MULTI-SCALE RESIDUAL CONVOLUTIONAL ENCODER DECODER WITH BIDIRECTIONAL LONG SHORT-TERM MEMORY FOR SINGLE CHANNEL SPEECH ENHANCEMENT

Yang Xian[1], *Yang Sun*[2], *Wenwu Wang*[3], *Syed Mohsen Naqvi*[1]

[1]Intelligent Sensing and Communications Research Group, Newcastle University, UK
[2]Big Data Institute, University of Oxford, UK
[3]Centre for Vision, Speech and Signal Processing, University of Surrey, UK

ABSTRACT

The existing convolutional neural network (CNN) based methods still have limitations in model accuracy, latency and computational cost for single channel speech enhancement. In order to address these limitations, we propose a multi-scale convolutional bidirectional long short-term memory (BLSTM) recurrent neural network, which is named as McbNet, a deep learning framework for end-to-end single channel speech enhancement. The proposed McbNet enlarges the receptive fields in two aspects. Firstly, every convolutional layer employs filters with varied dimensions to capture local and global information. Secondly, the BLSTM is applied to evaluate the interdependency of past, current and future temporal frames. The experimental results confirm the proposed McbNet offers consistent improvement over the state-of-the-art methods and public datasets.

Index Terms— CNN, single channel, speech enhancement, BLSTM, McbNet, receptive field

1. INTRODUCTION

In the last decade, statistical signal processing [1], [2] and computational auditory scene analysis (CASA) based methods [3], [4] have been introduced for speech separation and enhancement. Single channel speech enhancement is the task of enhancing the intelligibility and quality of the target speech extracted from noisy speech mixture that is recorded by a single-microphone. It has been exploited in many real-world applications such as mobile speech communication, speech recognition and robotics [5]–[8].

In recent years, deep neural network (DNN) has become a popular solution to single channel speech enhancement. The DNN is introduced to estimate the mapping relation between time-frequency (T-F) features of noisy mixture and clean speech [9]. Furthermore, the DNN is employed as a non-linear regression function to estimate the mapping relation between the noisy mixture and target speech, which ensures the powerful modelling capability [10]. The DNN with skip connection is proposed to learn the residual relation between the noisy mixture and target speech [11]. Alternatively, based on W-disjoint orthogonality, only one source is predominantly active at each T-F point [12], speech enhancement can be achieved by computing the weight function (mask) of each T-F bin. This has led to a variety of methods using, the ideal binary mask (IBM), ideal ratio mask (IRM) and derverberation mask (DM) [13], [14]. Recently, the long-short term memory (LSTM) RNN provides better generalization of speaker independent speech enhancement by exploiting the interdependency of past and current temporal frames [15]. The magnitude spectrogram is simply treated as an image by the convolutional encoder decoder (CED). It is used to map the spectrogram of noisy speech mixture to that of clean speech [16]. The LSTM has been incorporated with the CNN provides a consistent improvement over the CED and LSTM methods, which is named as CRN [17].

The aforementioned methods have several limitations. For example, the kernel size of CED and CRN is fixed, which only captures the local information, while the global information about the interdependency between long-term temporal frames is not well utilized. Therefore the information flow between input layer and output layer. In this paper, we propose a multi-scale convolutional bidirectional LSTM (BLSTM), in short as McbNet. Firstly, we introduce a multi-scale CNN, where filters with varied sizes are employed in every convolutional sub-layer, thus offering high-level feature in different scales, which captures the interdependency between the local and global information. Secondly, the dependency of past, current and future temporal frames is captured by exploiting forward and backward LSTMs in each BLSTM layers.

The remainder of the paper is organized as follows. Section 2 describes the proposed McbNet method. The experimental settings and results are given in Section 3. Section 4 concludes the paper.

