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

A Noise-Aware Memory-Attention Network Architecture for Regression-Based Speech Enhancement

Yu-Xuan Wang[1]*, Jun Du*[1]*, Li Chai*[1]*, Chin-Hui Lee*[2]*, Jia Pan*[1]

[1]University of Science and Technology of China, Hefei, Anhui, P.R.China
[2]Georgia Institute of Technology, Atlanta, GA, USA

yxwang1@mail.ustc.edu.cn, jundu@ustc.edu.cn, cl122@mail.ustc.edu.cn,
chl@ece.gatech.edu, jiapan@iflytek.com

Abstract

We propose a novel noise-aware memory-attention network (NAMAN) for regression-based speech enhancement, aiming at improving quality of enhanced speech in unseen noise conditions. The NAMAN architecture consists of three parts, a main regression network, a memory block and an attention block. First, a long short-term memory recurrent neural network (LSTM-RNN) is adopted as the main network to well model the acoustic context of neighboring frames. Next, the memory block is built with an extensive set of noise feature vectors as the prior noise bases. Finally, the attention block serves as an auxiliary network to improve the noise awareness of the main network by encoding the dynamic noise information at frame level through additional features obtained by weighing the existing noise basis vectors in the memory block. Our experiments show that the proposed NAMAN framework is compact and outperforms the state-of-the-art dynamic noise-aware training approaches in low SNR conditions.

Index Terms: attention mechanism, memory block, noise-aware training, LSTM-RNN, speech enhancement

1. Introduction

Single-channel speech enhancement (SE) is a widely studied problem in signal processing which aims at enhancing noisy speech to improve speech quality and intelligibility [1]. Notable conventional algorithms include spectral subtraction [2, 3], Wiener filtering [4, 5], MMSE estimator [6, 7], and OM-LSA speech estimator [8]. In recent years, most supervised SE techniques have been based on deep neural network (DNN) architectures [9], which show strong regression capabilities of mapping from the input noisy log-power spectra (LPS) features to the target clean LPS features. Although DNN-based SE algorithms have achieved considerable success, more and more research efforts are made to further improve the speech enhancement performance.

On the one hand, due to the fully-connected structure, DNN cannot fully utilize the relationship between the neighbouring frames under long-term acoustic contexts even with the help of frame expansion [9, 10]. As an alternative, long short-term memory recurrent neural network (LSTM-RNN) makes a full use of the information between the current and the previous frames by adding the memory cells and a series of "gates" to determine the retention and deletion information of previous frames [11, 12]. LSTM-RNN also achieves better generalization at low signal-to-noise ratios (SNRs) than DNN [13, 14]. More recently, inspired by the success of attention models in various sequence-to-sequence learning tasks [15, 16, 17], an attention mechanism can also be added to LSTM-RNN [18] or bidirectional long short term memory (BLSTM) [19] for the SE task. It is proved to have a better generalization ability. Besides LSTM-RNN, other powerful structures, such as convolutional neural network (CNN) [20], convolutional-recurrent neural network (CRNN) [21], generative adversarial network (GAN) [22], have also been proposed.

On the other hand, it is noted that the DNN performance deteriorates when a mismatch exists between the training and testing sets [23]. Many noise types have been added to the training set to resolve this issue in [24], but it cannot always improve the speech quality. Noise-aware training (NAT) attains state-of-the-art noisy speech recognition results on the Aurora-4 task [25], and has been applied successfully to speech enhancement. Static noise aware training (SNAT) predicts the noise information, and appends the same information to each frame by assuming that the noise signal during the whole utterance is stationary [10]. However noise is changing greatly in most realistic environments. Accordingly dynamic noise aware training (DNAT) estimates the noise signal in a dynamic manner, and is able to deal with the non-stationary scenes [26]. Further improvements and deformations, such as post-processing, turning full-band features into sub-band features and interpolating SNAT & DNAT [27], are considered as DNAT extensions. Similarly, an SNR-aware model is adopted to predict SNR levels [28], and speaker-aware denoising autoencoder (SaDAE) predicts the speaker identities [29]. Other studies in [30] and [31] use the framework of denoising auto-encoders (DAE) to learn the transformation, but follow almost the same idea as DNAT.

