AN ATTENTION-BASED NEURAL NETWORK APPROACH FOR SINGLE CHANNEL SPEECH ENHANCEMENT

Xiang Hao[⋆]*, Changhao Shan*[⋆]*, Yong Xu*[†]*, Sining Sun*[⋆]*, Lei Xie*[⋆]

[⋆]Shaanxi Provincial Key Laboratory of Speech and Image Information Processing,
School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[†]Tencent AI Lab, Bellevue, USA

ABSTRACT

This paper proposes an attention-based neural network approach for single channel speech enhancement. Our work is inspired by the recent success of attention models in sequence-to-sequence learning. It is intuitive to use attention mechanism in speech enhancement as humans are able to focus on the important speech components in an audio stream with "high attention" while perceiving the unimportant region (e.g., noise or interference) in "low attention", and thus adjust the focal point over time. Specifically, taking noisy spectrum as input, our model is composed of an LSTM based encoder, an attention mechanism and a speech generator, resulting in enhanced spectrum. Experiments show that, as compared with OM-LSA and the LSTM baseline, the proposed attention approach can consistently achieve better performance in terms of speech quality (PESQ) and intelligibility (STOI). More promisingly, the attention-based approach has better generalization ability to unseen noise conditions.

Index Terms— speech enhancement, neural networks, attention mechanism

1. INTRODUCTION

Speech enhancement, aiming to improve speech quality and intelligibility, has been widely investigated for many years in both academia and industry [1]. This technique is desired in many applications, e.g., mobile telecommunication, hearing aids and automatic speech recognition, etc. In these applications, the hearing experience or speech recognition performance will be degraded drastically when noise exists. In this paper, we study the single channel speech enhancement problem in which spatial information is unavailable, making the problem more challenging.

Over the past several decades, a large variety of algorithms have been proposed. In general, these algorithms can be classified into two categories, namely statistical-based approaches and data-driven approaches. Statistical-based techniques include spectral subtraction [2], Wiener filtering [3], the minimum mean-square error log-spectral method [4], etc. Typical data-driven approaches include non-negative matrix factorization (NMF) [5] and neural network (NN). Recently, with the development of deep learning (DL), deep neural network (DNN) have become a popular method.

DNN-based speech enhancement can be regarded as a regression or prediction task. Given a set of manually-prepared clean-noisy speech pairs, a neural network learns to transform noisy magnitude spectra to their clean equivalents [6, 7] or corresponding masks such as ideal binary mask (IBM) [8] and ideal ratio mask (IRM) [9]. Various DNN structures have been explored, such as deep autoencoder [10], convolutional neural network (CNN) [11], long short term memory (LSTM) network [12] and their combinations [13]. Popular machine learning strategies have also been employed, including multi-task learning [14], progressive learning [15] and reinforcement learning [16]. Several studies combine the spectral and phase information when the algorithm re-synthesizes the predicted features back into time-domain waveforms [17]. As another option of input, raw waveforms can be directly fed into a neural network as well. In this way, CNN [18], generative adversarial network (GAN) [19] and the popular audio generation model, WaveNet [20], have been studied. In the above DNN-based approaches, the loss function of most neural networks is mean-square error (MSE). Recently, perceptual evaluation criteria, e.g. PESQ, has been used as the optimization objective [21] as well.

In this paper, we explore the *attention*-based neural network structures for improving the performance of speech enhancement. Our work is inspired by the recent success of attention models in various sequence-to-sequence learning tasks, including machine translation [22], speech recognition [23] and keyword spotting [24, 25]. It is intuitive to use attention mechanism in speech enhancement: humans are able to focus on a certain region of an audio stream with "high attention" (e.g., the target speech) while perceiving the surrounding audio (e.g., noise or interference) in "low attention", and then adjust the focal point over time. Specifically, we adopt attention mechanism on LSTM-RNN models which have shown superior speech enhancement performances by modeling important sequential information [12]. While the RNN model learns weights of past input features *implicitly* when predicting enhanced frame, the attention mechanism calculates correlations between past frames and the current frame to be enhanced and give weights to past frames *explicitly*. Experiments show that, as compared with the LSTM baseline, the proposed attention approach can consistently achieve better performance in terms of speech quality (PESQ) and intelligibility (STOI). More promisingly, the attention-based approach has better generalization ability to unseen noise conditions.

