Single-Channel Target Speaker Extraction System with Attention Enhancement

Yen-Ting Lai[1], Yi-En Lin[1], Pao-Chi Chang[1], and Jia-Ching Wang[2]
[1] Dept. Communication Engineering, National Central University, Taiwan
[2] Dept. Computer Science and Information Engineering, National Central University, Taiwan

***Abstract*--In this paper, we propose a system for single-channel target speaker extraction. We adopt a Temporal Convolutional Network (TCN) architecture as speech extraction model. We also import an attention enhancement to provide system more rich and efficient information. This let the extraction model estimate mask of the target better. With the better mask, the quality of the target speaker extraction can be improve.**

I. Introduction

Talking with other people is a kind of important human communication and social method. When we record the sounds of conversation, we may face complex auditory scenes, especially in noisy and multi-talker environments. For human hearing system, we human have ability to distinguish out the sound we want to focus on, but for Automatic Speech Recognition (ASR) system, the interference may reduce the accuracy seriously. This kind of the problem also called cocktail party problem. It is necessary to apply preprocessing mechanism (e.g., speech separation, target speaker extraction) before ASR system, in order to separate each talker's speech.

Our proposed system is for target speaker extraction. Different from speech separation, target speaker extraction only extracts the target speaker's speech by given a reference speech which can be collected by individual smart device. It can avoid several difficulty in speech separation, such as do not need to determine the number of speakers in the mixture, and suffer from the masks permutation problem. Because we only focus on the target and only it's mask will be generated.

In practical application, we used time-domain model that based on Temporal Convolutional Network (TCN). TCN make Convolutional Neural Network (CNN) can deal with time series modeling, and can be constructed in different model length. For the reference speech, our proposed attention enhancement method help the system generate the target embedding vectors with more rich information. For the loss function, we use scale-dependent signal to distortion ratio (SD-SDR) [1] to replace scale-invariant signal to distortion ratio (SI-SDR) as new reconstruction loss. In the experiments, the results show that the methods mentioned above can improve the extracted speech quality.

II. Proposed System

A. Diagram of proposed system

Our proposed system is referred to the baseline model SpEx+ [2]. The diagram is showed in Fig.1 and the system can be divided into few part. We will introduce their functions below.

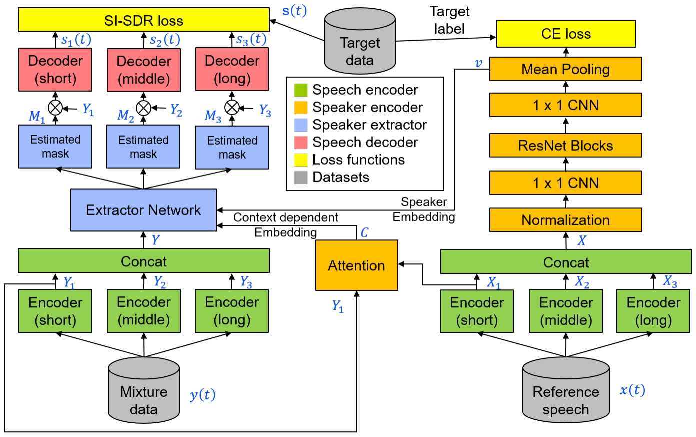

Fig. 1. Diagram of proposed system.

B. Speech encoder

The speech encoder projects the mixture speech $y(t)$ or reference speech $x(t)$ into a latent feature space and are formed by 3 parallel 1-D convolutional layer with different kernel size, L_1 (short), L_2 (middle), L_3 (long). We call multi-scale encoders. The weight sharing strategy is adopted in order to make both sides be in same latent feature space. It can make speech extraction have better performance.

C. Speaker encoder

Speaker encoder extracts speaker embedding of the target speaker from reference speech. In the baseline, it generates a single speaker embedding vector v by ResNet. However, the speaker embedding vector is independent of input data and does not allow to carry rich representation of all the information. So we provide an attention enhancement method to compute time-varying, context-dependent embedding vectors C [3]. These vectors can have interaction with the mixture and be different per frame. Equation (1)-(3) show the formula. Where the X_{n} and Y_{n} are coefficients matrix after encoding, T_{r} and T_{m} are number of frames, respectively. We concatenate both the vectors as new target embedding vectors.

$$d_{t,i} = {Y_{n,t}}^T X_{n,i} \tag{1}$$

$$w_{t,i} = \frac{exp(d_{t,i})}{\sum_{j=1}^{T_r} exp(d_{t,j})} \tag{2}$$

$$C_t = \sum_{i=1}^{T_r} w_{t,i} X_{n,i} \tag{3}$$

D. Speaker extractor

Speaker extractor utilize the embedding vectors to estimate target mask. The details of the extractor are showed in Fig.2.