2. THE PROPOSED METHOD

2.1. Problem Statement

In single channel speech enhancement, the noisy speech mixture can be written as:

$$y(m) = s(m) + n(m) \tag{1}$$

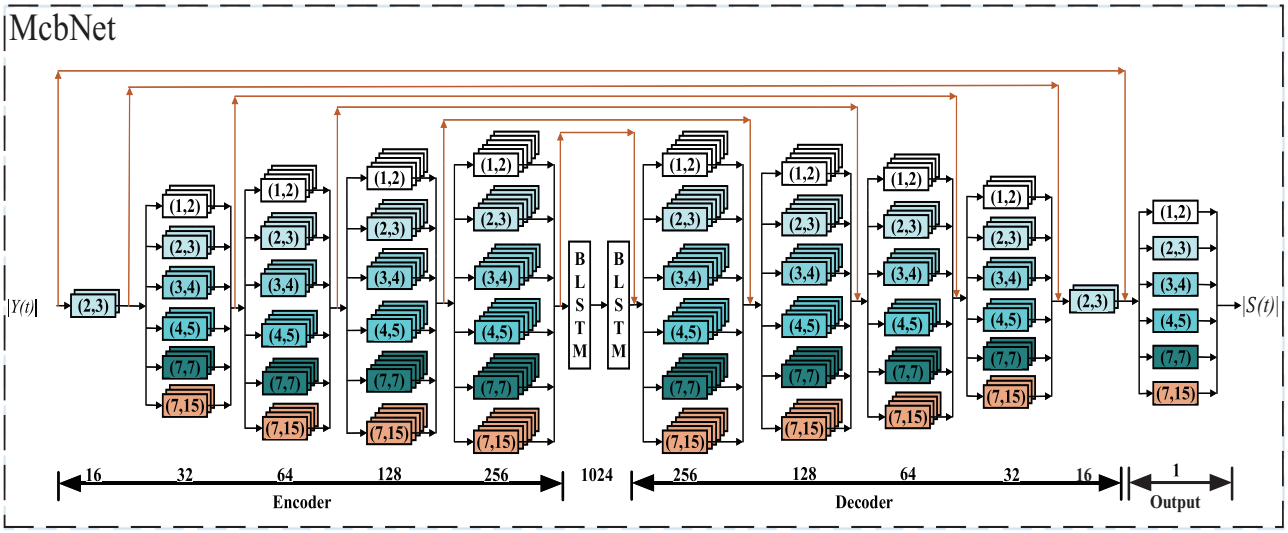

Fig. 1: Block diagram of the proposed McbNet architecture. Different colors represent different filter (kernel) sizes, which can be seen on every block. The number of channels are listed on the bottom of diagram. The black lines denote the standard feed-forward transmission, and the red lines denote the skip connection.

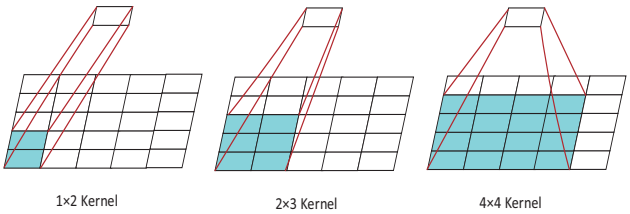

Fig. 2: Block diagram of the three kernel sizes, the shaded blocks represent the receptive fields of filter. The $1\times$ 2 kernel size is used to capture the information for adjacent two cells, which is more appropriate for local information processing. The 4×4 kernel has larger repetitive filed, and global information inside the 4×4 units is captured, with the trade off local information.

where $y(m)$ denotes the noisy speech, $s(m)$ and $n(m)$ represent the clean speech signal and noise at time m, respectively. By using the short time Fourier Transform (STFT), the spectrogram of noisy speech mixture at time frame t and frequency bin f is obtained as:

$$Y(t,f) = S(t,f) + N(t,f) \tag{2}$$

where $S(t,f)$ and $N(t,f)$ are the spectrograms of clean speech signal and noise, respectively. The neural network model is trained to find the mapping relation G_θ between the magnitude spectrograms of clean speech signal $|S(t,f)|$ and noisy speech mixture $|Y(t,f)|$. Where G_θ is parametrized by θ. The mapping function is estimated by optimizing the loss function as:

$$\begin{aligned} Loss &= \min_\theta \sum_t \sum_f \left(G_\theta(|Y(t,f)|) - |S(t,f)|\right)^2 \\ &= \min_\theta \sum_t \sum_f \left(|\hat{S}(t,f)| - |S(t,f)|\right)^2 \end{aligned} \tag{3}$$

where $|\hat{S}(t,f)|$ is the magnitude spectrogram of the estimated target speech, which is combined with phase information of the noisy mixture to estimate the target speech.