2. ATTENTION-BASED SPEECH ENHANCEMENT

2.1. System Overview

The schematic diagram of our attention-based approach is illustrated in Fig 1. The input is $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T]$, where $\mathbf{x}_t$ represents the magnitude spectrum of noisy speech at frame t and T represents the total number of frames of a speech segment. Specifically, our system contains three modules: the encoder, the attention

This research work is supported by the National Natural Science Foundation of China (No.61571363).

⋆Lei Xie is the corresponding author.

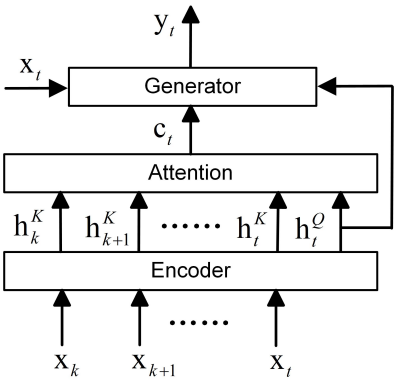

Fig. 1. Schematic diagram of the proposed attention-based model for speech enhancement.

mechanism and the generator.

The encoder extracts a high-level feature representation $\mathbf{h}$ from the input speech feature $\mathbf{x}$:

$$\mathbf{h}^Q, \mathbf{h}^K = \text{Encoder}(\mathbf{x}), \tag{1}$$

where the $\mathbf{h}^Q$ and $\mathbf{h}^K$ are the query and key, respectively [26]. In our work, we adopt LSTM as the encoder that has strong sequential modeling ability leading to superior performances in speech enhancement [12]. Then the attention mechanism takes the query and key as input and forms a fixed-length context vector:

$$\mathbf{c}_t = \text{Attention}(\mathbf{h}^K, \mathbf{h}_t^Q). \tag{2}$$

Finally, the enhanced speech $\mathbf{y}$ is the output of the generator which takes the context vector $\mathbf{c}_t$, the noisy speech $\mathbf{x}_t$ and the encoder's output $\mathbf{h}_t^Q$ as inputs:

$$\mathbf{y}_t = \text{Generator}(\mathbf{c}_t, \mathbf{h}_t^Q, \mathbf{x}_t). \tag{3}$$

We detail the encoder, the attention mechanism and the generator in the following subsections.

2.2. Encoder

The encoder aims to give a high-level representation from the input feature. As shown in Fig 2, we explore two different structures for the encoder: *expanded* and *stacked*. The difference between the two encoders lies in the generation of query $\mathbf{h}^Q$.

For the expanded encoder as shown in Fig 2 (a), the input feature $\mathbf{x}_t$ is first fed into a fully-connected layer

$$\hat{\mathbf{x}}_t = \tanh(\mathbf{W}_s \mathbf{x}_t + \mathbf{b}_s), \tag{4}$$

Then, $\hat{\mathbf{x}}_t$ is regarded as the input of the LSTM cell and

$$\mathbf{h}_t^K = f(\hat{\mathbf{x}}_t), \tag{5}$$

where $f(\cdot)$ represents function of LSTM cell and $\mathbf{h}_t^K$ is the output of LSTM cell. Through the same process, we can get the output $\mathbf{h}_t^Q$ as well.

