SPEECH DE-NOISING USING DNN-BASED ENHANCEMENT MODEL AND KALMAN FILTERING

Candy Olivia Mawalim, Shogo Okada, and Masashi Unoki

Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
Email: {candylim, okada-s, unoki}@jaist.ac.jp

ABSTRACT

This paper describes our proposed speech de-noising system (E023) submitted to the ICASSP SP Clarity Challenge (Speech Enhancement for Hearing Aids). The aim of this challenge is to improve the speech intelligibility while maintaining the quality of a given speech-in-noise with the main concentration on de-noising task. Our system was developed by a hybrid approach using a single-channel DNN-based enhancement network and Kalman filtering as a post processing. The preliminary evaluation on development set showed that our proposed method could improve the speech intelligibility in terms of STOI, but reduces the quality in terms of PESQ score.

Index Terms— Speech enhancement, denoising, hybrid approach, kalman filter

1. INTRODUCTION

One of the main tasks in speech enhancement is de-noising which is to remove noise from a given speech-in-noise (SPIN). The ICASSP SP Clarity Challenge aims to find the optimal speech enhancement method for hearing aids, especially for de-noising, in the more realistic scenario of speech signal[1]. The collected scenes are overlapping with the second Clarity Enhancement Challenge (CEC2) which are more difficult than the first challenge (CEC1) [1], such as, more variety of noise sources, head is moving while talking, and the onset timing is less predictable.

Various methods from signal processing based approaches to machine learning based approches have been proposed for de-noising speech. For instance, the speech enhancement using Wiener filtering [2] which attempts to estimate the clean speech by removing the noise while preserving the spectral information. Meanwhile, Kalman filtering [3] applies a mathematical model to estimate the signal and noise components to recursively update the estimated signal. In past few years, some studies utilized deep neural network (DNN) model and reported a significant improvement in speech enhancement [4, 5]. The DNN with short-time Fourier transform (STFT) can improve the time-frequency representation of the speech signal which has been degraded by noise. This report proposed a hybrid approach of DNN-based model and Kalman filtering for de-noising the input SPIN.

2. METHOD

Figure 1 shows the block diagram of our proposed system. Generally, it consists of the parallel processing of two DNN-based enhancement models for left and right input signals and the post processing by Kalman filtering. The post processing aims to further suppress the interference noise.

The input SPIN is split into left and right signals. We performed STFT with 4 ms frame length and 50% overlapping to obtain the magnitude and phase from each signal. The output magnitude is then input to the DNN-based enhancement model. The DNN-based enhancement model is constructed using two layers bidirectional LSTM network with self-attention and a feed-forward linear layer. The output of this model is the amplitude spectrogram of the clean speech. After obtaining the modified spectrogram, we reconstruct the signals by using inverse STFT and the phase from the STFT. Finally, we pass the output signals to kalman filter before concatenate them to a stereo enhanced speech.

More detail explanation and figure of DNN (TBD).

3. EXPERIMENT

3.1. Dataset

Our proposed DNN-based enhancement network was trained on the training subset of the Clarity speech dataset [6]. It consists of 6000 training scenes. We did not utilize the head rotation information. At this point, we simplify the task by handling single-channel and downsampling the signals from 44.1 kHz to 8 kHz to reduce the training time and space complexity. To ensure the latency requirement (less than 5 ms), we set the frame length as 4 ms and hopping size is half of the

[1] https://claritychallenge.org/docs/icassp2023/icassp2023_intro

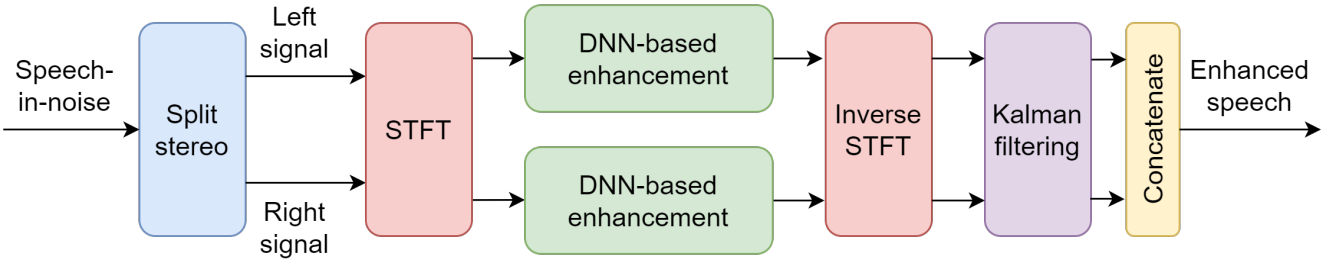

Fig. 1. Proposed method.

Table 1. Evaluation results using development set

Metric	Channel	
	Left	Right
PESQ	1.35	1.47
	1.14	1.24
STOI	0.36	0.20
	0.50	0.34

frame length. For the post processing, we utilized the Kalman filtering that implemented in filterpy library[2].

3.2. Results

We conduct a preliminary objective evaluation by selecting 10 signals from the development set. The perceptual evaluation of speech quality (PESQ)[7] score and the STOI [8] are calculated from the given SPIN and the enhanced speech with the clean speech as reference. Table 1 shows the mean results of PESQ and STOI from each left and right channels. The overall results showed that our proposed method could improve the intelligibility (STOI increased) of the noisy signal but it reduces the quality (PESQ decreased).

4. LIMITATION AND FUTURE WORK

Our DNN model was trained using single-channel 8 kHz signals. Thus, it introduces some disadvantages, such as the reduced quality of the reconstructed signals and produces signals with a limited frequency range. Our proposed method has not yet considered the head turning data and the speaker information. Speaker adaptation techniques were reported to be effective for separating speech signals in a mixture with multiple speakers (such as the given data) [9, 4]. Thus, there are several cases which our model failed to capture the main target speaker (mixed with other speakers). In the future, we will train our model with higher sampling frequency data and incorporate the speaker adaptation for improving the performance of enhancement. Moreover, the considerably insufficient evaluation of this report will also be addressed in our future work.

[2] `https://filterpy.readthedocs.io/en/latest/`

5. REFERENCES

[1] S. Graetzer et al., "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *Proc. of Interspeech*. 2021, pp. 686–690, ISCA.

[2] P. Scalart and J.V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 1996, vol. 2, pp. 629–632 vol. 2.

[3] K. Paliwal and A. Basu, "A speech enhancement method based on kalman filtering," in *ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1987, vol. 12, pp. 177–180.

[4] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[5] Daiki Takeuchi, Kohei Yatabe, Yuma Koizumi, Yasuhiro Oikawa, and Noboru Harada, "Real-time speech enhancement using equilibriated RNN," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. 2020, pp. 851–855, IEEE.

[6] Simone Graetzer, Michael A. Akeroyd, Jon Barker, Trevor J. Cox, John F. Culling, Graham Naylor, Eszter Porter, and Rhoddy Viveros-Muñoz, "Dataset of british english speech recordings for psychoacoustics and speech processing research: The clarity speech corpus," *Data in Brief*, vol. 41, pp. 107951, 2022.

[7] 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)*, 2001, vol. 2, pp. 749–752 vol.2.

[8] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.

[9] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong, "Attention is all you need in speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. 2021, pp. 21–25, IEEE.