INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Monaural speech enhancement with dilated convolutions

Shadi Pirhosseinloo[1], *Jonathan S. Brumberg*[1,2]

[1]Department of Electrical Engineering and Computer Science, University of Kansas, USA
[2]Department of Speech-Language-Hearing, University of Kansas, USA

shadi@ku.edu, brumberg@ku.edu

Abstract

In this study, we propose a novel dilated convolutional neural network for enhancing speech in noisy and reverberant environments. The proposed model incorporates dilated convolutions for tracking a target speaker through context aggregations, skip connections, and residual learning for mapping-based monaural speech enhancement. The performance of our model was evaluated in a variety of simulated environments having different reverberation times and quantified using two objective measures. Experimental results show that the proposed model outperforms a long short-term memory (LSTM), a gated residual network (GRN) and convolutional recurrent network (CRN) model in terms of objective speech intelligibility and speech quality in noisy and reverberant environments. Compared to LSTM, CRN and GRN, our method has improved generalization to untrained speakers and noise, and has fewer training parameters resulting in greater computational efficiency.

Index Terms: noise and speaker independent speech enhancement, dilated convolutions, residual learning, deep neural networks

1. Introduction

In real world environments, speech signals are corrupted by nonspeech noise, interfering speech, and room reverberation. The presence of such acoustic interference and reverberation has a negative effect on speech quality and speech intelligibility in speech processing applications such as speaker identification systems and automatic speech recognition (ASR) as well as normal hearing and hearing-impaired listeners, especially when the level of signal-to-noise ratio (SNR) is low. Monaural speech separation algorithms can improve the performance of ASR, hearing aid design and mobile communication in adverse environments by separating target speech from background noise and reverberation using a single channel recording. In this study, we focus on speech enhancement which aims to separate target speaker from background noise and reverberation.

For decades, monaural speech separation has been studied for speech processing applications. Inspired by Time-Frequency (T-F) masking based on computational auditory scene analysis (CASA) [1, 2], monaural speech separation has been recently treated as a supervised learning problem. The choice of training model, acoustic features, and training target are all important factors for accurate supervised speech separation. The ideal binary mask (IBM) [3], ideal ratio mask (IRM) [4] were proposed as training targets for masking based supervised speech separation, while target magnitude spectrum (TMS) [5] was used as a training target in mapping based supervised speech separation. Furthermore, recent studies have examined the effect of different input acoustic features (e.g., gammatone based features versus spectral features) [6, 7] on supervised speech separation in noisy and reverberant condition.

Noise generalization and speaker generalization are also very important in supervised speech separation. Training speech enhancement models (such as neural networks) with different types of noise can solve the noise generalization problem. The speaker generalization problem can be similarly solved with training models with a large number of speakers. However, recent studies [8, 9] show that feedforward deep neural networks (DNN), which currently enjoy a popularity in the field, have limited capacity for modeling large numbers of speakers and in the presence of different speakers in training dataset; DNN have difficulty tracking a target speaker. Moreover, when the background noise includes speech components (like babble), DNN models mistake the background noise for target speech [9]. One reason may be due to the small context window typically used in DNN to include temporal context for predicting the target value for each frame; however, DNN are not able to use long-term contexts to track the target speaker [9]. As an alternative, Chen et al. [8] suggest a sequence to sequence mapping to leverage long-term contexts for speech enhancement. A long-short term memory (LSTM) neural network [10] was proposed in [8, 9] to solve speaker generalization problem by using different noises and speakers in training and LSTM model outperforms DNN based model on unseen speakers.

In recent years, convolutional neural networks (CNN), recurrent neural networks (RNN), and their combinations have been used for speech enhancement with noise and speaker independent modeling. For example, an LSTM network was proposed as a noise and speaker independent model that had better performance to unseen speakers compared to DNN [9]. Two additional studies for speech enhancement proposed gated residual networks (GRN) [11] and convolutional recurrent neural networks (CRN) with encoder-decoder structure [12], which resulted in greater generalization for untrained speakers and fewer trainable parameters compared to LSTM. However, most of the proposed methods were only trained and tested in noisy environments with no reverberation. Furthermore, GRN and CRN have large number of training parameters which increase the training time and computational cost.