Differently, in the stacked encoder as shown in Fig 2 (b), $\mathbf{h}_t^Q$ is calculated as

$$\mathbf{h}_t^Q = f(\mathbf{h}_t^K). \tag{6}$$

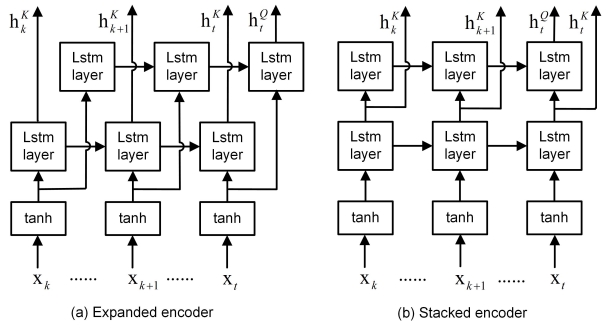

Fig. 2. Two types of encoder structure.

2.3. Attention Mechanism

Recall that the attention mechanism forms a context vector $\mathbf{c}$ that contains the information of key $\mathbf{h}^K$. Notably, the attention mechanism can use both past and future frames of a time sequence. In practice, we usually treat speech enhancement as a *casual* problem, or we only use the past frames to avoid any latency.

Here we use two strategies: *causal dynamic attention* and *causal local attention*. Specifically, in the former, if we denoise a frame, $\mathbf{x}_t$, we calculate attention weights using $[\mathbf{x}_1, \cdots, \mathbf{x}_t]$. This means we consider all the past frames when we enhance the current frame. If the utterance is too long, the attention weights of many previous frames may be nearly zero. So, in the latter, if we enhance the current input $\mathbf{x}_t$, we only use $[\mathbf{x}_{t-w}, \cdots, \mathbf{x}_t]$ to calculate the attention weight, where the value of w can be set as a constant.

Thus, a normalized attention weight α is learned:

$$\alpha_{tk} = \frac{\exp(\text{score}(\mathbf{h}_k^K, \mathbf{h}_t^Q))}{\sum_{k=n}^{t} \exp(\text{score}(\mathbf{h}_k^K, \mathbf{h}_t^Q))}. \tag{7}$$

??

When we use causal dynamic attention, $n = 1$. For causal local attention, $n = (t - w)$. We follow the correlation calculation in [27], so $\text{score}(\mathbf{h}_k^K, \mathbf{h}_t^Q) = \mathbf{h}_k^{K\top} \mathbf{W} \mathbf{h}_t^Q$. Finally, we compute the context vector as the weighted average of $\mathbf{h}_k^K$:

$$\mathbf{c}_t = \sum_{k=n}^{t} \alpha_{tk} \mathbf{h}_k^K. \tag{8}$$

2.4. Generator

The output of the speech enhancement system is generated using the context vector $\mathbf{c}_t$, the encoder's output $\mathbf{h}_t^Q$ and the input feature $\mathbf{x}_t$. Given $\mathbf{c}_t$ and $\mathbf{x}_t$, we first learn an enhancement vector $\mathbf{e}_t$:

$$\mathbf{e}_t = \tanh(\mathbf{W}_e[\mathbf{c}_t; \mathbf{h}_t^Q] + \mathbf{b}_e), \tag{9}$$

where the $[\cdot;\cdot]$ denotes the concatenation of two vectors. Finally, similar to [28], we form a 'hidden' mask of the input feature $\mathbf{x}_t$, and the final enhanced speech $\mathbf{y}_t$ is

$$\mathbf{y}_t = \mathbf{x}_t \odot \text{sigmoid}(\mathbf{W}_m \mathbf{e}_t + \mathbf{b}_m). \tag{10}$$

3. EXPERIMENTS

3.1. Datasets

To evaluate the performance of the proposed method, we create some synthetic datasets. The conditions of all the datasets are summarized

Table 1. Conditions of noise and SNR for datasets. Utterances in Train, Valid and Test-0 are from the same multi-speaker set; while utterances in Test-1,2,3 and 4 from another set of speakers.

