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

Environment-dependent Attention-driven Recurrent Convolutional Neural Network for Robust Speech Enhancement

Meng Ge[1], *Longbiao Wang*[1,*], *Nan Li*[1], *Hao Shi*[1], *Jianwu Dang*[1,2], *Xiangang Li*[3]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[3]Didi Chuxing, Beijing, China

{gemeng, longbiao_wang, tju_linan, hshi_cca}@tju.edu.cn, jdang@jaist.ac.jp

Abstract

Speech enhancement aims to keep the real speech signal and reduce noise for building robust communication systems. Under the success of DNN, significant progress has been made. Nevertheless, accuracy of the speech enhancement system is not satisfactory due to insufficient consideration of varied environmental and contextual information in complex cases. To address these problems, this research proposes an end-to-end environment-dependent attention-driven approach. The local frequency-temporal pattern via convolutional neural network is fully employed without pooling operation. It then integrates an attention mechanism into bidirectional long short-term memory to acquire the weighted dynamic context between consecutive frames. Furthermore, dynamic environment estimation and phase correction further improve the generalization ability. Extensive experimental results on REVERB challenge demonstrated that the proposed approach outperformed existing methods, improving PESQ from 2.56 to 2.87 and SRMR from 4.95 to 5.50 compared with conventional DNN.

Index Terms: environment-dependent, attention, convolutional network, recurrent network, speech enhancement

1. Introduction

Speech enhancement is one of the corner stones for development of robust automatic speech recognition and communication systems [1, 2]. Notably, the existing systems of speech enhancement are often built by using data-driven approaches based on large scale deep neural networks (DNNs) [3, 4]. However, its accuracy is limited by environment variability and context variability, which result from the mismatch between training and test environments, as well as inadequate consideration of contextual information in real speech respectively.

Great efforts have been made to alleviate these issues for preserving real signals and reducing noises. Conventional statistical signal processing approaches, including multi-step linear prediction (MSLP) [5], weighted prediction error (WPE) [6], etc., analyze the statistical characteristics of noise in various environments. They aim to estimate pattern of the noise and find the way to suppress it. This statistical strategy is applicable to changeable noise environment. However, preservation of real signals is restricted by the strict distribution assumptions.

Recently, data-driven approaches have attracted more interests and achieved better performance than conventional statistical approaches. They view the enhancement problem as a regression one, where the nonlinear regression function is parametrized by deep network, such as DNNs [3, 7], RNNs [8, 9] and CNNs [10, 11]. For robustness improvement in varied environments, simulated or real noise is augmented into the network to enhance the generalization ability of the model [12]. However, such diverse contexts are not fully utilized to preserve the real signal and varied environmental factors are not estimated online for noise reduction.

To address above problems, we propose an end-to-end robust Attention-driven Network (ANet) and its environment-dependent version EDANet. Specifically, ANet consists of three components. The encoder-decoder convolutional component exploits the local frequency-temporal context patterns in the spectrogram. The attention-driven bidirectional recurrent component models different contribution from consecutive frames as well as dynamic correlations between the contextual frames. The final component is a fully-connected layer with dropout layer that further reduces the noises and predicts the clean spectrograms. To improve the generalization ability, we introduce EDANet to better enhance speech signal as well as WPE method to dynamically estimate the environment and correct speech phase corrupted by reverberation. The experiments are verified based on REVERB challenge database [13]. The results obtained from various experiment demonstrate that our model achieves better performance than all the competitors.

2. Baseline Model

One priority choice for speech enhancement is DNN [12], as presented in Fig. 1, which has been extensively explored in the past few years. First, the DNN model is trained from a collection of noisy and clean speech represented by the log spectra features. Then, the well-trained DNN model is fed with the features of noisy speech for generation of the enhanced log spectra features. Afterwards, the additional phase information from the noisy speech is combined with the enhanced log spectra features to reconstruct the new speech.

There are two main limitations of this model. First, it fails to exploit the rich contextual patterns existing in spectrograms, thereby hindering the recovery of real signals. Second, it ignores noises and speech phase distortions caused by various environments, leading to great difficulty in noise removal.

3. Environment-Dependent Attention-driven Neural Network

The proposed approach is presented to make improvement both in signal preservation and noise reduction, as shown in Fig. 2. For signal preservation, Attention-driven Network (ANet)

[*] corresponding author.