In this paper, we propose a novel noise-aware memory-attention network (NAMAN) for single-channel speech enhancement. Unlike the way attention model embedded into the backbone of the neural network structure [18], we utilize the attention mechanism in a side branch, which is designed to learn the similarities between the current frame and the existing noise basis vectors in the static memory block instead of the previous frames. The clustered acoustic feature vectors, namely Mel-frequency cepstral coefficients (MFCCs) of noise signals, are extracted as the prior noise information and stored in the memory block. The dynamically predicted noise features are obtained by combining the weights learned from the attention mechanism and MFCCs in the memory block together, and are then attached to the noisy features during training. With the help of memory block, our noise-aware training is carried out jointly with the process of denoising. This one-stage model training design significantly simplifies the complicated two-stage design of DNAT [26]. Moreover, DNAT can achieve a good performance over DNN, but when it turns to LSTM-RNN, which has a more powerful modeling ability on the acoustic context of neighboring frames, the performance gain is less significant. The experimental results show that our NAMAN model can still maintain significant improvements under the LSTM-RNN setting.

http://dx.doi.org/10.21437/Interspeech.2020-2037

2. Proposed Deep NAMAN Architecture

Figure 1 illustrates the NAMAN structure consisting of the main network, the memory block and the attention block. The key of the proposed framework is to generate predicted features incorporating the noise information embedded in the current noisy speech frame by a weighted combination of the noise basis vectors in the memory block. With the help of the attention mechanism and LSTM-RNN, the predicted noise vectors can provide useful information for speech enhancement. The details are elaborated in the following subsections.

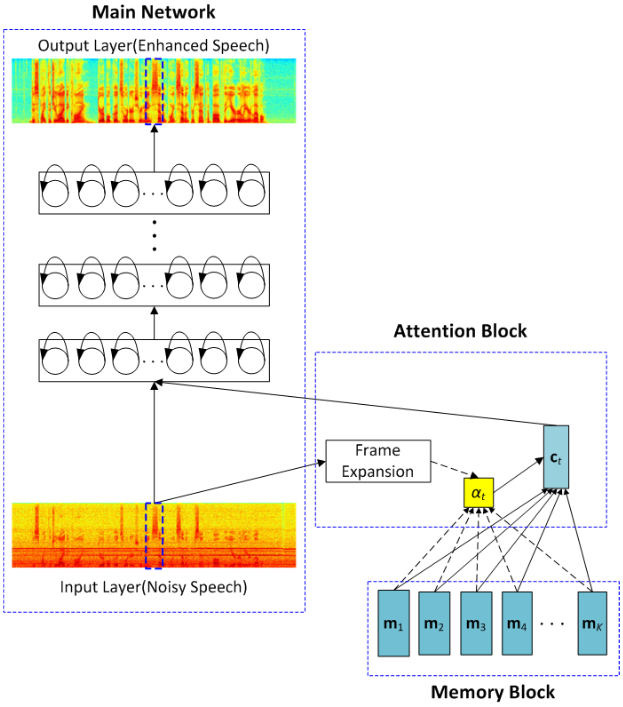

Figure 1: *The structure of NAMAN, including the main network, the attention block and the memory block.*

2.1. Main Speech Enhancement Regression Network

The main network has two effects on the whole framework: denoising and exchanging information with the attention block. Denoising aims to remove the noise from noisy speech to get the enhanced speech. On the other hand, the attention block needs the noise information to pick up the most relevant vectors from the memory block. With the layers increasing, the noise is removed gradually. Here, we append the output from the attention block as auxiliary noise information to the input features to be fed into the NAMAN input layer for subsequent processing.

Given a noisy utterance with T frames, the input noisy LPS features are represented by

$$X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}, \tag{1}$$

where $\mathbf{x}_t$ denotes the noisy feature vector at frame t. Here LSTM-RNN is adopted as the main network for its congenital advantage of sequence representation and temporal contexts acquisition. A detailed calculation in the LSTM-RNN cells is implemented as follows:

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i), \tag{2}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f), \tag{3}$$

$$\mathbf{c}_t = \mathbf{f}_t \otimes \mathbf{c}_{t-1} + \mathbf{i}_t \otimes \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c), \tag{4}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o), \tag{5}$$

$$\mathbf{h}_t = \mathbf{o}_t \otimes \tanh(\mathbf{c}_t), \tag{6}$$

where **i**, **f**, **o** represent the “input gate”, “forget gate” and “output gate”, respectively. **c** is the cell activation vector, and **h** is the hidden vector. **W** and **b** stand for the weight matrices and bias vectors from the cell to gate. σ is the logistic sigmoid function, and $\otimes$ denotes element-wise multiplication. The corresponding outputs of the l-th hidden layer of the main network are:

$$H^l = \{\mathbf{h}_1^l, \mathbf{h}_2^l, ..., \mathbf{h}_T^l\}, \tag{7}$$

where $\mathbf{h}_t^l$ denotes the output of the l-th hidden layer at frame t.