Set	Train	Valid	Test-0	Test-1	Test-2	Test-3	Test-4
Noise	Musan	Musan	Musan	Musan	Musan	CHIME3	CHIME3
SNR	0-20dB	0-20dB	0-20dB	0-20dB	-5-0dB	0-20dB	-5-0dB

in Table 1. Specifically, first we randomly select a clean speech file from a multiple-speaker speech corpus which has 21407 utterances (about 24.5 hours) and a noise file from the Musan corpus [29]. Then we randomly select an SNR between 0dB and 20dB, and mix these two files to create a noisy file according to the selected SNR. We divide the generated 21407 noisy files into 13407, 4000 and 3000 for training, validation and testing. Neural network training is conducted using the training set and the loss on the validation set is examined as the convergence condition. The original testing set is named as Test-0. We further create 4 new test sets (Test-1,2,3,4) each with 3000 utterances from another set of speakers for real scenarios. In Test-1,2, interferer noises are from the noise pool, Musan, as Test-0. These Test-3,4 are generated by adding different noises (with the training set) from CHIME3 [30]. The SNR of these test sets are illustrated as Table 1.

3.2. Experimental Setups

The sampling rate of the speech files is 16 kHz. We use short-time Fourier transform (STFT) to extract the spectrum from each speech utterance. A Hanning window with 512 points and overlap interval of 128 are used. The fast Fourier transform (FFT) is taken at 512 points and the first 257 FFT points are used as our spectral feature. Hence for all neural networks in this paper, the input and the output are both magnitude spectrum with 257 dimensions. In the evaluation stage, we combine the output of neural network and phase information calculated from the corresponding noisy speech and use inverse Fourier transform (IFT) to get the enhanced wave file.

We compare our approach with OM-LSA [31] and an LSTM approach without attention mechanism. Here as a typical statistical method, OM-LSA is treated as the baseline. The LSTM has two layers and each layer has 128/256/512 cells. Note that in order to make performance comparison when the models have similar size of parameters, the cell size of the LSTM encoder in the proposed attention approach is set to 112/224/448, accordingly. These models are initialized with the normalized initialization and trained using Adam [32]. The loss function used is mean square error (MSE). All parameters are learned automatically when we train the neural network. The learning rate is set to 0.0005 at the beginning and decays with a rate of 0.5 when the loss of current epoch is greater than the previous epoch. The batch size is 128. Dropout regularization is also used [33]. Perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) are used as evaluation criteria.

3.3. Experiment Results

We first analyze the performance of the baseline methods and the proposed method on Test-0 using casual dynamic attention. The results of PESQ and STOI are summarized in Table 2. We can clearly see that the attention-based method outperforms the two baselines consistently for different size of parameters, which indicates that introducing attention mechanism to neural network based speech enhancement is beneficial. We also notice that the stacked encoder structure has slightly better performance than the expanded encoder structure but the difference is not significant.

Table 2. PESQ and STOI (in percent) of different models using causal dynamic attention on Test-0.

rnn-size	LSTM		att-expanded		att-stacked		OM-LSA		NOISY	
(lstm/att)	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
128/112	2.860	92.2	2.906	93.0	**2.928**	**93.1**	2.688	90.5	2.492	90.6
256/224	2.957	93.1	3.000	93.8	**3.016**	**93.9**				
512/448	3.014	93.6	3.063	94.2	**3.084**	**94.3**				

Table 3. PESQ and STOI (in percent) of different models using causal local attention on Test-0.

w	rnn-size	LSTM		att-expanded		att-stacked		OM-LSA		NOISY	
	(lstm/att)	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
5	128/112	2.860	92.2	2.928	93.2	2.905	93.0	2.688	90.5	2.492	90.6
	256/224	2.957	93.1	3.019	93.8	3.013	93.8				
	512/448	3.014	93.6	3.104	94.4	**3.105**	**94.4**				
15	128/112	2.860	92.2	2.913	93.1	2.879	93.0				
	256/224	2.957	93.1	3.025	93.9	3.010	93.8				
	512/448	3.014	93.6	3.077	94.3	**3.103**	**94.4**				
30	128/112	2.860	92.2	2.918	93.1	2.920	93.1				
	256/224	2.957	93.1	3.015	93.8	3.009	93.9				
	512/448	3.014	93.6	3.080	94.3	**3.091**	**94.4**				