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-1477

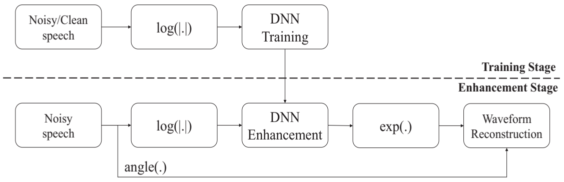


Figure 1: *The diagram of DNN-based framework.*

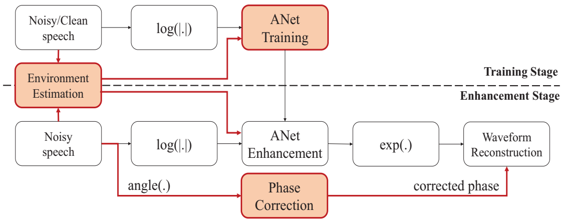


Figure 2: *The diagram of EDANet.*

is designed to exploit more comprehensive contextual features from the input spectrogram as shown in Fig. 3. The convolutional component applies into the local temporal-frequency context; the attention-driven bidirectional recurrent component models the contributions of each frames and dynamic correlations between the contextual frames. In terms of noise reduction, EDANet integrates the strategy of online environment estimation and phase correction. This suppresses different noise patterns in various environments and greatly alleviates the phase distortion caused by reverberation, aiming to achieve better performance in noise reduction.

3.1. Basic attention-driven network (ANet) training

Formally, $\mathbf{x} \in \mathbb{R}^{d\times t}$ represents the noisy spectrogram and $\mathbf{y} \in \mathbb{R}^{d\times t}$ represents its corresponding clean version, where d is the dimension of each frame, i.e., number of frequency bins in the spectrogram, and t is the spectrogram length. Given a training set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$ of the pairs of noisy and clean spectrograms, the problem of speech enhancement is formalized as finding a mapping $g_\theta : \mathbb{R}^{d\times t} \to \mathbb{R}^{d\times t}$ that maps a noisy utterance to a clean one, where g_θ is parameterized by θ. Then the following optimization problem is solved for obtaining the best model parameter θ:

$$\hat{\theta} = \arg\min_{\theta} \sum\nolimits_{i=1}^{n} \|g_\theta(\mathbf{x}_i) - \mathbf{y}_i\|_F^2. \tag{1}$$

Under this setting, ANet is designed as mapping function g_θ for enhancement. It is mainly made up by convolutional, recurrent and full-connected components as shown in Fig. 3.

3.1.1. Encoder-decoder convolutional component

To exploit more accurate local context pattern from given spectrograms, the encoder-decoder CNN [14] is utilized in ANet. The encoder blocks capture the entire context; while the decoder blocks recover details of the time-frequency structure in the spectrogram. This structure enables the network to efficiently model both long contexts and fine-grained structures while maintaining the advantages of CNN [15].

Specifically, $\mathbf{z} \in \mathbb{R}^{b\times w}$ stands for a convolutional kernel of size $b \times w$. A feature map $h^l_{\mathbf{z}}$ is defined as the convolution of the spectrogram $\mathbf{x}$ with kernel $\mathbf{z}$ in the layer l of the encoder-decoder CNN. It is followed by an elementwise nonlinear mapping $\sigma : h^l_{\mathbf{z}}(\mathbf{x}) = \sigma(h^{l-1}_{\mathbf{z}}(\mathbf{x}) * \mathbf{z})$ and $h^0_{\mathbf{z}}(\mathbf{x}) = \mathbf{x}$. Throughout the paper, $\sigma(a) = max\{a, 0\}$ is chosen as the rectified linear function (ReLU), as its effectiveness in alleviation of the notorious gradient vanishing problem has been extensively verified in practice. Each of such convolutional kernel $\mathbf{z}$ produces a 2D feature map, k is applied to separate convolutional kernels to the input spectrogram, resulting in a collection of 2D feature maps $\{h^l_{\mathbf{z}_j}(\mathbf{x})\}_{j=1}^{k}$.

It is worth pointing out in our convolutional component that the pooling operations are removed due to the window size of input spectral (e.g., 7 frames are set in this paper) is small and the pooling operation will quickly reduce the width. Besides, in order to recover the original speech signal, zero-padding operation is applied to guarantee the final prediction of the model have exactly the same length in the time dimension as the input spectrogram.

