On the contrary, the virtual signal output of the VACE-WPE (z_v) revealed considerably different characteristics relative to the rest of the dereverberated signals. In fact, z_v showed the worst performance compared to all the other candidates in terms of all of the evaluation metrics, and was even inferior to the unprocessed signal. To visualize the spectral characteristics of the virtual signals, we plotted sample spectrograms of the input and output signals of the VACE-WPE in the large room environment in Fig. 3. As seen in the figure, the spectrograms of x_v and z_v do not exhibit the spectral patterns similar to those of the spectrograms of x_1 and z_1, respectively. Nonetheless, the "traces" of the spectra of x_1 have been reduced in z_1 over the entire frequency region.

Secondly, in the small room environment, the overall signal quality was much better than that measured in the medium or large room conditions. The proposed VACE-WPE (z_1) showed comparable performance to the single-channel and the actual dual-channel WPE methods in terms of the SRMR, while being slightly superior with regard to the PESQ. In contrast, the reverberated signal (x_1) exhibited the lowest CD and LLR metrics, which may be because the LP orders of the WPE algorithms set for the inference are too large, hence leading to overestimation of reverberation as well as speech distortion; another possible reason is that the small room acoustics are unseen during the training of the LPSNet and VACENet, which may somehow deteriorate the neural WPE systems. Meanwhile, z_v showed a similar pattern to that described in the larger rooms.

In summary, we conjecture that the VACENet simply learns to generate an auxiliary signal that can assist the single-channel signal in such a way as to obtain a better estimate of the late reverberation component within the MIMO WPE framework, rather than learning to generate a secondary signal that is observable within a specific microphone arrangement. This may be possible because the underlying algorithm of the MIMO WPE is the MCLP, which is clearly different from that of the single-channel WPE, thus providing a room for improvement over the single-channel counterpart and enabling the neural network to solve the many-to-one mapping problem of finding the early arriving target speech via the MCLP given the reverberated observations.

6. Conclusions

In this study, we investigated the feasibility of improving the single-channel neural WPE dereverberation algorithm by augmenting the monaural speech signal with a virtual signal generated by a neural network, and then passing them through the dual-channel WPE. Regularizing the training of the VACENet by reinforcing the training loss functions with physically meaningful constraints or extending the current VACE framework to beamformers for speech enhancement could be directions for future work.

7. Acknowledgements

This research was supported and funded by the Korean National Police Agency. [Project Name: Real-time speaker recognition via voiceprint analysis / Project Number: PA-J000001-2017-101]

8. References

[1] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[2] K. Kinoshita, M. Delcroix, H. Kwon, T. Hori, and T. Nakatani, "Neural network based spectrum estimation for online WPE dereverberation," in *Proc. INTERSPEECH*, 2017, pp. 384–388.

[3] L. Drude, C. Boeddeker, J. Heymann, R. Haeb-Umbach, K. Kinoshita, M. Delcroix, and T. Nakatani, "Integrating neural network based beamforming and weighted prediction error dereverberation," in *Proc. INTERSPEECH*, 2018, pp. 3043–3047.

[4] H. Katahira, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Virtually increasing microphone array elements by interpolation in complex-logarithmic domain," in *Proc. European Signal Processing Conference* (EUSIPCO), 2013, pp. 1–5.

[5] ——, "Generalized amplitude interpolation by β-divergence for virtual microphone array," in *Proc. International Workshop on Acoustic Signal Enhancement* (IWAENC), 2014, pp. 149–153.

[6] K. Yamaoka, L. Li, N. Ono, S. Makino, and T. Yamada, "CNN-based virtual microphone signal estimation for MPDR beamforming in underdetermined situations," in *Proc. European Signal Processing Conference* (EUSIPCO), 2019, pp. 1–5.

[7] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, "A fully convolutional neural network for complex spectrogram processing in speech enhancement," in *Proc. International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2019, pp. 5756–5760.

[8] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. International Workshop on Machine Learning for Signal Processing* (MLSP), 2017, pp. 1–6.

[9] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "PHASEN: A phase-and-harmonics-aware speech enhancement network," *arXiv preprint arXiv:1911.04697*, 2019.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention* (MICCAI), 2015, pp. 234–241.

[11] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. International Conference on Machine Learning* (ICML), 2017, pp. 933–941.

[12] J. S. Garofolo, "TIMIT acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[13] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2017, pp. 5220–5224.

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. Automatic Speech Recognition and Understanding Workshop* (ASRU), 2011.

[15] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas *et al.*, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[16] S. Pirhosseinloo and J. S. Brumberg, "Monaural speech enhancement with dilated convolutions," in *Proc. INTERSPEECH*, 2019, pp. 3143–3147.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

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

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[20] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. International Conference on Machine Learning* (ICML), 2013, pp. 1310–1318.

[21] 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 *Proc. International Conference on Acoustics, Speech, and Signal Processing* (ICASSP), 2001, pp. 749–752.

[22] 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.

[23] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *Proc. International Workshop on Acoustic Signal Enhancement* (IWAENC), 2014, pp. 55–59.