2.2. Noise-Basis Memory Block

The memory block provides the prior noise information for the attention block. It consists of an extensive set of basis vectors, which contain rich noise information and can represent a new noise type by weighted combination. Moreover, the vectors are bound to be quite distinguishable from each other by its corresponding noise type. In view of the random and abrupt nature of noise signals, we adopt the frame-level MFCC features extracted from noise signals.

The procedure of memory block generation is illustrated in Algorithm 1. First, we need to collect different types of noise waveforms from different environments, such as fax machine noises, car idling, footsteps, paper rustling, rain, animal noises, etc. Second, we cut the noise waveforms into frames and extract the MFCC features. Next, based on those noise MFCC feature frames, we can cluster them to form a compact set of K distinguishable noise basis vectors, using a K-means algorithm with a cosine distance shown below:

$$d_{\cos}(\mathbf{n}_i, \mathbf{n}_j) = \frac{\mathbf{n}_i \cdot \mathbf{n}_j}{||\mathbf{n}_i|| \cdot ||\mathbf{n}_j||}, \tag{8}$$

where $d_{\cos}(\mathbf{n}_i, \mathbf{n}_j)$ is exactly the cosine distance between $\mathbf{n}_i$ and $\mathbf{n}_j$, and $\mathbf{n}_i$ stands for the i-th noise feature vector. Finally, the K cluster centers are stored as the memory block defined as:

$$\mathbf{M} = \{\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_K\}, \tag{9}$$

where $\mathbf{m}_k$ is the k-th noise basis vector.

Algorithm 1 Procedure of Memory Block Generation.

Step1: Noise Sources Collection
collect different noise types as many as possible.
Step2: Feature Extraction
extract frame-level MFCC features from all noise waveforms.
Step3: Clustering
luster all the noise feature vectors into K clusters.
Step4: Memory Block Generation
form the memory block **M** with the K cluster centroids.

It is noted that the memory vectors are static, and should not be updated during either the training or testing step.

2.3. Memory-Aware Attention Block

The attention block is another important part of the whole architecture, it focuses on selecting the basis vectors from the memory block, which are the most relevant to the noise information embedded in the current speech frame [15]. To gather accurate information for the attention model, not only placing the attention block close to the input, but also performing frame expansion on the input features as follows:

$$\mathbf{f}_t = [\mathbf{x}_{t-\tau}, ..., \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}, ..., \mathbf{x}_{t+\tau}], \tag{10}$$

where $\mathbf{f}_t$ stands for the feature vector after frame expanding at frame t, and τ controls how many history and future frames are involved. Frame expansion is just a simple step which can avoid overfitting and really contribute to collecting noise and responding to mutation of signals.

The attention model takes $\mathbf{f}_t$ and $\mathbf{m}_k$ as input and combines them to a vector with the learned weights. A small neural network is designed to learn the similarity scores between $\mathbf{f}_t$ and $\mathbf{m}_k$, which can be defined by the general formula [32]:

$$e_{t,k} = \mathbf{m}_k^\top \mathbf{W}_a \mathbf{f}_t, \tag{11}$$

where $e_{t,k}$ scores the similarity between $\mathbf{f}_t$ and $\mathbf{m}_k$. The matrix $\mathbf{W}_a$ contains the the attention model parameters. The attention value $\alpha_{t,k}$ is then calculated by $\mathbf{f}_t$ and $\mathbf{m}_k$ through a softmax operation, as shown in the dashed arrows of Figure 1:

$$\alpha_{t,k} = \frac{\exp(e_{t,k})}{\sum_{i=1}^{K} \exp(e_{t,i})}. \tag{12}$$

After normalization, the value $\alpha_{t,k}$ is regarded as a weight, and multiplied by $\mathbf{m}_k$, as shown in the solid arrows of Figure 1:

$$\mathbf{c}_t = \sum_{k=1}^{K} \alpha_{t,k} \, \mathbf{m}_k, \tag{13}$$

where $\mathbf{c}_t$ is the predicted noise vector by the attention model at frame t. So $\mathbf{c}_t$ is a weighted sum of all the basis vectors $\mathbf{m}_k$, and is then concatenated to the input vector:

$$\bar{\mathbf{x}}_t = [\mathbf{x}_t \; \mathbf{c}_t]^\top, \tag{14}$$

where the new vector $\bar{\mathbf{x}}_t$ is fed to the first hidden layer.