3.1.2. Attention-driven bidirectional recurrent component

In the convolution operation, however, the same kernel is used across the whole spectrogram, while the contribution from each frame (and each frequency bin) often varies with its distance to the current frame [16]. For speech enhancement task, not all frames in a sequence are equally informative; and speech fragments that are too quiet or noisy contribute little to the current real signal. To incorporate these observations, we introduce the Bidirectional Long Short-Term Memory (BLSTM) [17] with an attention mechanism [18] into ANet. It aims to leverage upon the memory structure capable of capturing some temporal constrains that are not fully utilized in the DNN architecture.

Specifically, the output of the convolutional component is a collection of k feature maps $\{h^l_{\mathbf{z}_j}(\mathbf{x})\}_{j=1}^{k}, h^l_{\mathbf{z}_j} \in \mathbb{R}^{p\times t}$. Before fed into attention-driven BLSTM, those maps are first vertically concatenated into a 2D feature map:

$$H(\mathbf{x}) = \left[h^l_{\mathbf{z}_1}, h^l_{\mathbf{z}_2}, \ldots, h^l_{\mathbf{z}_k}\right] \in \mathbb{R}^{pk\times t}. \tag{2}$$

At each time step t, given input $H_t := H_t(\mathbf{x})$, the contribution value α_t of H_t to the target frame is calculated as

$$\alpha_t = \frac{exp(H_t)}{\sum_{i=t-(s-1)/2}^{t+(s-1)/2} exp(H_i)}, \tag{3}$$

where s is the length of each segment. To further model the dynamic correlations between weighted adjacent frames in the noisy spectrogram, we feed the weighted adjacent frames, i.e., $\hat{H}(\mathbf{x}) = \alpha H(\mathbf{x})$, into the following BLSTM. Thus, the hidden representation $\mathrm{V}(\mathbf{x})$ of target frames is generated as:

$$\mathrm{V}(\mathbf{x}) = BLSTM\{\hat{H}(\mathbf{x})\} \in \mathbb{R}^{q\times t}, \tag{4}$$

3.1.3. Fully-connected and dropout component

To avoid the over-fitting problem and better reduce the useless noises in the various scenes, we adopt a fully-connected layer with dropout strategy for further improving the generalization ability of our model. Formally, for each t, we have:

$$\hat{\mathbf{y}} = max\{0, W\mathrm{V}(\mathbf{x}) + b_W\} \in \mathbb{R}^{d\times t}, \tag{5}$$

where $W \in \mathbb{R}^{d\times q}$ and $b_W \in \mathbb{R}^{d}$ are the parameters. As shown in Eq. (1), the last step is to define the mean-squared error between the predicted spectrogram $\hat{\mathbf{y}}$ and the clean one $\mathbf{y}$, and to optimize all parameters simultaneously. Specifically, AdaDelta is used to ensure a stationary solution.

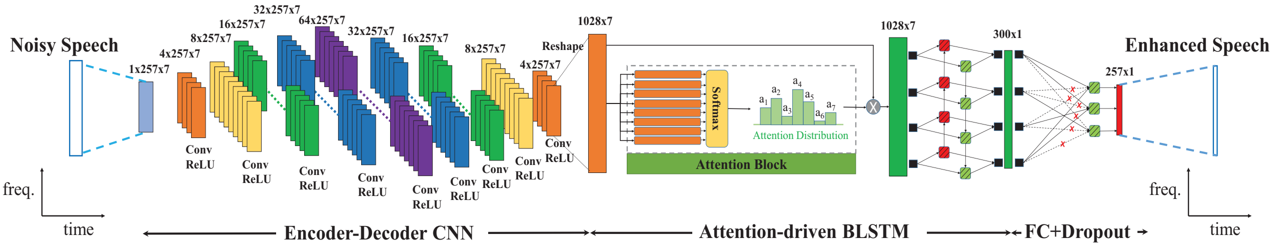


Figure 3: *The architecture of the proposed end-to-end basic Attention-driven Network (ANet) training.*

