CONTEXT-AWAREIDEAL RATIO MASK ESTIMATION FOR SPEECH DENOISING USING DEEP NEURAL NETWORKS

[1]HANNAH LEE, [2]JAEGYU CHOI, [3]DEOKGYU YUN, [4]SEUNG HO CHOI

[1,2,4]Dept. of Electronic and IT Media Engineering, Seoul National University of Science and Technology
[3]Dept. of Electronic Engineering, Seoul National University of Science and Technology
Email: [1]hannha9656@gmail.com, [2]gyuuu_@naver.com, [3]deokkyyun@gmail.com, [4]shchoi@snut.ac.kr

Abstract - In this paper, we propose a context-aware approachfor speech denoising based on deep learning. In most studies, single deep neural network(DNN) is trained in various noisy environments. However, large variations of noisy environments can cause performance degradation. We develop environmentally-optimized models trained in each noisy environment, which is effective in restoring clean speech spectra under different noisy conditions.In this research work, we first classify noisy spectra through DNN models trained for classifying noise environment before noisy speech spectra are processed. We apply the proposed method to the estimation of ideal ratio mask (IRM) that is used for denoising in spectral domain. Experimental results showed that the proposed context-aware method performs better than that using non-context-awareness model.

Keywords - Speech denoising, Context-awareness,Deep learning, Deep neural network,Ideal ratio mask

I. INTRODUCTION

Many speech denoising techniques have been studied based on the mathematical assumptions that describe the mapping relations between noisy andclean speech signal,including Wiener filtering and signal subspace method [1], MMSE-based spectral amplitudeestimator [2].In recent years, with the progress of deeplearning techniques in signal processing [3], deep neural networks (DNNs)are utilized for speech and audioprocessing. The DNN-based noise reduction algorithmshave also achieved better performancesthan the conventional methods in speech enhancement [4] and speech separation[5]. DNNs containing multiple hidden layers of nonlinearityhave shown great potential to capture the complexrelationships between noisy and clean utterances acrossvarious speakers, noise types and noise levels.In most of these studies, however, noisy signals with different noise environments are trained with single model.

The use of situational information or context is expected to improve the performance of DNN-based denoising. Information of specific noise environment can be used as a contextual cue that improves a performance of system. There have been a number of studies focusing on context-awareness. For example, in [6], five noise eventswere classified based on HMM. Our proposed method uses DNNs to classify the noise environment and estimate ideal ration mask (IRM) for speech enhancement in time-frequency domain [7].

II. THE PROPOSED METHOD

The IRMwas adopted as the learning target of neural networks for speech enhancement in the previous study[7]. The concept of IRM is extended from the ideal binary mask, and it performs well as a soft mask defined as:

$$z^{IRM}(t,f) = \frac{S(t,f)}{S(t,f)+N(t,f)} \quad (1)$$

where $z^{IRM}(t,f)$ is the IRM, and S(t, f), N(t, f) represents the power spectra of the speech and noise signals at the time-frequency unit [8].In this study, we develop DNN model to estimate IRM as network output. First, in this work, we classify the noise environment from noisy speech by passing the spectrum magnitude of the speech through a DNN classification modelas shown in Fig. 1. Afterthe noise environment is classified, the spectral magnitudes of the noisy speech entertheenvironment-wiseDNN model. We obtain anestimate of the IRM for each noisy spectrum andthe estimated IRM is multiplied by the spectrum magnitude of the noisy speech to generate the estimated clean speech.

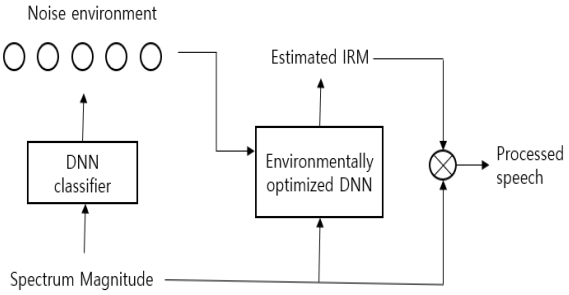


Figure 1Block diagram of the proposed method.