Inspired by the dilated convolutions method [13], which aggregates multiscale contextual information without losing resolution, we develop a fully-convolutional neural network that utilizes 2D convolutional networks with pooling layers for feature extraction. Furthermore, the network includes two blocks of stacked dilated 1D convolutions with skip connections for speech enhancement in noisy-reverberant environment. The results of our analyses show that the proposed dilated convolutional network has better performance and lower computational cost compared to LSTM and previous convolutional networks for speech enhancement.

The paper is organized as follows: the new dilated convolutional neural network with skip connections and residual learning are described in section 2. The experimental setup and

http://dx.doi.org/10.21437/Interspeech.2019-2782

results are presented in sections 3 and 4, followed by a brief conclusion in section 5.

2. System description

The goal of our proposed monaural speech enhancement model is to estimate the target speaker from the single microphone noisy-reverberant mixture. Supervised speech enhancement can be treated as a regression problem that maps acoustic features from noisy-reverberant mixtures to a training target, which can be either time frequency mask or a spectral representation of a target speaker. The acoustic features and training target are passed to the neural network for training and in the testing stage, the estimated output and the phase of noisy-reverberant mixture are then resynthesized into a time-domain speech waveform.

In this study, the noisy-reverberant mixtures were generated by convolving the target speaker audio with an impulse response (IR) for different room locations and reverberations, then mixed with different types of non-stationary background noises [14]. The signal was segmented into time frames with a 20 ms hamming window and 10 ms overlap (50%). We used a 320-point short-time Fourier transform (STFT) to calculate the magnitude spectra, which was used as input acoustic features with the dimension of 161.

2.1. Proposed model

Dilated convolutions were developed for semantic segmentation that aggregates multi-scale contextual information and supports exponential expansion of the receptive field without loss of resolution [13]. The 1-dilated convolution is the conventional convolution and its receptive field scale increases linearly with the layer depth. However, the scale of receptive field in dilated convolution increases exponentially with the depth of the layer when the kernels are stacked with exponentially increasing dilation rates [13].

In this work, we first stack 4 layers of 2D convolution with exponential linear unit (ELU) activation function [15], batch normalization (BN) [16], and the maxpooling layer with the of size (1,2) for feature extraction to capture contextual information in both time and frequency axis. The learned features from 2D convolutions are then reshaped and are given as an input for 1D convolutions. Our model includes two residual blocks that perform 1D convolution with a kernel size of 3 and 256 output channels and 1D dilated convolutions with exponentially increasing dilation factors as 2,4,8,16,32,64,128 with the kernel size of 3 and 16 output channels. The dilated convolutions enable our model to deepen the network by increasing the dilation factor instead of increasing the length of the filter. Stacking a set of layers with dilation to the maximum dilation factor can create context stacks. Moreover, the output of each dilated block is given to another 1D convolution with 256 output channels and finally passed through a sigmoid activation function that operates as a soft mask where it is then multiplied with the learned features from previous 1D convolution layer. Furthermore, we have employed skip connections that provide advantages in training deep models and allow the network to use features that are extracted from different hierarchical levels for final prediction. These skip connections preserve and integrate the knowledge learned by each stacked layer. The activation function on the output layer is rectified linear unit (ReLU) [17] for mapping based speech enhancement models, while sigmoid is used as an activation function for masking based speech enhancement.