3.2. EDANet training

The basic ANet training achieves effectiveness in signal preservation and noise reduction. However, it is not able to deal with mismatched conditions of the training and test environment. To enable this environment awareness, we further propose the EDANet. Dynamical estimation is conducted on the environment information $\mathbf{e} \in \mathbb{R}^{d\times t}$ from the noisy spectrogram $\mathbf{x} \in \mathbb{R}^{d\times t}$. Then both $\mathbf{x}$ and $\mathbf{e}$ are fed into ANet for prediction of the clean spectrogram $\mathbf{y} \in \mathbb{R}^{d\times t}$. Thus, in the training stage, Eq. (1) is transformed into the following optimization problem:

$$\hat{\theta} = \arg\min_{\theta} \sum\nolimits_{i=1}^{n} \|g_\theta(\mathbf{x}_i, \mathbf{e}_i) - \mathbf{y}_i\|_F^2. \tag{6}$$

Here, the online environment information $\mathbf{e}$ is estimated by WPE method [6], because of its effectiveness in reverberation estimation. In the enhancement stage, it is also used to online estimate the speech phase ϕ. Further predicted amplitude $\hat{\mathbf{y}}$ is combined to synthesize new speech signal. Thus, these two steps contribute to noise reduction and phase correction.

Specifically, WPE assumes that the desired signal $\mathcal{S}$ is obtained from the noise signal $\mathbf{y}$ through a linear filter G. It follows a zero-mean complex Gaussian with variance λ as $N_{\mathbb{C}}(\mathcal{S}; 0, \lambda)$. Through maximization of the log likelihood

$$L = \max_{G,\lambda} \prod\nolimits_{n=1}^{N} N_{\mathbb{C}}(\mathcal{S}; 0, \lambda) = \min_{G,\lambda} \sum\nolimits_{n=1}^{N} \frac{|\mathcal{S}|^2}{\lambda} + \log \pi\lambda,$$

The parameter G, λ and the desired signal $\mathcal{S}$ are obtained. Here, the environment information $\mathbf{e}$ and corrected phase ϕ are calculated as follows:

$$\mathbf{e} = \mathcal{S}, \quad \phi = angle(\mathcal{S}). \tag{7}$$

4. Experiments

To evaluate the effectiveness of our approach, we tested it experimentally on the REVERB challenge dataset [13]. The REVERB challenge dataset contained simulated and real utterances; the training data only included simulated recordings. The simulated and real recordings on test data were used for evaluation. At the preprocessing, the audio signal was transformed to frames using STFT with a frame length of 512, a frameshift of 256 and the dimension of the log spectral feature vector at 257. The architecture of EDANet was described as follows: the convolutional component consisted of 9 convolutional layers, where the number of filters for each convolution layer was 4, 8, 16, 32, 64, 32, 16, 8 and 4 respectively; and the size of all filters was fixed at 3×3. Two layers of BLSTMs following the convolutional component were adopted, each with 300 hidden units. The back-propagation algorithm was improved by the dropout regularization, with the corruption level is 0.2. To measure the enhancement quality, the PESQ [12] and SRMR [19] measure were applied into evaluation of different models.

4.1. Effect of environment-dependent training

To evaluate the effectiveness of WPE in solving environment variability, the environmental information data were modeled and compared in different ways: 1) **MSLP late**: the entire late reverberation data from MSLP method [5]; 2) **MSLP audio**: the dereverberated audio from MSLP method; 3) **WPE audio**: the dereverberated audio from WPE method.

Based on the comparison results, “WPE audio” achieved the best performance among all the competitors in Table 1. It proved the necessity and effectiveness of WPE method in environment estimation. Specifically, “WPE audio” and “MSLP audio” made the performance as good as that of “MSLP late”. In other words, the derverberated audio better enhanced the generalization ability of model compared with reverberation estimation only. In addition, “WPE audio” outperformed “MSLP audio” in performance improvement by achieving higher quality speech dereverberation. Therefore, “WPE audio” as the best choice for environment-dependent training, was selected to model environment information for speech enhancement task.