In the experiments for causal local attention, the value of w was set to 5, 15 and 30. The results of PESQ and STOI on Test-0 are shown in Table 3. Previous observation is still valid for causal local attention: the attention-based model outperforms the baseline systems. We also notice that bigger value of w for causal local attention results in no further improvement and the best performance is achieved when $w = 5$. At the same time, comparing Table 2 and Table 3, the causal local attention models have comparative or even better performance than the causal dynamic attention models. The above two observations verify the conjecture that we do not need consider very long historical information in the speech enhancement task. This is reasonable because noise conditions (types and SNRs) may change time to time.

To make an intuitive understanding of the attention mechanism, we have visualized the attention weights of the two attention-based approaches with stacked encoder ($w = 5$) on a testing speech clip (100 frames), as shown in Fig 3. The x axis represents $\mathbf{h}^K$ and the y axis represents $\mathbf{h}^Q$. The point, (x, y) represent attention weight. We can see that the attention-based model gives different weights (or different 'attention levels') to the contextual frames explicitly during speech enhancement. To further showcase the speech enhancement ability of different methods, a speech utterance (spectrum) from Test-0 is shown in Fig 4. The original speech is contaminated with 'dog-barking' noise. The traditional OM-LSA method cannot handle this kind of non-stationary noise properly: most of the noise components still exist and some of the speech components are mistakenly removed. Although the LSTM approach can significantly remove the noise components in this example, there still exists noise residuals (in the marked rectangular). On the contrast, the attention-based method can remove the noise properly and the speech components are almost completely restored. Some of our testing examples can be found from `https://xhoare.github.io/index.html`.

Experimental results on Test-0,1,2,3,4 are summarized in Table 4, which shows the generalization capability of different approaches. Here for the attention model, the casual local attention

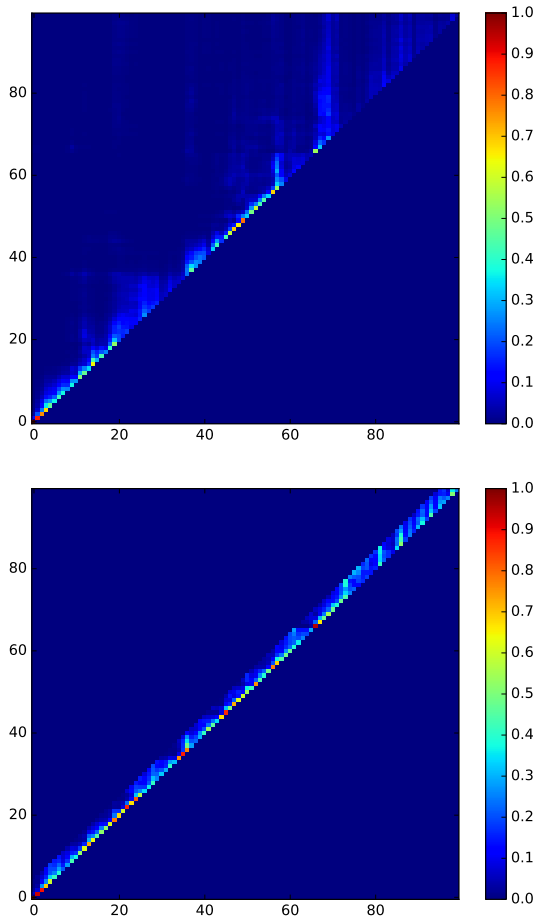

Fig. 3. Visualization of the alignment weights on a test speech clip. Above: casual dynamic attention; Below: causal local attention ($w = 5$). Stacked encoder is used and LSTM cell size is 448.