2.2. Multi-scale Encoder Decoder

The block diagram of the proposed McbNet is shown in Fig. 1. The proposed McbNet is an end-to-end framework that estimates the spectral magnitude of target speech by using non-linear mapping. The McbNet mainly includes two blocks: multi-scale convolutional encoder-decoder, and BLSTM layers. More specifically, the multi-scale convolutional encoder is employed to extract high-level representation from the magnitude spectrum of noisy speech mixture. The multi-scale convolutional encoder (MCE) is exploited to project this high-level representation back to lower dimension. Empirically, every MCE layer consists of six 2-D convolutional sub-layers with the different filter (kernel) sizes. The filters of small size such as (1,2), (2,3) and (3,4) are used to capture the local dependency between the adjacent temporal frames, which is good at extracting feature from the short duration vocal information. The filters of large size (4,5), (7,7) and (7,15) are exploited to capture the global dependency of different frames, which has an advantage in feature extraction from the long duration speech. Then, the the output of convolutional sub-layers are added together to obtain the output of one encoder layer. The batch-normalization is followed after each convolutional sub-layer. Different from the standard CNN that uses the ReLU activation function, the proposed McbNet utilizes the advanced activation function (leakyReLU), which provides better generalization ability and accelerates the convergence speed [18]. The multi-scale convolutional decoder (MCD) has a symmetric structure with MCE, and every MCD layer consists of multi-scale decovolutional sub-layers with varied filter sizes. Also, the batch-normalization and LeakyReLU are applied in MCD. The MCE and MCD are connected by skip connection and two BLSTM layers. Note, the kernel sizes and number of channels can be found in Fig. 1, and the stride size of McbNet is fixed to (1,2).

Table 1: Speech enhancement performance comparisons in terms of STOI over three noises with different state-of-the-art methods and SNR levels. Each result is the average value of 100 experiments. *Italic* text refers to the proposed methods. **Bold** number indicates the best performance.

Measure	STOI(%)											
Scenarios	Babble				Leopard				N56			
SNR / Methods	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture	53.86	63.05	71.66	62.86	71.68	75.57	78.92	75.39	57.07	68.26	78.43	67.92
DNN [10]	66.36	72.91	79.26	72.83	77.11	80.55	83.26	80.30	72.59	79.44	84.07	78.70
S-DNN [11]	66.80	73.66	79.63	73.36	78.34	81.54	83.98	81.29	72.79	79.45	84.07	78.77
LSTM [15]	68.78	75.81	81.54	75.38	80.76	83.32	85.40	83.16	76.99	82.49	86.47	81.98
CRN [17]	70.10	76.95	81.88	76.31	81.20	84.02	85.80	83.67	78.49	83.37	87.09	82.98
CRN-BLSTM	70.30	77.08	81.96	76.45	81.20	84.20	85.90	83.77	78.62	83.53	87.20	83.11
McbNet	**72.80**	**79.15**	**84.15**	**78.70**	**83.51**	**85.40**	**87.03**	**85.31**	**80.79**	**85.54**	**89.06**	**85.13**