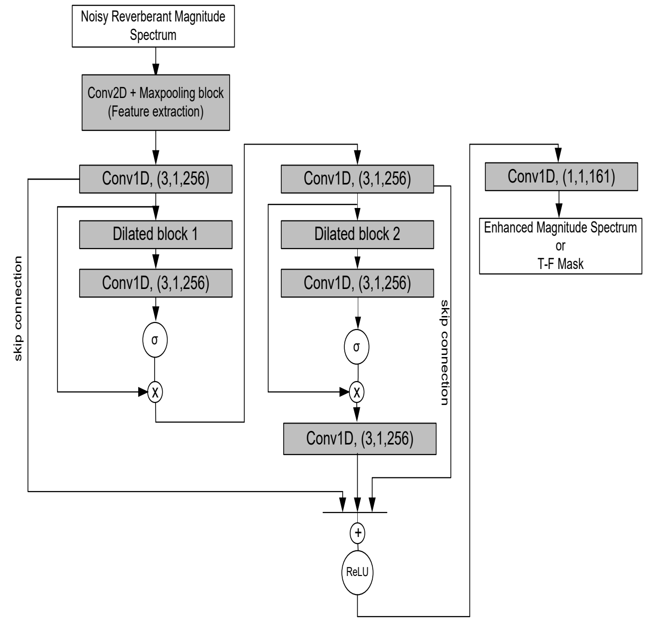

Figure 1: *Proposed dilated convolutional network architecture.*

Our proposed dilated convolutional network is shown in Fig.1 and the detailed description of it is given in Table 1. The hyperparameters for 2D convolution layers are shown as (kernel size, (strides), output channel), while the hyperparameters for 1D convolution are given as (kernel size, dilation rate, output channels) in Table 1. Note that the zero-padding is applied to both 2D convolution and 1D convolution layers. We compared our network with three different baselines of LSTM, CRN and GRN explained as follows:
LSTM baseline: The LSTM includes four hidden layers with the size of 1024 units while the input and output layers have 161 units.
CRN baseline: The CRN [12] includes an encoder and decoder with five convolutional layers and five deconvolutional layers, respectively. Moreover, the convolutional layers are causal and have a stride of 2 along frequency. In addition, 2 layers of LSTM are stacked between encoder and decoder to model temporal dynamics of speech and skip connections are used to concatenate the output of each encoder layer to the input of each decoder layer.
GRN baseline: GRN is a 62 convolutional layer network, constructed with a stack of frequency-dilated convolutional layers and multiple gated residual blocks with different dilation rates [11].

2.2. Training targets

In this work, we examine two different training targets of clean speech: ideal ratio mask (IRM) [4] and target magnitude spectrum (TMS) [5]. IRM has been widely used in supervised speech separation, which can interpreted as a smooth version of IBM [3] but with improved performance in comparison to IBM [4]:

$$\mathrm{IRM}(\omega,\tau)=\sqrt{\frac{s^2(\omega,\tau)}{s^2(\omega,\tau)+n^2(\omega,\tau)}} \tag{1}$$

Table 1: *Architecture of our proposed dilated convolutional network.*

Block name	**Layer name**	**hyperparameters**
Conv2d+Maxpooling	conv2d_1	$5 \times 5, (1 \times 1), 32$, BN, ELU
	conv2d_2	$9 \times 9, (1 \times 1), 32$, BN, ELU
	maxpool_1	$1 \times 2, (1 \times 2)$
	conv2d_3	$5 \times 5, (1 \times 1), 64$, BN, ELU
	conv2d_4	$9 \times 9, (1 \times 1), 64$, BN, ELU
	maxpool_2	$1 \times 2, (1 \times 2)$
Conv1d	conv1d_1	$3, 1, 256$, BN, ELU
Dilated block1	conv1d_d1	$3, 2, 16$, ELU
	conv1d_d2	$3, 4, 16$, ELU
	conv1d_d3	$3, 8, 16$, ELU
	conv1d_d4	$3, 16, 16$, ELU
	conv1d_d5	$3, 32, 16$, ELU
	conv1d_d6	$3, 64, 16$, ELU
	conv1d_d7	$3, 128, 16$, ELU
Conv1d	conv1d_2	$3, 1, 256$, ELU
Conv1d	conv1d_3	$3, 1, 256$, BN, ELU
Dilated block2	conv1d_d8	$3, 2, 16$, ELU
	conv1d_d9	$3, 4, 16$, ELU
	conv1d_d10	$3, 8, 16$, ELU
	conv1d_d11	$3, 16, 16$, ELU
	conv1d_d12	$3, 32, 16$, ELU
	conv1d_d13	$3, 64, 16$, ELU
	conv1d_d14	$3, 128, 16$, ELU
Conv1d	conv1d_4	$3, 1, 256$, ELU
Conv1d	conv1d_5	$3, 1, 256$, ELU
Conv1d	conv1d_6	$1, 1, 161$, ReLU/Sigmoid

where $n^2(\omega, \tau)$ and $s^2(\omega, \tau)$ are the energy of noise and speech in each T-F unit, respectively. In masking-based speech separation, the estimated IRM is multiplied with the magnitude spectrum of noisy mixture to calculate the magnitude of separated speech. The result is then used with the phase of the noisy mixture to reconstruct an audio waveform of the processed speech. In mapping-based speech separation, TMS is the training target where the estimated magnitude spectrum is used with noisy phase to construct the processed speech waveform.