2.4. Training and Testing

With the outputs of final LSTM layer shown in Eq. (7), we adopt a linear layer on top of it to generate the outputs of the main network, namely, the LPS features of enhanced speech. Then the parameters of NAMAN are optimized with a minimum mean squared error (MMSE) criterion:

$$E = \frac{1}{T} \sum_{t=1}^{T} ||\hat{\mathbf{s}}_t - \mathbf{s}_t||_2^2, \tag{15}$$

where $\hat{\mathbf{s}}_t$ and $\mathbf{s}_t$ are the t-th LPS feature vectors of estimated and clean reference utterances, respectively.

In the training stage, the parameters in both the main network and the attention block are jointly optimized. In the testing or enhancement stage, the predicted noise vector concatenating to the input of the main network can be obtained by the attention mechanism for each frame to improve the performance of output enhanced speech.

3. Experiments and Result Analysis

3.1. Database

In order to improve the generalization capacity of unseen environments, 958 noise types including 100 noise types [33], 15 home-made noise types and 843 noise types from Free Sound part of the MUSAN corpus [34] were selected as the noise database for training. All 7138 utterances from the training set of WSJ0 corpus were corrupted with the above-mentioned 958 noise types at six levels of SNRs (-5dB, 0dB, 5dB, 10dB, 15dB and 20dB) to build a 36-hour multi-condition training set composed of pairs of clean and noisy speech utterances. Approximately 200 sentences randomly selected from the 36-hour data set were used as the cross-validation set. Similarly, the 330 utterances from the core test set of WSJ0 corpus were used to construct the test set for each combination of noise types and SNR levels (-5dB, 0dB and 5dB). As we only conducted the evaluation of mismatched non-stationary noise types in this study, three unseen noise types, namely Buccaneer1, Destroyer engine and HF channel, were adopted for testing, which were all collected from the NOISEX-92 corpus [35].

3.2. Experimental Setting

As for the front-end, all the speech waveforms were sampled at 16kHz, and the frame length was set to 512 samples with a frame shift of 256 samples. A short-time Fourier transform (STFT) was used to compute the spectra of each overlapping windowed frame. Thus, the 257-dimensional LPS features were produced to train the neural network. Both the input and the reference feature vectors were normalized by global mean and variance before feeding into the networks. In the memory block, the 12-dimensional MFCC features of the noise waveforms in the training set, with their first and second order derivatives, were extracted, and were clustered into 500 classes at the frame level by the K-means algorithm.

For the main network, on top of the input layer there were 2 stacked LSTM layers with projection, each hidden layer had 1024 memory cells and the output layer had 257 units. To make our predictions more accurate, we expanded the input to the attention block 3 frames forward and backward, respectively. All the networks were initialized with random weights. The learning rate for the fine-tuning was set to 0.1 for the first 6 epochs and declined at a rate of 90% after every 6 epochs. Original phase of noisy speech was adopted with the enhanced LPS for the waveform reconstruction.

In this experiment, two other noise-aware models, denoted as SNAT and DNAT, were used for performance comparison. SNAT and DNAT had the same network configurations as our model, i.e. 2 LSTM hidden layers with 1024 cells per layer, other model parameters were consistent with “SNAT” and “DNAT3” in [26]. We also provided the oracle experiment assuming the real noise spectrum was known on the test set as the upper bound, approximatively. The enhancement performance was assessed by using perceptual evaluation of speech quality (PESQ) [36] for measuring speech quality, short-time objective intelligibility (STOI) [37] for measuring speech intelligibility, and log-spectral distortion (LSD) (in dB) [38] for evaluating signal differences in the frequency domain.