2.3. BLSTM Layers and Residual Multi-scale Output layers

The BLSTM layer contains forward and backward LSTM layers. Using the input sequence from $c(1)$ to $c(T)$, the forward LSTM outputs $\overrightarrow{h}(1)$ to $\overrightarrow{h}(T)$. Each forward LSTM block not only receives sequence from encoder, but also receives sequence from the previous forward LSTM block within the same layer. Therefore, the forward LSTM layer is capable of utilizing the interdependency of past and current temporal frames. Similarly, the backward LSTM can capture the interdependency of current and future temporal frames, and outputs $\overleftarrow{h}(1)$ to $\overleftarrow{h}(T)$. Then, the $\overrightarrow{h}(t)$ and $\overleftarrow{h}(t)$ are fed into the merge block, which gives the final output of BLSTM layers. Empirically, we select the summation function, so the output of BLSTM layer becomes:

$$h(t) = \overrightarrow{h}(t) + \overleftarrow{h}(t) \tag{4}$$

To overcome the potential overfitting, we add the skip connection from the input to the multi-scale output layers. Therefore, the multi-scale output layer can estimate the magnitude of the target speech from the information flow of previous layer and input magnitude of the noisy mixture. Similarly, the multi-scale output layer is 2D-deconvolutional layer, which contains six sub-layers, and the kernel sizes of these sub-layers are different. Thus, the multi-scale output layer utilizes the local and global information. The stride size of the output layer is set to (1,1), and batch-normalization and linear activation are followed.

3. EXPERIMENTAL EVALUATIONS

3.1. Datasets

In our experiments, 1000 and 100 clean utterances are used to generate the training and testing datasets, which are randomly selected from the TIMIT corpus [19]. The TIMIT database contains 6300 utterances which are spoken by 630 speakers. For training, 20 training noise interferences are randomly selected from the None-Speech Sounds [20] and NOISEX-92 [21] datasets. The testing noise interferences are categorized into two scenarios, the seen noise interferences (Babble, Leopard) and the unseen noise interference (N56). The noisy mixtures are generated by mixing the clean utterances and noise interferences at -5dB, 0dB and 5dB signal-to-noise ratio (SNR) levels. In total, 20,000 noisy mixtures in the training dataset, and 300 noisy mixtures in the testing dataset are used.

The signal to distortion ratio improvement (ΔSDR) [22], perceptual evaluation of speech quality (PESQ) [23] and short-time objective intelligibility (STOI) [24] are used to measure the performance. The ΔSDR is equal to SDR of the estimated speech minus SDR of the unprocessed noisy mixture, which is used to evaluate the overall enhancement performance. The PESQ ranges from zero to five, which indicates the intelligibility score of speech. The STOI ranges from zero to one, which indicates human speech quality score. The higher value of measurement indicates a better enhancement performance.

3.2. Baselines and Parameters

The proposed McbNet is compared with four state-of-the-art methods, the standard DNN method in [10]; the DNN method with skip connection S-DNN in [11]; and the LSTM model used in [15], these all methods have four hidden layers that contain 1024 units. The parameters of the CRN model are set by following [17]. The input and output layers for all methods are 257 units. The Adam optimization algorithm and mean square error (MSE) are employed in the baseline and the proposed methods. The dropout rate is fixed to 0.2. The sample rate is 16,000 Hz, and the window length is 512. Further parameters is provided in Fig. 1.

3.3. Experimental Results

Fig. 3 and Tables 1 & 2 provide experimental results in terms of ΔSDR, STOI and PESQ for the baseline and the proposed methods with unseen and seen noises. The testing speakers are different from the training speakers.

The DNN generates, on average, 5.36dB ΔSDR, 77.27% STOI and 2.04 PESQ, which provides the lowest improvements over the unprocessed noisy mixture. These results show that the generalization of DNN remains insufficient. The S-DNN slightly outperforms the DNN. Since the S-DNN employs the skip connection to learn the residual mapping form the noisy mixture.

The LSTM generates, on average, 6.81dB ΔSDR, 80.27% STOI and 2.30 PESQ score, which shows better

Table 2: Speech enhancement performance comparisons in terms of PESQ over three noises with different state-of-the-art methods and SNR levels. Each result is the average value of 100 experiments. *Italic* text refers to the proposed method. **Bold** number indicates the best performance.