3. Experimental setup

3.1. Dataset

We used speech material from the TIMIT corpus [18] for training and testing. A total number of 1700 utterances spoken by 462 speakers (male and female) were used for training along with 300 different background noises (15 hours) from the FreeSFX [19] and Freesound [20] corpora. Finally, the Babble and Cafeteria noises drawn from the NOISEX [21] and DEMAND [22] repositories were used as unseen noises during testing.

The noisy-reverberant mixtures (Total=34000, training=30600, development=3400) used for training the neural networks in this study were generated first by convolving the clean speech signal with a randomly selected room impulse response (IR). Each IR simulated reverberant room conditions according to the image method technique [23] followed by adding a random selection of the different background noises. The IRs with different reverberation times (T_{60}) in the range of $\{0.3 - 0.9\}s$ and an anechoic condition with $T_{60} = 0.0s$ were created by the room impulse response generator [24]. Room dimension was $(10 \times 9 \times 8)$ m and the microphone was located at $(3, 4, 1.5)$ m. The target speaker was located at a random position at different distances in the range of 0.5–2.5 m from the microphone. A random SNR in the range [-5,5] dB was used for the training set and was based on the reverberant target speech instead of anechoic speech.

We generated two test sets to investigate generalization of speakers and noise in each of the models. Briefly, one set used 6 trained speakers (3 males, 3 females) and the other test set used 6 untrained speakers (3 males, 3 females). The noisy-reverberant mixtures were generated by convolving the clean speech signal with real IRs recorded at the University of Surrey [25]. The Surrey database includes IRs recorded from four reverberant rooms with T_{60} of 0.32 s, 0.47 s, 0.68 s and 0.89 s, respectively and the distance between the sound source and the microphone is 1.5 m.

- Test Set 1: 1000 mixtures were created from 6 trained speakers convolved with random Surrey IRs and mixed with random segment of background noise of Cafeteria and Babble.
- Test Set 2: 1000 mixtures were created from 6 untrained speakers convolved with random Surrey IRs and mixed with a random segment of background noise of Cafeteria and Babble.

The SNR for both test sets was in three levels of -5, 0 and 5 dB. No similar noise segment, speech utterance or IR are included in both training and testing data sets and the noises, speech utterances, and IRs were resampled to 16 KHz.

The neural network cost function for all models was chosen as the mean square error (MSE) and the Adam algorithm [26] was selected for backpropagation optimization. We trained the convolutional models with the mini-batch size of 16 while the batch size in LSTM was 256. The number of training epochs was 50 and all features were normalized to have zero mean and unit variance on each frequency channel before feeding to the networks. Batch normalization was applied on hidden layers for faster training [16].

3.2. Evaluation criteria

The short-time objective intelligibility (STOI) [27] and perceptual evaluation of speech quality (PESQ) [28] were used to evaluate speech enhancement model performance. The correlation between the temporal envelopes of clean and processed signals is calculated in STOI, resulting in a score ranging from 0 to 1 with higher score indicating greater objective speech intelligibility [27]. PESQ measures objective speech quality between the processed speech and clean speech, and results in a score between [-0.5,4.5] where higher values are reflective of greater speech quality [28].

4. Evaluation results

Table 2 shows the STOI and PESQ scores for unprocessed and processed signals for trained speakers using mapping based (TSM) and masking based (IRM) speech enhancement models. The four different neural network structures are compared in terms of STOI and PESQ in three SNR levels for two different noises of Babble (Bab) and Cafeteria (Caf). The proposed dilated convolutional neural network consistently outperforms LSTM, CRN and GRN in both criteria. The average of STOI improvements for proposed model with TSM over unprocessed, LSTM processed, CRN processed and GRN processed are 14.48%, 8.33%, 7.17% and 1.88%, respectively. Furthermore, this model has higher PESQ score compared to other three networks. In masking-based speech enhancement, the dilated convolutions network still outperforms other network structures in terms of STOI with improvement of 10.82%,

Table 2: *Model and Training target comparison in terms of STOI and PESQ on trained speakers.*

