Time Domain Speech Enhancement Using SNR Prediction and Robust Speaker Classification

任继刚 毛启容 贾洪杰 陈静静
江苏大学计算机科学与通信工程学院

优胜奖

Abstract

Abstract—Recently, deep neural networks (DNNs) have been successfully used for speech enhancement. Most of them are single-task models, and the extracted features are limited by the speech enhancement model. To extract more speech information, we propose a muti-task learning method for speech enhancement using signal to noise ratio(SNR) prediction. SNR prediction task extracts features of the relationship between noise and speech that cannot be obtained by the speech enhancement model. In addition, we added the task of robust speaker recognition to extract speaker features. By sending SNR features and speaker features to the speech enhancement task, the model can obtain more speech-related information to improve the effect of speech enhancement. Experimental results on a public dataset show that our method achieves the state-of-the-art performance and also gets good scores in terms of subjective quality in time-domain speech enhancement.

Index Terms—speech enhancement, muti-task learning, signal to noise ratio *prediction, speaker classification*

Network and Strategy

1. Total method

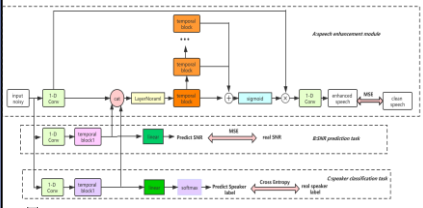

图1. SNR prediction and robust speaker classification muti-task module

2. Speech enhancement method

Conv-TasNet is a state-of-art method for time-domain speech separation and speech enhancement[17]. The noisy speech signal $\boldsymbol{X}$ is transformed into N-dimensional representations $K \in R^{N*T}$ for all the frames by multiplying with a trainable encoder matrix $W \in R^{N*L}$ as follow:

$$\boldsymbol{K} = \boldsymbol{WX} \qquad (2)$$

Then we fed K into separation block W_c,which is composed of a bottleneck layer, temporal convolutional layers, and a mask estimate layer. The $\boldsymbol{c}$ is number of speakers.

$$\mathbf{M_c} = \mathbf{W_c(K)} \qquad (3)$$

where M_c are the masks of c sources, then the masks are multiplied with K to get the separated sources representations:

$$\mathbf{V_c} = \mathbf{M_c} * \mathbf{K} \qquad (4)$$

Finally, V_c is multiplied with a trainable linear decoder $A \in R^{L*N}$

$$\widehat{S_c} = \mathbf{AV_c} \qquad (5)$$

where $\widehat{S_c}$ contains the estimated time-domain signal frames of each source. There are some details when we use Conv-TasNet to implement speech enhancement. First, speech enhancement aims to separate the clean speech from noisy speech, so the number of output source is one, corresponding c=1. Second, the Conv-TasNet model's parameters is learned by SI-SNR loss. This loss is very common in speech separation, but it is not commonly used in speech enhancement, so we use Mean square error (MSE) loss to optimize the parameters of the model, which is more often used in speech enhancement. The MSE loss function is described as follows:

$$\mathbf{L_{MSE=}} \left\| \hat{\boldsymbol{S}} - \boldsymbol{S} \right\|_2^2 \qquad (6)$$

where $\hat{S}$ is the estimate frame and S is the clean signal frame. We refer to the speech enhancement module as Conv-MSE.

3. SNR prediction method

SNR is an important index to measure speech quality, and it is also an important characteristic of noisy speech. The calculation formula is as follows:

$$\boldsymbol{SNR = 10 * log_{10}(\|S\|_2^2 / \|N\|_2^2)} \qquad (7)$$

where S is the clean speech signal frame and N is the noise signal frame. as shown in this formula, SNR includes both speech information and noise information. We train a CNN model to predict SNR. The speech signal is transformed into N-dimensional representations $K \in R^{N1*T}$ by an encoder $W \in R^{N1*L}$,which is similar to (2). Motivated by the temporal convolutional network (TCN) [18,19,20]. we use 1-D dilated convolutional blocks(temporal block1) to capture the connection in time sequence. We also use ResNet[21] between each temporal block to avoid vanishing gradient and exploding gradient.