Table 1: *Results of DNN with different environment on the simulated data. The best result is highlighted in bold.*

Environment	PESQ			SRMR		
	Far	Near	Avg.	Far	Near	Avg.
No (Baseline) [12]	2.42	2.69	2.56	4.34	4.87	4.61
MSLP late	2.42	2.70	2.56	4.38	5.30	4.84
MSLP audio	2.49	2.79	2.64	4.42	5.23	4.83
WPE audio	**2.52**	**2.85**	**2.69**	**4.50**	**5.35**	**4.93**

4.2. Effect of attention-driven network

To verify the effectiveness of the proposed attention-driven network (ANet) in context information preservation, comparisons with several network structures were conducted. The results are summarized in Table 2, with conclusions detailed below: 1) BLSTM system significantly outperformed DNN system in terms of PESQ and SRMR. To be more specific, compared with conventional independent modeling of frame static context in DNN system, the dynamic context modeling in both directions could exploit richer context information to preserve the real signal. 2) The encoder-decoder CNN with BLSTM (EDCNN-BLSTM) made a better performance than BLSTM system, since the former exploited local frequency-temporal context and complemented dynamic temporal context for further speech enhancement. 3) The proposed EDCNN-Attn-BLSTM approach achieved better results than EDCNN-BLSTM system. The possible explanation is that the attention mechanism enabled the model to learn a better alignment between the input frames with different contributions and the output target frame.

Table 2: *Results of different networks on the simulated data.*

Network Structure	PESQ			SRMR		
	Far	Near	Avg.	Far	Near	Avg.
DNN (Baseline) [12]	2.42	2.69	2.56	4.34	4.87	4.61
BLSTM	2.53	2.89	2.71	4.58	4.95	4.77
EDCNN-BLSTM	2.57	2.95	2.76	4.70	4.97	4.84
EDCNN-Attn-BLSTM (ANet)	**2.61**	**2.98**	**2.80**	**4.74**	**4.99**	**4.87**

4.3. Results and discussion

Finally, the estimated environment information was integrated into the attention-driven network. The results for the simulated data are summed up in Table 3, while the comparison with the real data is illustrated in Table 4. In the case of real data, only SRMR measure is available for evaluation since PESQ measure requires a reference or clean speech to evaluate.

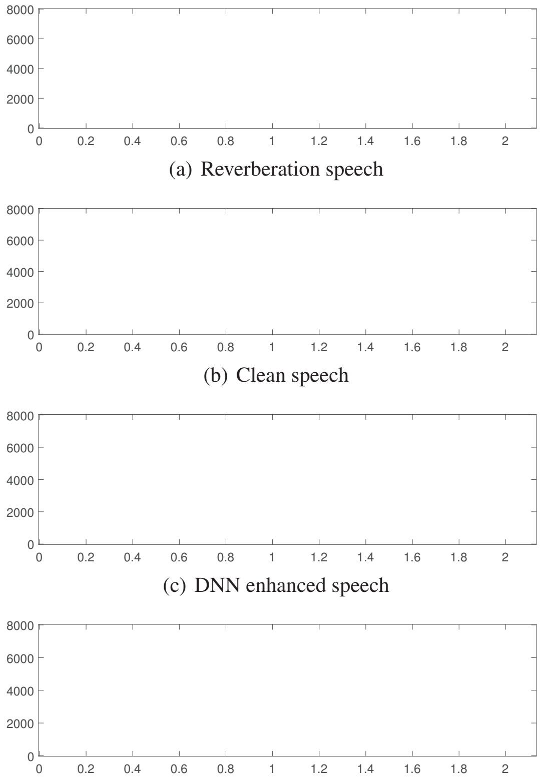
It was observed that incorporating environment information (i.e., WPE audio) allowed the ANet model to generate better enhanced speech. The explanations are given from two aspects. First, various complementary contexts in noisy spectrogram were fully exploited to preserve the real signal. Second, online environmental estimation improved the generalization to reduce different noise. In addition, the proposed EDANet with corrected phase (i.e., WPE phase) made better performance on the simulated data and real data. It proved the alleviation of phase distortion and speech enhancement by corrected phase.

Table 3: *Results (PESQ and SRMR) for the simulated data.*

Method	PESQ			SRMR		
	Far	Near	Avg.	Far	Near	Avg.
No process	2.16	2.59	2.38	3.46	3.96	3.71
MSLP [5]	2.14	2.53	2.34	3.04	3.32	3.18
WPE [6]	2.26	2.72	2.49	3.76	4.27	4.02
DNN (Baseline) [12]	2.42	2.69	2.56	4.34	4.87	4.61
BLSTM	2.53	2.89	2.71	4.58	4.95	4.77
EDCNN-Attn-BLSTM (ANet)	2.61	2.98	2.80	4.74	4.99	4.87
+ WPE audio	2.64	3.05	2.85	4.78	5.04	4.91
+ WPE audio + WPE phase (EDANet)	**2.66**	**3.08**	**2.87**	**4.81**	**5.07**	**4.94**