Metrics	**STOI (%)**									**PESQ**								
SNR	**-5 dB**			**0 dB**			**5 dB**			**-5 dB**			**0 dB**			**5 dB**		
Noise	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***
Unprocessed	51.11	55.08	53.09	58.29	60.74	59.51	64.83	67.50	66.17	1.31	1.48	1.39	1.57	1.77	1.67	1.87	2.02	1.94
LSTM+TSM	53.95	60.69	57.32	64.98	68.29	66.63	72.48	74.05	73.26	1.37	1.62	1.49	1.77	1.96	1.86	2.04	2.15	2.09
CRN+TSM	57.65	62.20	59.92	66.21	68.49	67.35	72.71	74.12	73.41	1.48	1.67	1.57	1.81	1.96	1.88	2.05	2.16	2.10
GRN+TSM	62.19	68.39	65.29	71.25	74.34	72.75	77.77	79.20	78.48	1.58	1.81	1.69	1.92	2.07	1.99	2.17	2.27	2.22
Proposed+TSM	**64.96**	**70.58**	**67.77**	**73.36**	**75.90**	**74.63**	**79.28**	**80.36**	**79.82**	**1.64**	**1.85**	**1.74**	**1.98**	**2.13**	**2.05**	**2.22**	**2.33**	**2.27**
LSTM+IRM	54.06	59.83	56.94	63.95	66.72	65.33	70.44	72.49	71.46	1.45	1.67	1.56	1.84	2.01	1.92	2.10	2.25	2.17
CRN+IRM	56.23	60.83	58.53	64.97	67.32	66.14	71.38	72.20	71.79	1.49	1.71	1.6	1.85	2.02	1.93	2.12	2.24	2.18
GRN+IRM	58.51	64.26	61.38	67.35	69.85	68.60	73.42	74.95	74.18	1.54	1.77	1.65	1.93	2.10	2.01	2.20	2.33	2.26
Proposed+IRM	**62.35**	**67.23**	**64.79**	**70.14**	**71.60**	**70.87**	**75.28**	**75.88**	**75.58**	**1.62**	**1.87**	**1.74**	**2.00**	**2.16**	**2.08**	**2.26**	**2.39**	**2.32**

Table 3: *Model and Training target comparison in terms of STOI and PESQ on untrained speakers.*

Metrics	**STOI (%)**									**PESQ**								
SNR	**-5 dB**			**0 dB**			**5 dB**			**-5 dB**			**0 dB**			**5 dB**		
Noise	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***	***Bab***	***Caf***	***Avg.***
Unprocessed	50.31	54.31	52.31	57.10	61.12	59.11	63.89	65.69	64.79	1.25	1.44	1.34	1.53	1.75	1.64	1.84	1.97	1.90
LSTM+TSM	56.26	61.76	59.01	65.04	69.24	67.14	71.07	71.96	71.51	1.39	1.66	1.52	1.73	1.93	1.83	2.00	2.08	2.04
CRN+TSM	58.51	63.10	60.80	65.97	69.99	67.98	71.89	72.57	72.23	1.44	1.67	1.55	1.75	1.95	1.85	2.02	2.10	2.06
GRN+TSM	61.74	67.26	64.50	69.79	73.97	71.88	75.67	76.75	76.21	1.49	1.72	1.60	1.81	2.01	1.91	2.06	2.14	2.10
Proposed+TSM	**65.81**	**70.33**	**68.07**	**72.71**	**76.04**	**74.37**	**78.07**	**78.49**	**78.28**	**1.59**	**1.81**	**1.70**	**1.90**	**2.07**	**1.98**	**2.13**	**2.23**	**2.18**
LSTM+IRM	55.71	60.49	58.10	63.91	67.38	65.64	69.77	69.91	69.84	1.41	1.69	1.55	1.80	1.99	1.89	2.10	2.17	2.13
CRN+IRM	57.78	61.45	59.61	64.33	67.81	66.07	69.80	69.94	69.87	1.45	1.70	1.57	1.79	2.00	1.89	2.09	2.17	2.13
GRN+IRM	61.31	64.99	63.15	67.90	70.89	69.39	73.02	72.83	72.92	1.56	1.79	1.67	1.91	2.10	2.00	2.21	2.29	2.25
Proposed+IRM	**63.83**	**66.97**	**65.40**	**69.53**	**71.97**	**70.75**	**74.22**	**73.54**	**73.88**	**1.59**	**1.84**	**1.71**	**1.95**	**2.14**	**2.04**	**2.25**	**2.32**	**2.28**

5.83%, 4.92% and 2.35% compared to unprocessed, LSTM, CRN and GRN, respectively. Overall, the mapping-based methods (TSM) have higher STOI score compared to masking based approach (IRM), though the IRM based networks have slightly better PESQ compared to TSM methods.