Measure	PESQ											
Scenarios	Babble				Leopard				N56			
SNR / Methods	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture	1.28	1.52	1.81	1.53	1.75	1.99	2.22	1.97	1.14	1.31	1.57	1.34
DNN [10]	1.58	1.90	2.20	1.89	2.03	2.31	2.50	2.28	1.70	1.98	2.16	1.95
S-DNN [11]	1.69	2.00	2.28	1.99	2.25	2.45	2.67	2.46	1.74	2.00	2.20	1.98
LSTM [15]	1.82	2.15	2.44	2.14	2.41	2.61	2.80	2.61	1.93	2.17	2.36	2.15
CRN [17]	1.92	2.22	2.49	2.21	2.49	2.70	2.89	2.69	1.99	2.22	2.40	2.20
CRN-BLSTM	1.93	2.23	2.50	2.22	2.51	2.72	2.90	2.71	2.01	2.25	2.44	2.23
McbNet	**2.10**	**2.35**	**2.59**	**2.35**	**2.67**	**2.85**	**3.01**	**2.84**	**2.12**	**2.32**	**2.52**	**2.32**

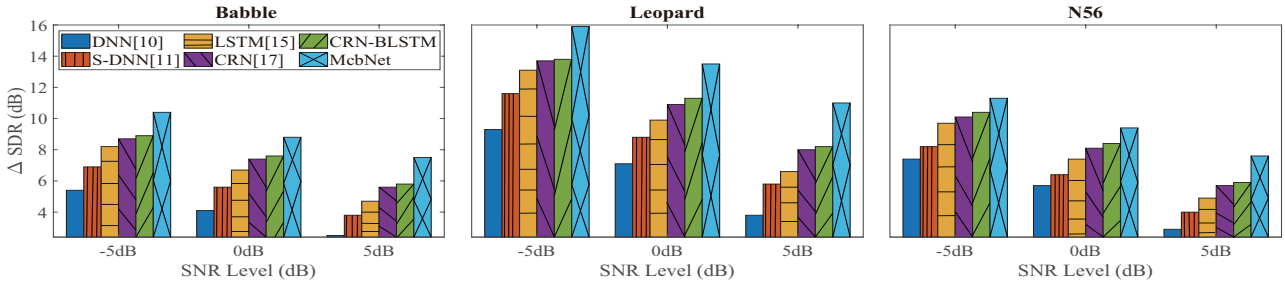

Fig. 3: Speech enhancement performance comparison in terms of ΔSDR for three types of noises with different methods and SNR levels. Each result is the average value of 100 experiments.

generalization ability over the DNN and S-DNN. Unlike the DNN and S-DNN, the LSTM receives not only sequence of the previous layer, but also the hidden state of previous block within the same layer. Therefore, the LSTM exploits combined information of past and current temporal frames.

The CRN obtains, on average, 8.70dB ΔSDR, 80.98% STOI and 2.36 PESQ, which provides higher improvements over the DNN, S-DNN and LSTM methods. Since the local spatial patterns of the input magnitude spectrum are captured by CRN, it is capable of leveraging the T-F structure of magnitude spectrum. Moreover, the LSTM layers inside the CRN exploit the temporal dependency by using past and current temporal frames. The CRN-BLSTM offers improvements over the CRN, because the past, current, future temporal frames are utilized by BLSTM layers.

The proposed McbNet gets the highest improvements over the baseline methods, and it obtains, on average, 10.59dB ΔSDR, 83.05% STOI and 2.50 PESQ. The McbNet gets almost 1.89dB ΔSDR, 2.07% STOI improvement and 0.14 PESQ improvement over the CRN method. The McbNet using the MC to encode the input magnitude spectrum in different scales. The local interdependency is captured by the convolutional sub-layers with small kernel sizes. The convolutional sub-layers with large kernel sizes is used to find the interdependency between the remote frames. By using the small and large size filters, the receptive field of McbNet is enlarged and the T-F structure of the magnitude spectrum is leveraged. Furthermore, the BLSTM layers are introduced to connect the MCE and MCD, which are capable of exploiting the interdependency of past, current and future temporal frames. Besides, the raw data is fed to the output layer of the McbNet to learn the residual mapping relation.

For the noise-independent case (N56), the McbNet provides consistent improvement over baseline methods, which shows that the McbNet has better generalization ability to unseen speaker and noise.