Table 4: *Results (SRMR) for the real data.*

Method	SRMR		
	Far	Near	Avg.
No process	3.19	3.17	3.18
MSLP [5]	3.04	3.19	3.07
WPE [6]	3.58	3.42	3.50
DNN (Baseline) [12]	4.97	4.92	4.95
BLSTM	5.28	5.07	5.18
EDCNN-Attn-BLSTM (ANet)	5.50	5.28	5.39
+WPE audio	5.56	5.36	5.46
+ WPE audio + WPE phase (EDANet)	**5.59**	**5.40**	**5.50**



(a) Reverberation speech

(b) Clean speech

(c) DNN enhanced speech

(d) EDANet enhanced speech

Figure 4: *Spectrograms of an utterance example.*

To have a better understanding on the experimental results, a case study was carried out by visualizing an utterance example corrupted in reverberation environment. Fig. 4(a) and 4(b) show the log magnitude spectrogram of the reverberant speech and the clean speech. The corresponding DNN-enhanced output is shown in Fig. 4(c). DNN approach achieved good results in reducing the background noise and preserving the low frequency part of the signal. Nevertheless, the approach tended to suppress the high frequency part of the real signal. It may explain the difficulty for the approach in overcoming performance limitations in Table 3 and Table 4. Fig. 4(d) is the log magnitude spectrogram from the enhanced speech using proposed EDANet approach. Compared with Fig. 4(c), the high frequency part of the spectrogram in Fig. 4(d) was better preserved. Hence, EDANet exhibited its advantages in preserving high/low-frequency context information, thereby producing better enhanced spectrogram for speech enhancement. According to our case study, although our proposed EDANet is effective in signal presevation, it does not lead to significant improvement for noise suppression in silent segment as shown in Fig. 4(d). This explains the reason why SRMR score improvements in near-field case are less than in far-field case in Table 3 and Table 4. This should be explored in future work.

5. Conclusion

We proposed an end-to-end environment-dependent approach EDANet for speech enhancement. Compared with the existing methods, the proposed approach exhibited two distinctive features. First, various context in spectrogram could be more comprehensively exploited to preserve the signal, including local frequency-temporal context via encoder-decoder CNN structure, dynamic temporal context via BLSTM, as well as the contribution from contextual frame via an attention mechanism. Second, online environment estimation and phase correction improve generalization and robustness. This enabled our model could better remove noise in complex scenes. Extensive experimental results showed a superior performance of our new approach on speech enhancement.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China (61771333) and Tianjin Municipal Science and Technology Project (18ZXZNGX00330). Sponsored by DiDi GAIA Research Collaboration Initiative.

7. References

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

[2] I. Tashev, *Sound Capture and Processing: Practical Approaches*. John Wiley and Sons, 2009.

[3] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2014.

[4] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T. Sainath, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[5] K. Kinoshita and T. Naktani, "Speech dereverberation using linear prediction," *NTT Technical Review*, vol. 9, no. 7, pp. 1–7, 2011.

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

[7] D. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1492–1501, 2017.

[8] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. Hershey, and B. Schuller, *Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR*. Springer International Publishing, 2015.

[9] A. Maas, Q. Le, T. O'Neil, O. Vinyals, P. Nguyen, and A. Ng, "Recurrent neural networks for noise reduction in robust asr," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

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

[11] O. Ernst, S. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," *arXiv preprint arXiv:1803.08243*, 2018.

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

[13] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, and B. Raj, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[14] D. Wang, Y. Zou, and W. Shi, "A deep convolutional encoder-decoder model for robust speech dereverberation," in *Digital Signal Processing (DSP)*. IEEE, 2017, pp. 1–5.

[15] Y. Lecun and Y. Bengio, *Convolutional networks for images, speech, and time series*. MIT Press, 1998.

[16] D. Yu, W. Xiong, J. Droppo, A. Stolcke, G. Ye, J. Li, and G. Zweig, "Deep convolutional neural networks with layer-wise context expansion and attention," in *Interspeech*, 2016, pp. 17–21.

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

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *Computer Science*, 2014.

[19] T. Falk, C. Zheng, and W. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio Speech & Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

View publication stats