We repeated our analysis for untrained speakers to evaluate whether the proposed network generalizes well to untrained speakers. Table 3 compares the performance of LSTM, CRN, GRN and proposed model using TSM and IRM for untrained speakers in terms of STOI and PESQ. The rate of increase in STOI for proposed model with TSM is 14.83%, 7.68%, 6.57% and 2.71% compared to the unprocessed, LSTM, CRN and GRN, while in proposed model with IRM, the change in STOI is 11.27%, 5.48%, 4.82% and 1.52%. In the most challenging condition (SNR=-5 dB and Babble noise) of untrained speakers, the proposed model increases STOI by 15.50% compared to unprocessed mixture, while LSTM, CRN and GRN have 5.95%, 8.2%, 11.43% improvement.

One advantage of the proposed model is that it leverages contexts in both frequency and time axis similar to CRN and GRN which leads to modeling more complex temporal dependency. Moreover, the dilated blocks deepen the network more efficiently than increasing the length of the filter and these blocks work as a soft mask on the learned features of previous layers. Another advantage of this model is its higher computational efficiency. The number of trainable parameters in this model is 2.92 million while the number of parameters in LSTM, CRN and GRN are 25.51, 17.22 and 3.11 millions, respectively.

5. Conclusions

This paper presents a novel dilated convolutional neural network structure for noise and speaker independent speech enhancement. The evaluation of speech enhancement performance indicates that the proposed model outperforms LSTM, CRN and GRN in terms of STOI and PESQ for both trained and untrained speakers. Furthermore, the proposed model has a smaller number of trainable parameters compared to recurrent neural network and other proposed convolutional networks. The proposed model has significant improvement in monaural speech enhancement and it has better generalization to different types of unseen noises and untrained speakers compared to previous models.

6. References

[1] D. L. Wang, and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications.* Hoboken, NJ: Wiley/IEEE Press, 2006.

[2] S. Pirhosseinloo, and K. Kokkinakis, "Time-frequency masking for blind source separation with preserved spatial cues," In *Proceedings of Interspeech*, 2017, pp. 1188–1192.

[3] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, pp. 181–197, 2005.

[4] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

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

[6] M. Delfarah, and D. L. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1085–1094, 2017.

[7] S. Pirhosseinloo, and J. S. Brumberg, "A new feature set for masking-based monaural speech separation," in *52^{nd} Asilomar Conference on Signals, Systems, and Computers*, 2018, pp. 828–832.

[8] J. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[9] J. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," in *Proceedings of Interspeech*, 2016, pp. 3314–3318.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] K. Tan, J. Chen, and D. L. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 189–198, 2019.

[12] K. Tan and D. L. Wang, "A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement," in *Proceedings of Interspeech*, 2018, pp. 3229–3233.

[13] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2016.

[14] S. Pirhosseinloo, and K. Kokkinakis, "An interaural magnification algorithm for enhancement of naturally-occurring level differences," in *Proceedings of Interspeech*, 2016, pp. 2558–2561.

[15] D. A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[16] S. Ioffe, and C. Szegedy, " Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICML)*, 2015, pp. 448–456.

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

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, " TIMIT acoustic phonetic continuous speech corpus LDC93S1 (web download)," 1993.

[19] "FreeSFX," http://www.freesfx.co.uk/, 2017.

[20] "Freesound," http://freesound.org/, 2015.

[21] A. Varga, and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.

[22] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multichannel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of Acoustical Society of America*, vol. 133, pp. 3591–3591, 2013.

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

[24] E. Habets, "Room impulse response generator," 2010.[Online]. http://home.tiscali.nl/ehabets/rir generator.html

[25] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1867–1871, 2010.

[26] D. Kingma, and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[27] 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, no. 7, pp. 2125–2136, 2011.

[28] A. Rix, J. Beerends, M. Hollier, and A. 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 (ICASSP)*, 2001, vol. 2, pp. 749–752.