3.3. Experimental Results

Table 1 lists the average PESQ, STOI and LSD performance comparison of different models on the test set. “Noisy” denotes noisy speech with no processing. “Mapping” represents the original LSTM-based regression model using the direct mapping approach without noise-aware training, "SNAT" and "DNAT" refer to "SNAT" and "DNAT3" in [26], respectively. "NAMAN" denotes our proposed approach. "Oracle" means the real noise spectrum is known [26]. Three low SNRs (-5dB, 0dB, 5dB) are selected where the enhancement task is hard and necessary to carry out. Both the two improved models (SNAT, DNAT) outperform the direct mapping system (Mapping) on all the three measures and SNR levels, and severely underperform Oracle, which leaves a lot of room to further improve the performance. Besides, NAMAN performs much better than Mapping, achieving an average PESQ gain of 0.177 (from 2.093 to 2.27), an average STOI gain of 0.029 (from 0.77 to 0.799) and an average LSD decrease of 0.635 (from 4.394 to 3.759). NAMAN also yields better results than SNAT, which can not well handle the non-stationary noise types. What's more, even compared with the powerful model DNAT, NAMAN can also keep the consistent superiority, which is more obvious for low SNRs, PESQ improves from 1.683 to 1.828 with the gain of 0.145, STOI increases from 0.67 to 0.696 with the gain of 0.029 at SNR=-5dB.

Table 1: *Performance comparison on the test set at different SNRs of the three unseen noise environments, among: Noisy, Mapping, SNAT, DNAT, NAMAN and Oracle. Ave denotes the average of three SNRs (-5dB, 0dB and 5dB).*

SNR(dB)		-5	0	5	Ave
PESQ	Noisy	1.300	1.509	1.783	1.531
	Mapping	1.592	2.115	2.572	2.093
	SNAT	1.669	2.196	2.606	2.157
	DNAT	1.683	2.214	2.635	2.177
	NAMAN	1.828	2.304	2.677	**2.270**
	Oracle	2.295	2.681	2.985	2.654
STOI	Noisy	0.596	0.714	0.823	0.711
	Mapping	0.650	0.785	0.875	0.770
	SNAT	0.667	0.800	0.879	0.782
	DNAT	0.670	0.806	0.887	0.788
	NAMAN	0.696	0.814	0.887	**0.799**
	Oracle	0.814	0.879	0.924	0.872
LSD	Noisy	15.826	12.228	9.102	12.385
	Mapping	4.846	4.395	3.941	4.394
	SNAT	4.698	3.992	3.565	4.085
	DNAT	4.664	3.755	3.047	3.822
	NAMAN	4.554	3.691	3.031	**3.759**
	Oracle	3.666	3.320	2.985	3.324

Figure 2 shows an utterance example corrupted by Buccaneer1 noise at SNR=0dB. DNAT successfully removes most of the noise in noisy speech. NAMAN not only reconstructs more speech details compared with DNAT (shown in the dashed rectangular boxes), but also restores more information during high-frequency bands through the whole fragment (shown in the dashed oval boxes). Hence NAMAN can obtain higher scores, which estimates noise by attention mechanism, performs better in speech restoration and achieves less speech distortions.

Table 2 compares the run-time latency and the model size of different models. A set of 500 noisy test utterances are selected randomly and fed to the network to estimate the latency and model size which are normalized by the corresponding values of the Mapping model. From the last two rows we can observe that, for both latency and model size, NAMAN uses only about a half of those values in DNAT.

Table 2: *A comparison among Mapping, SNAT, DNAT and NAMAN. N_T and N_M are the run-time latency and model size, respectively, normalized by Mapping.*