III. EXPERIMENTS AND RESULTS

For the experiments, we used the TIMIT [9] database, sampled at a rate of 16 kHz. As the input data for the neural networks, we used a 257-dimensional spectral magnitude vector.We considered six kinds of

experimental environments, one clean and five noisy ones, in training and test. The noises consist of white Gaussian, babble, factory, car and siren noise. The SNR ranges from -10 to 10dB for the noisy conditions. The DNNs that we used for training and test have three hidden layers that have 512 nodes each. We used adaptive moment estimation (ADAM) for the stochastic optimization thatis a powerful optimization algorithm [10]. As our first step in this study, we considered whether context-awareness model accurately classifythe noise types. Accuracy of context-awareness model was 99%, so it can be seen that noise types of noisy signals were successfully classified. Log spectral distance (LSD) was used to measure the performance of proposed method. The LSDs relativeto clean speech in noisy conditions are measured and summarized as shown in Table 1.

TABLE 1:
AVERAGED LSD VALUES BETWEEN CLEAN SPEECH AND THE RESULTS OF DENOISING

Noise type	Averaged LSD		Difference
	Noisy	Processed	
white	24.81	15.03	-9.78
Factory	16.59	12.18	-4.41
Babble	12.68	10.95	-1.73
Car hon	10.04	6.93	-3.11
Siren	8.40	6.98	-1.42

TABLE 2:
COMPARISON RESULTS OF EACH METHOD IN NOISY CONDITIONS

Method	Averaged LSD	Difference
Non-Context-awareness	10.37	-3.55
Proposed	9.94	-3.98

Table 2 shows the LSDs between clean speech and the results of each method. As the table indicates, the proposed method showed better performance than the non-context-awareness method.

CONCLUSIONS

This paper proposed a context-aware approach for speech denoising based on deep learning, which first classify noisy spectra through DNN models before noisy speech spectra are processed. We applied the proposed method to the estimation of ideal ratio maskthat is used for denoising in the time-spectrum domain. Experimental results showed that theproposed method is effective in reducing noise in various noisy environments. We can conclude that the context-awareness method using DNN can be successfully used for removing noise.

ACKNOWLEDGMENTS

This research project was supported by Ministry of Culture, Sports and Tourism(MCST) and from Korea Copyright Commission in 2017. [Development of predictive detection technology for the search for the related works and the prevention of copyright infringement]

REFERENCES

[1] Philipos C. Loizou,Speech Enhancement: Theory and Practice, CRC press, 2013.
[2] Y. Ephraim and D. Malah, “Speech enhancement using a minimummean-square error log-spectral amplitude estimator,” IEEE Transactionson Acoustics, Speech, and Signal Processing, vol. 33, no. 2, pp.443–445, 1985.
[3] Y. Bengio, P, Lamblin, D.Popovici, and H. Larochelle, “Greedy layer-wise training of deep networks,” Advances in Neural Information Processing Systems, 19: 153-160, MIT Press, Cambridge, 2007.
[4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath et al., “Deep neural networks for acoustic modelingin speech recognition: The shared views of four research groups,” Signal Processing Magazine, IEEE, vol. 29, no. 6, pp. 82–97, 2012.
[5] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Deep learning for monaural speech separation,”in ICASSP, 2014.
[6] Gaunard P., Mubikangiey C. G., Couvreur C. and Fontaine V., “Automatic Classification Of Environmental Noise Events By Hidden Markov Models,”Applied Acoustics, 1998.
[7] Sun, L., Du, J., Dai, L. R., & Lee, C. H.,“Multiple-target deep learning for LSTM-RNN based speech enhancement,”Hands-free Speech Communications and Microphone Arrays (HSCMA), 2017.
[8] C. Hummersone, T. Stokes, and T. Brookes, “On the ideal ratio mask as the goal of computational auditory scene analysis,” Blind Source Separation, pp. 349–368. Springer, 2014.
[9] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, N. (1990). “DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM,” National Institute of Standards and Technology, NTIS Order No. PB91-505065.
[10] D. Kingma, and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint: 1412. 6980, 2014

