2021 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)

Dual-Masking Wind Noise Reduction System Based on Recurrent Neural Network

Wei-Hung Liu[1] whliu.vaplab@gmail.com
Yen-Ting Lai[1] ytlai.vaplab@gmail.com
Kai-Wen Liang[1] kwistron@gmail.com
Jia-Ching Wang[2] jiacwang@gmail.com
Pao-Chi Chang[1] pcchang@ce.ncu.edu.tw

[1]Department of Communication Engineering
National Central University
Taoyuan, Taiwan

[2]Department of Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan

***Abstract*—In this paper, we utilize the architecture of permutation invariant training (PIT) model for wind noise reduction. The proposed system takes the advantage of the dual mask features of the speech separation architecture and properly combines the results of the two masks to synthesize a better signal. We use bidirectional gated recurrent unit (BGRU) to find appropriate weights for the features of short time Fourier transform (STFT). One mask finds the signal that we want to keep. Another mask finds the unwanted signals, which can be removed from the mixed signal. Compared with the traditional method for eliminating wind noise, our proposed method can achieve better noise reduction for non-stationary and non-periodic wind noise.**

Keywords—Wind noise, Noise reduction, Deep learning, Speech separation, Dual mask

I. INTRODUCTION

In the natural environment, noise is always unavoidable. In recording, it is usually necessary to reduce interference that utilizes physical methods e.g., windshield and shock mounting. Although there is pre-protection, it is still necessary to further process those more complex signals in the next stage. The signal processing is divided into two ways. One is noise reduction for noise processing. The other is speech enhancement to strengthen the signal. With the development of deep learning, the problems of these tasks under non-stationary noise are better solved. In this paper, we use the advantage of multiple masks of the speech separation model that permutation invariant training[1](PIT) model. Use short-time Fourier transform(STFT) to extract features and send them to the bidirectional gated recurrent unit(BGRU) for training. While estimating the speech mask, wind noise mask is also used to assist. In the loss calculation, we also consider two losses together to make the training standard better.

II. PROPOSED METHOD

In this paper, the deep learning structure of proposed wind noise reduction system is shown in Fig.1. The system mainly includes four parts, the preprocessing, the training, the mask, and the loss function.

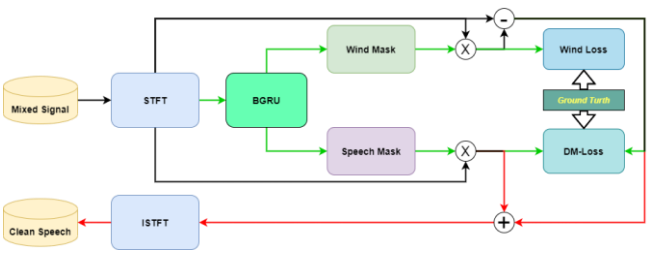

Fig. 1. The flow diagram of the proposed system.

A. Preprocessing

At the beginning, we mix the speech and wind noise. Then we use short-time Fourier transform(STFT) to extract the features of 16KHz mixed signal. The Fig.2 show the flow chart of STFT.

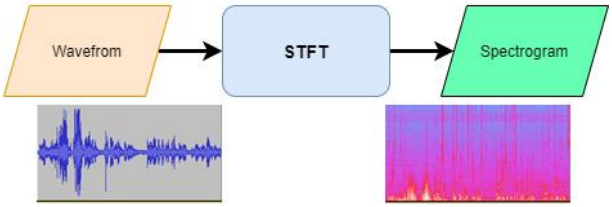

Fig. 2. The flow chart of STFT.

In spectrogram, the horizontal axis is time steps and the vertical axis is frequency. In our system, STFT is computed based on a Hanning window that has the sample size 256 samples and an overlap of 50% each step. For each STFT, we take 128 bins to represent the frequency distribution. STFT can be decomposed into real part and imaginary part. We keep the imaginary part to restore the signal and take the absolute value of the two parts as the training features,

$$X_{feature} = \sqrt{X_{real}{}^2 + X_{imag}{}^2} \quad (1)$$

where $X_{feature}$, X_{real}, andX_{imag} denote the feature part, real part ,and imaginary part, respectively.

B. Training

The neural network model of the proposed system is shown as in Fig. 3. We use BGRU to train the model that has 129 frequency bins for each sequence, 512 GRU cells for each layers and 2 GRU layers.

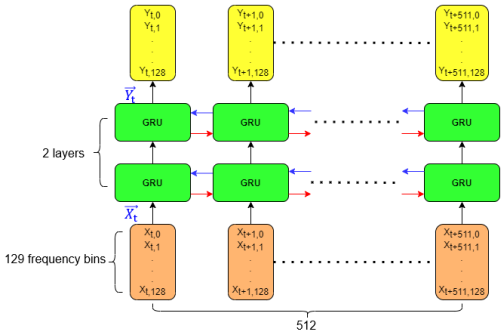

Fig. 3. The neural network model of the proposed system.

978-1-6654-1951-2/21/$31.00 ©2021 IEEE

where $\overrightarrow{X_t}$ and $\overrightarrow{Y_t}$ denote the input sequence and the output sequence of time step t, respectively.

C. Mask

The mask is a weighting matrix. The training weight connect different weights for frequency bins to make two mask for wind and speech. The masking process can be express in Fig.4,

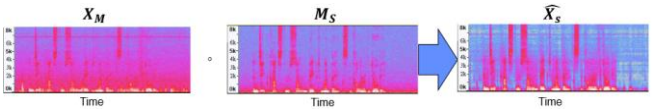

Fig. 4. The speech of masking process

where $\boldsymbol{X_M}$ is the spectrogram of mixed signal, $\boldsymbol{M_S}$ is the speech mask, and $\widehat{\boldsymbol{X_S}}$ is the estimation of speech. The $\boldsymbol{X_M}$ and $\boldsymbol{M_S}$ are element-wise production.