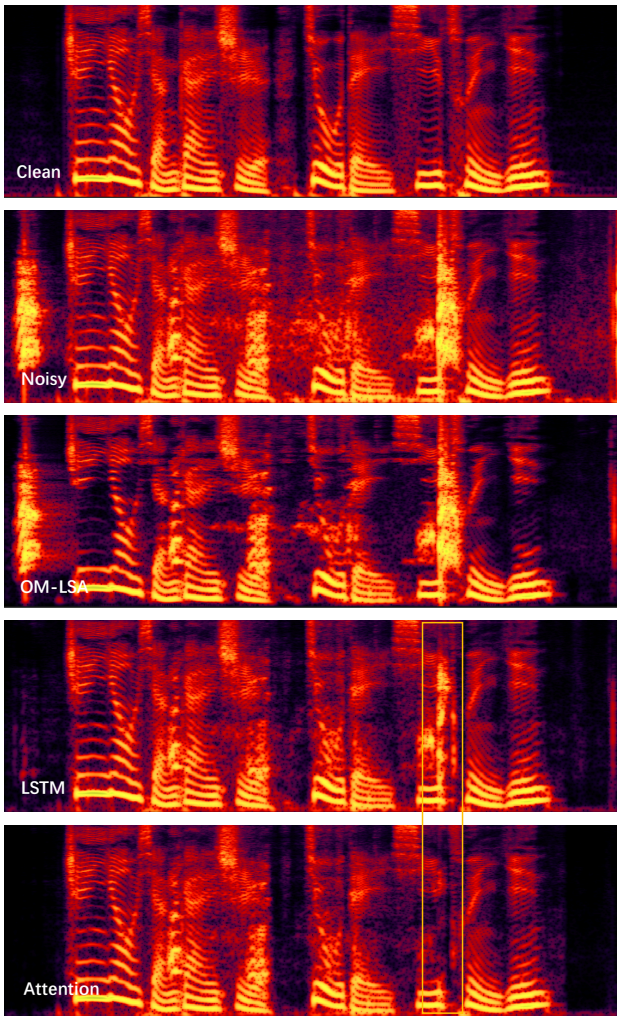

Fig. 4. A test example before and after speech enhancement.

strategy is used and the attention window $w = 5$. First, we can see that, on all the test sets, the neural network methods (LSTM and attention) achieve superior performances over the traditional OM-LSA method and attention-based models consistently achieve the best performance. We also notice that all the neural network models have performance degradation when tested on the mismatched sets (Test-1,2,3,4 vs. Test-0). Moreover, the models trained on the data with 0-20dB SNR have significant performance degradation on the test set with -5-0dB SNR (Test-2). Further with both mismatched noises and SNR conditions, large performance degradation can be observed on Test-4. But among all the methods, our attention models have shown better generalization ability. For example on Test-4, the attention model with stacked encoder has achieve 0.173 and 0.295 absolute PESQ gains as compared with the ordinary LSTM model and the OM-LSA method, respectively. These results further prove that the proposed method using attention mechanism is promising in improving speech enhancement performance.

Table 4. PESQ and STOI (in percent) of different approaches on Test-1,2,3,4. Causal local attention is used for attention approach and $w = 5$. LSTM cell size is 512 and 448 for LSTM model and attention model, respectively.