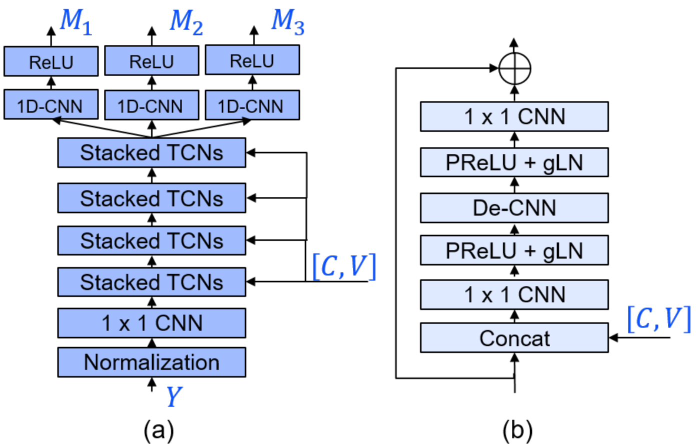

Fig. 2. Diagram of proposed system. Part (a) is the structure of speaker extractor. Part (b) is the details of TCN blocks.

Each stacked TCNs contains B = 8 TCN blocks. The target embedding vectors will only be input to the first block of each stacked TCN. In TCN block, the dilated depth-wise separable convolution (DE-CNN) is used. The interval of kernels are controlled by dilation factor 2^b (b ∈ {0,...,−1}). With the growth dilation of intervals, TCN can have different scale on receptive field and reduce the computation. After the TCN, 3 masks which are respective to multi-scale encoding will be generated.

E. *Speech decoder*

The speech decoders are formed by 3 transposed 1-D convolutional layer corresponding to speech encoders. The decoders utilize the masks to estimate the responses of target and reconstruct the responses back to waveform as follows:

$$s_i = de(M_i \otimes Y_i)\ ,\ i = 1, 2, 3 \tag{4}$$

The symbol ⊗ means element-wise multiplication.

F. *Loss function*

The loss function of our system is the combination of two part, multi-scale reconstruction loss (MSR loss) and cross-entropy loss (CE loss), as (5). γ is a scaling parameter. CE loss evaluates discriminative ability of the speaker embedding vector and is defined as (6). Where N_s is the number of speakers, I_i is the true class label of a target speaker. MSR loss evaluates quality of the estimated speech and (7) shows the formula. α *and* β are the weights to different scales. In the baseline, ρ function takes SI-SDR as the formula. The scaling factor $\mu = \hat{s}^T s / s^T s$ determines the scale of the ground truth s to make it be orthogonal to distortion error. We consider the scaling as an error. It is the sum of two terms, $||\mu s - \hat{s}||^2$ accounting for the residual energy, and $||s - \mu s||^2$ accounting for the rescaling error, and the result is $||s - \hat{s}||^2$.

$$\mathcal{L}_{total} = \mathcal{L}_{reconst} + \gamma \mathcal{L}_{CE} \tag{5}$$

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N_s} I_i \log(\sigma(W \cdot v)_i) \tag{6}$$

$$\mathcal{L}_{reconst} = -[(1-\alpha-\beta)\rho(s_1, s) + \alpha\rho(s_2, s) + \beta\rho(s_3, s)],$$

$$\rho(\hat{s}, s) = \begin{cases} 20\log_{10}\frac{||\mu \cdot s||}{||\mu \cdot s - \hat{s}||} & \text{(SI-SDR)} \\ 20\log_{10}\frac{||\mu \cdot s||}{||s - \hat{s}||} & \text{(SD-SDR)} \end{cases} \tag{7}$$

III. EXPERIMENTS

A. *Datasets*

We use Libri speech corpus [4] for model training and testing. The dataset contains 95 speakers and was divided into 3 set: training set (37828), development set (7188), and testing set (3648). We random select utterances from 2 speakers to generate a mixture with relative SNR between 0 to 5 dB. The relative reference speech is also selected randomly and is not same with the utterance used in mixture. The speakers used in testing set are unseen in training. All the utterances are down-sampled to 8k Hz for reducing computation.

B. *Experiment results*

TABLE I
COMPARISON OF OUR SYSTEM WITH OTHER METHODS

Methods	ρ formula	SI-SDR	PESQ
Mixture	-	2.486	2.1604
SBF-MTSAL-Concat [5]	-	10.228	2.6142
SpEx [6]	-	13.136	2.8866
SpEx+ (Baseline)	SI-SDR	14.267	3.0143
Proposed method	SI-SDR	14.974	3.0732
Proposed method	SD-SDR	**15.041**	**3.1465**

The kernel sizes are L_1 = 2.5ms, L_2 = 10ms, L_3 = 20ms. SI-SDR and PESQ are used for evaluating the quality of estimated target speech. Table.1 shows the comparison of our system with other speaker extraction models. Proposed method has better score than baseline on SI-SDR and PESQ. It means that proposed attention enhancement can help for improving the performance. We also compare the different formula of reconstruction loss, SI-SDR and SD-SDR. The result shows the SD-SDR loss performs better.

EXAMPLES OF REFERENCE STYLES

[1] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?", ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 626-630.

[2] Meng Ge, Chenglin Xu, Longbiao Wang, Eng Siong Chng, Jianwu Dang and Haizhou Li, "SpEx+: A Complete Time Domain Speaker Extraction Network", in Proc. of INTERSPEECH 2020, pp 1406-1410.

[3] X. Xiao et al., "Single-channel Speech Extraction Using Speaker Inventory and Attention Network," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 86-90.

[4] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206-5210.

[5] C. Xu, W. Rao, E. S. Chng, and H. Li, "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6990–6994.

[6] Chenglin Xu, Wei Rao, Eng Siong Chng and Haizhou Li, "SpEx: Multi-Scale Time Domain Speaker Extraction Network," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1370-1384, 2020.