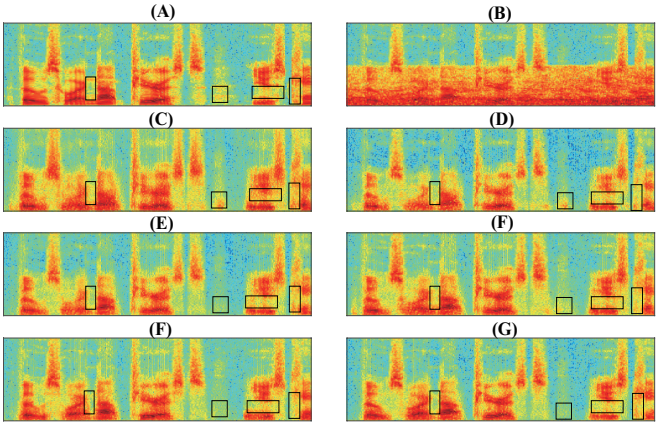

Fig. 4: Spectrograms of different signals: (A) clean speech; (B) noisy speech mixture; (C) enhanced speech by DNN [10]; (D) enhanced speech by S-DNN [11]; (E) enhanced speech by LSTM [15]. (F) enhanced speech by CRN [17]. (G) enhanced speech by CRN-BLSTM. (H) enhanced speech by the proposed McbNet.

4. CONCLUSIONS

In this paper, an McbNet framework was proposed to solve speaker independent single channel speech enhancement with *seen* and *unseen* noises. The McbNet with varied kernel sizes was used to find interdependencies between the temporal frames. More specifically, the small size kernel was applied to capture the local information, which includes

interdependency of adjacent temporal frames. Moreover, large size kernel was used to capture the global information, which contains interdependency between long-term temporal frames. Furthermore, the BLSTM layers and residual learning were introduce to utilize maximum information flow. The experimental results show that the proposed McbNet enlarged the receptive fields and outperformed the state-of-the-art methods with unseen speakers and noise.

5. REFERENCES

[1] A. Hyvarinen and E. Oja, *Independent Component analysis*. Wiley Press, 2001.

[2] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Processing Magazine*, vol. 19, no. 7, pp. 2125 – 2136, 2011.

[3] M. I. Mandel, R. J. Weiss, and D. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[4] M. S. Salman, S. M. Naqvi, A. Rehman, W. Wang, and J. A. Chambers, "Video-aided model-based source separation in real reverberant rooms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1900–1912, 2013.

[5] D. L. Wang, "Deep learning reinvents the hearing aid," *IEEE Spectrum*, vol. March Issue, pp. 32–37, 2017.

[6] P. C. Loizou, *Speech enhancement : theory and practice*. CRC Press, 2013.

[7] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 895–910, 2010.

[8] Y. Sun, Y. Xian, W. Wang, and S. M. Naqvi, "Monaural source separation in complex domain with long short-term memory neural network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 359–369, 2019.

[9] Y. X. 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, no. 12, pp. 1849–1858, 2014.

[10] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[11] M. Tu and X. X. Zhang, "Speech enhancement based on deep neural networks with skip connections," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[12] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002.

[13] Y. Jiang, D. L. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2112–2121, 2014.

[14] Y. Sun, W. Wang, J. A. Chambers, and S. M. Naqvi, "Two-stage monaural source separation in reverberant room environments using deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, no. 1, pp. 125–139, 2019.

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

[16] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1993–1997, 2017.

[17] K. Tan and D. L. Wang, "A convolutional recurrent neural network for real-time speech enhancement," *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3299–3233, 2018.

[18] A. L. Mass, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," *International Conference on Machine Learning (ICML)*, 2013.

[19] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic phonetic continuous speech corpus cdrom," 1993.

[20] G. N. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation." *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 2067–2079, 2010.

[21] A. Varga and H. 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 Commun*, no. 12, pp. 247–251, 1993.

[22] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transanctions on Audio Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[23] 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, 2008.

[24] 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.