Test set	NOISY		OM-LSA		LSTM		att-expanded		att-stacked	
	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
Test-0	2.492	90.6	2.688	90.5	3.014	93.6	3.063	94.2	**3.084**	**94.3**
Test-1	2.270	86.1	2.611	87.1	2.960	91.1	**3.061**	**91.9**	3.050	91.9
Test-2	1.256	72.9	1.596	73.1	2.049	78.3	2.111	78.6	**2.140**	**79.3**
Test-3	2.099	83.6	2.746	85.6	2.866	89.7	**2.970**	**90.7**	2.962	90.7
Test-4	0.963	70.5	1.684	70.8	1.806	73.9	1.931	75.7	**1.979**	**76.4**

4. CONCLUSION

In this work, we have shown the promising ability of introducing attention mechanism in neural network based speech enhancement. As compared with OM-LSA and an ordinary LSTM, the proposed approach can consistently achieve better performance in terms of PESQ and STOI. More promisingly, the attention-based approach has better generalization ability to unseen noise conditions. In the future, we will explore other new attention mechanisms [34] in the speech enhancement task. We will also study the abilities of the attention models in dealing with different types of noises.

5. REFERENCES

[1] Philipos C Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Inc., 2007.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1979.

[3] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *ICASSP*, 1996.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1984.

[5] Hao Teng Fan, Jeih Weih Hung, Xugang Lu, Syu Siang Wang, and Tsao Yu, "Speech enhancement using segmental nonnegative matrix factorization," in *ICASSP*, 2014.

[6] Yong Xu, Jun Du, Li Rong Dai, and Chin Hui Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, 2013.

[7] Yong Xu, Jun Du, Li Rong Dai, and Chin Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2015.

[8] Yuxuan Wang and De Liang Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, 2013.

[9] Arun Narayanan and De Liang Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *ICASSP*, 2013.

[10] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *INTERSPEECH*, 2013.

[11] Se Rim Park and Jinwon Lee, "A fully convolutional neural network for speech enhancement," *Evaluation*, 2017.

[12] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *Journal of the Acoustical Society of America*, 2017.

[13] Han Zhao, Shuayb Zarar, Ivan Tashev, and Chin Hui Lee, "Convolutional-recurrent neural networks for speech enhancement," in *ICASSP*, 2018.

[14] Yong Xu, Jun Du, Zhen Huang, Li Rong Dai, and Chin Hui Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2017.

[15] Tian Gao, Jun Du, Li Rong Dai, and Chin Hui Lee, "Densely connected progressive learning for lstm-based speech enhancement," in *ICASSP*, 2018.

[16] Yuma Koizumi and et al, "Dnn-based source enhancement self-optimized by reinforcement learning using sound quality measurements," in *ICASSP*, 2017.

[17] Kuldip Paliwal, Kamil Wjcicki, and Benjamin Shannon, "The importance of phase in speech enhancement," *Speech Communication*, 2011.

[18] Szu Wei Fu, Tsao Yu, Xugang Lu, and Hisashi Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2017.

[19] Santiago Pascual, Antonio Bonafonte, and Joan Serr, "Segan: Speech enhancement generative adversarial network," in *INTERSPEECH*, 2017.

[20] Dario Rethage, Jordi Pons, and Xavier Serra, "A wavenet for speech denoising," in *ICASSP*, 2018.

[21] Yan Zhao, Buye Xu, Girl Ritwik, and Zhang Tao, "Perceptually guided speech enhancement using deep neural networks," in *ICASSP*, 2018.

[22] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[23] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philmon Brakel, and Yoshua Bengio, "End-to-end attention-based large vocabulary speech recognition," in *ICASSP*, 2016.

[24] Changhao Shan, Junbo Zhang, Yujun Wang, and Lei Xie, "Attention-based end-to-end speech recognition on voice search," in *ICASSP*, 2017.

[25] Changhao Shan, Junbo Zhang, Yujun Wang, and Lei Xie, "Attention-based end-to-end models for small-footprint keyword spotting," in *ICASSP*, 2018.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[27] Minh Thang Luong, Hieu Pham, and Christopher D Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[28] Po Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2015.

[29] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[30] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third chimespeech separation and recognition challenge: Dataset, task and baselines," in *Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015.

[31] Cohen I., "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, 2003.

[32] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, 2014.

[34] Nitish Shirish Keskar Ahmed Karim and Socher Richard, "Weighted transformer network for machine translation," *arXiv preprint arXiv:1711.02132*, 2017.