The reasons why we choose 1-D dilated convolutional neural network to capture the temporal information rather than Long Short-term Memory(LSTM) can be listed as follows: First, although LSTM is commonly utilized when dealing with speech related task, the frame length it can process is limited and LSTM cannot be modeled on a long time series. The input frame length of our model is 64000, even after encoder the frame length is 4000. LSTM is difficult to model on such a long frame. Second, when we train the model, the fixed input is 4s speech, but the length of the speech frame in the verification and testing process is arbitrary. LSTM is suitable for calculating the timing relationship on a fixed length, but not suitable for variable length. However, the dilated convolution block can be processed on frames of any length. With the dilated rate increasing, the receptive field becomes larger, and the frame length that can be processed also becomes longer.

4. Speaker classification method

Generalization is a very important requirement for speech enhancement. Because the speakers we use for testing and training are not the same. In order to achieve the generalization, most models will utilize many samples spoken by many speakers. We can also use this trick to train the speaker classification model. When the speaker classification model is trained by many speakers, even if the tested speaker is not in the training set, we can still get the deep feature related to the tested speaker. We don't care about the result of classification during testing, what we need to use is the deep feature related to the speaker, and this deep feature satisfies the generalization. The deep feature is sent to the speech enhancement model as an auxiliary information. Thus, the speech enhancement model can obtain more speaker information [15].

Experiment

1. *VoiceBank-DEMAND dataset*

To evaluate the effectiveness of our proposed model. we utilized the VoiceBank-DEMAND dataset(VBD) constructed by Valentini et al. [22] which is openly available and frequently used in the state-of-art DNN-based speech enhancement models [11,23, 24,25]. The train and test sets consist of 28 and 2 speakers (11572 and 824 utterances), respectively[15]. We extract 300 pairs from the training set and utilize them as the validation set. We concatenate all the speech of each speaker together and cut them into 4s because the input dimension of Conv-MSE is 4s speech frame. After the data pre-process, the training dataset has 8213 utterances from 28 speakers totally.

2.result

Table II. Objective evaluation results on VBD dataset.

model	PESQ	CSIG	CBAK	COVL
Noisy	1.97	3.35	2.44	2.63
Open-Unmix [29,30]	2.39	3.12	3.19	2.73
DFL [31]	n/a	3.86	3.33	3.22
SEGAN[11]	2.16	3.48	2.94	2.80
Conv-MSE[17]	2.58	3.83	3.31	3.19
Conv-MSE+SNR	2.65	3.95	3.34	3.29
Conv-MSE+SNR+SPK	2.67	4.06	3.36	3.31

Conclusion

In this paper, we propose a novel muti-task method for time-domain speech enhancement, and introduce the SNR prediction task, which can further extracts the speech features as an auxiliary information. We use CNN to extract the auxiliary information and put this auxiliary information into speech enhancement model. Furthermore, the robust speaker classify-cation task is added to extract more speaker features. All the features are extracted directly from the time domain signal through DNN to avoid the loss of phase information when using manual features such as STFT. By introducing additional auxiliary information, the proposed method achieves a nearly 10% improvement on PESQ over the Conv-MSE, which is a state-of-art model in time-domain speech enhancement, and other metrics also achieve great improvement. Thus, we concluded that SNR prediction and robust speaker classification muti-task method is effective for time domain speech enhancement.

References

[1] A Li, C Zheng, L Cheng, R Peng, X Li "A Time-domain Monaural Speech Enhancement with Recursive Learning" arXiv preprint arXiv:2003.09815v2, 2020.
[2] Y Koyama, T Vuong, S Uhlich and B Raj1," Exploring the Best Loss Function for DNN-Based Low-latency Speech Enhancement with Temporal Convolutional Networks" arXiv preprint arXiv: 2005.11611v1, 2020.
[3] F. G. Germain, Q. Chen, and V. Koltun, "Speech Denoising with Deep Feature Losses," Proc. of Interspeech, 2019.
[4] S.W.Fu, C.F.Liao, Y.Tsao, and S.D.Lin,"MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement," Proc. of Int. Conf. on Machine Learning (ICML), 2019.

Contact

联系人：任继刚
电话：18012807579
邮箱：1720705359@qq.com