D. Loss function

In our system, we use mean square error to calculate loss. In the noise reduction mask, we hope we can reduce noise well but not affect the speech. Therefore, we decided to give this proportion to deep learning.

The two masks will calculate two losses respectively. Since the mixed signals are known signals, we can verify each other with the original signal under supervised learning. Assuming that when the mask is well trained, the mixed signal minus the wind signal will result in a clean voice signal as shown in the following equation (2),

$$\widehat{\boldsymbol{X_s}} = \boldsymbol{X_M} - \widehat{\boldsymbol{X_w}} \tag{2}$$

where $\widehat{\boldsymbol{X_s}}$ is residual speech, $\widehat{\boldsymbol{X_w}}$ is estimation of wind. In a perfect situation, $\widehat{\boldsymbol{X_s}}$ is indirectly related to the speech signal. Thus, we combine $\widehat{\boldsymbol{X_s}}$ and $\widehat{\boldsymbol{X_w}}$ into dual mask loss as following:

$$J_{DM} = \frac{1}{T\times F}\sum_{t=0}^{T-1}\sum_{f=0}^{F-1}\left\|w_1\widehat{\boldsymbol{X_s}}(t,f) + \boldsymbol{w_2}\widehat{\boldsymbol{X_s}}(t,f) - \boldsymbol{X_S}(t,f)\right\|^2 \tag{3}$$

where J_{DM} is dual mask loss, t is time step, f is frequency bins, $\boldsymbol{X_S}(t,f)$ is ground truth of speech, w_1 is the speech weighting and w_2 is the unwind weighting. Both weightings w_1 and w_2 are decided by training. However, there are too many variables in dual mask loss, which will cause the training to start to be unstable. Therefore, we combined wind loss and dual mask loss as total loss as the criterion as following,

$$J_W = \frac{1}{T\times F}\sum_{t=0}^{T-1}\sum_{f=0}^{F-1}\left\|\widehat{\boldsymbol{X_w}}(t,f) - \boldsymbol{X_W}(t,f)\right\|^2 \tag{4}$$

$$J_{total} = J_W + J_{DM} \tag{5}$$

where J_W is wind loss, J_{total} is total loss, and $\boldsymbol{X_W}(t,f)$ is ground truth of wind noise. The total loss can keep the training in the right direction because the estimation of wind noise can be relatively stable in the PIT at the beginning. Since the PIT has to match the answer with the label separately, the J_W with less variables can stabilize more quickly.

III. EXPERIMENTAL RESULTS

In the experiment, we mixed the librispeech[2] dataset and self-recorded wind noise with ReSpeaker 4-Mic Array. Each audio file has the sampling rate of 16KHz and the audio length of 5~15sec. We divide wind noise into three classes: weak wind (1~3m/s), medium wind (3~5m/s) and strong wind (5m/s~). In general experiment, we divide the mixed signal dataset into 1132 training sets, 204 validation sets, and 240 test sets. However, we found that strong winds have too much saturation. Therefore, we also conducted experiments to remove strong winds for comparison. This experiment that removes strong winds has 652 training sets, 140 validation sets.

In our test, we used scale-invariant signal-to-noise ratio(SI-SNR) and perceptual evaluation of speech quality(PESQ) as the standard for comparison. This symbol of "∗" means that training removes the strong wind set.

TABLE I. SI-SNR and PESQ comparison

	SI-SNR			PESQ		
	Weak	Medium	Strong	Weak	Medium	Strong
Mixed	-0.476	-2.458	-21.256	1.113	1.135	1.176
RNNoise[3]	7.095	6.989	-12.024	1.269	1.325	1.056
PIT	7.088	6.917	-15.118	1.252	1.314	1.102
PIT∗	7.161	7.083	-9.050	1.281	1.339	1.145
Propose method						
DMPIT-DM loss∗	6.850	6.467	-13.794	1.315	1.362	1.112
DMPIT	7.343	7.277	-6.575	1.329	1.378	1.071
DMPIT∗	7.322	7.278	-11.918	1.343	1.400	1.122

In TABLE I, we compare with RNNoise the general noise reduction model with GRU and the single mask method of PIT with BGRU. The method we propose can effectively improve the performance of SI-SNR and PESQ. In the case of dual masks, if only dual mask loss is used, its performance is significantly weaker than the case considered wind loss together. In the case of strong wind training, if it is training with strong wind, it will perform better in SI-SNR, otherwise it will perform better in PESQ.

IV. CONCLUSION

In this research, we propose an improved speech separation model to deal with the problem of noise reduction. This architecture uses the known speech and noise to separately train the desired part and the unwanted part, and then to combine the final results. It can be easily applied to two mask tasks but only requires one of the signals. It is more practical for us to collect real wind noise instead of simulated wind noise for experiments. The result of our method is better than the general noise reduction method and the single mask method, and the ideal of combination is more creative.

REFERENCE

[1] D. Yu et al., "Permutation Invariant Training of Deep Models for Speaker-Independent Multi-talker Speech Separation," in Proc. ICASSP, 2017, pp. 241–245

[2] Vassil Panayotov, Guoguo Chen, Daniel Povey, Sanjeev Khudanpur." Librispeech: An ASR corpus based on public domain audio books," in ICASSP, IEEE, 2015

[3] Jean-Marc Valin," A Hybrid DSP/Deep Learning Approach to Real-Time Full-Band Speech Enhancement," in IEEE Multimedia Signal Processing(MMSP), August, 2018