	Mapping	SNAT	DNAT	NAMAN
N_T	1	1.06	2.06	1.03
N_M	1	1.08	2.08	1.03

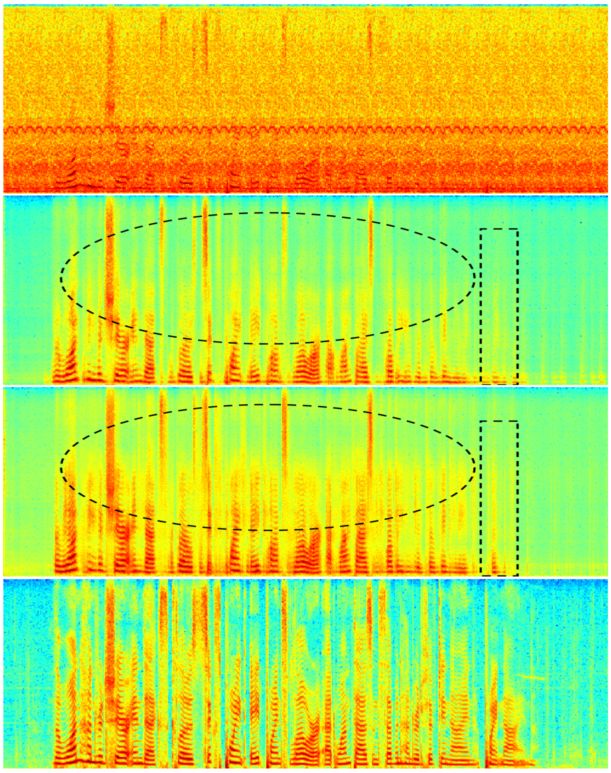

Figure 2: *Spectrograms of an utterance tested on* Buccaneer1 *noise at SNR=0 dB (from top to bottom): noisy speech, DNAT, NAMAN, clean speech.*

4. Conclusion

In this study, we have proposed a novel noise-aware memory-attention framework for regression-based speech enhancement. Compared with the two-stage DNAT model, NAMAN can predict noise information jointly with the denoising process. Experimental results show the proposed NAMAN approach consistently achieves better performances in low SNR conditions, in terms of PESQ, STOI and LSD, than those obtained with DNAT. Furthermore, NAMAN has the distinctive advantages of simple structures and better generalization ability on mismatched conditions. In future work, we plan to expand our model with SNR-aware and speaker-aware training, which may embody complementary capabilities for speech enhancement.

5. Acknowledgements

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grants No. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005. This work was also funded by Tencent.

6. References

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

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[3] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech communication*, vol. 52, no. 5, pp. 450–475, 2010.

[4] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[5] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[6] 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*, vol. 32, no. 6, pp. 1109–1121, 1984.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[8] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[9] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2013.

[10] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[11] D. Servan-Schreiber, A. Cleeremans, and J. L. McClelland, "Encoding sequential structure in simple recurrent networks," CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF PSYCHOLOGY, Tech. Rep., 1989.

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

[13] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3709–3713.

[14] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2014, pp. 577–581.

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[16] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[17] C. Shan, J. Zhang, Y. Wang, and L. Xie, "Attention-based end-to-end speech recognition on voice search," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4764–4768.

[18] X. Hao, C. Shan, Y. Xu, S. Sun, and L. Xie, "An attention-based neural network approach for single channel speech enhancement," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6895–6899.

[19] M. Ge, L. Wang, N. Li, H. Shi, J. Dang, and X. Li, "Environment-dependent attention-driven recurrent convolutional neural network for robust speech enhancement," *Proc. Interspeech 2019*, pp. 3153–3157, 2019.

[20] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*, 2016.

[21] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2401–2405.

[22] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[23] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Interspeech*, 2014, pp. 2685–2689.

[24] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[25] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7398–7402.

[26] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *Interspeech*, 2014, pp. 2670–2674.

[27] Q. Wang, J. Du, L.-R. Dai, and C.-H. Lee, "Joint noise and mask aware training for dnn-based speech enhancement with sub-band features," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, 2017, pp. 101–105.

[28] S.-W. Fu, Y. Tsao, and X. Lu, "Snr-aware convolutional neural network modeling for speech enhancement." in *Interspeech*, 2016, pp. 3768–3772.

[29] F.-K. Chuang, S.-S. Wang, J.-w. Hung, Y. Tsao, and S.-H. Fang, "Speaker-aware deep denoising autoencoder with embedded speaker identity for speech enhancement," *Proc. Interspeech 2019*, pp. 3173–3177, 2019.

[30] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, 2013, pp. 436–440.

[31] B. Xia and C. Bao, "Speech enhancement with weighted denoising auto-encoder." in *Interspeech*, 2013, pp. 3444–3448.

[32] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[33] G. Hu, "100 nonspeech environmental sounds," *The Ohio State University, Department of Computer Science and Engineering*, 2004.

[34] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[35] 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, no. 3, pp. 247–251, 1993.

[36] 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 *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[37] 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.

[38] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," in *Interspeech*, 2008, pp. 569–572.