Audio-Visual Single-Channel Signal Separation Based on Big Data Augmentation

Yifan Cui
Statistics Department
University of California, Santa Barbara
Santa Barbara, CA, US
yifanc2020@163.com

Yan Wang
Department of Statistical Sciences, Department of Economics, Faculty of Arts and Science
University of Toronto
Toronto, Ontario, Canada
sarahyy.wang@mail.utoronto.ca

2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI) | 978-1-7281-7738-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/IICSPI51290.2020.9332362

***Abstract*—Cocktail party problem has attracted more attentions in recent years in the speech community. Specially, the single-channel multi-talker speech separation and recognition has become a research hotspot. Also, the visual based information has also been adopted to improve the performance of speech separation and speech recognition. In this paper, we explored to improve the baseline permutation invariant training (PIT) based speech separation systems by two data augmentation methods. Firstly, the visual based information is selected to determine the permutation of separated speakers and also improve the separation performance. Besides, the SpecAugment was explored with big data augmentation method to improve the performance of separation. Finally, we achieved dB of SDR on a mixed dataset using TCD-TIMIT corpus.**

Keywords—audio-visual, speech separation, Big Data, permutation invariant training

I. INTRODUCTION

With the development of speech recognition and processing using deep learning, cocktail party problem has become a research hotspot. There are some related works focusing on single-channel multi-talker speech separation[1-4].

Deep clustering and deep attractor network projects the time-frequency (TF) units into a high-dimension space, where deep clustering use a cluster algorithm to generate a partition of TF units and then use it to separate the mixed speech, while deep attractor network use some attractor points to attract all the TF units corresponding to the target speaker[5-7].

Permutation invariant training (PIT) is an end-to-end speech separation method, which use a unified deep neural network to separate the mixed speech by considering all the possible permutation of outputs, since we don't know which output permutation corresponds to target speakers. PIT is also widely used in single-channel multi-talker speech recognition scenes. Meanwhile, the speech processing based on multi-mode has also been widely explored by the community, including audio-visual based speech recognition, speech separation and so on [8-12].

The videos with speakers' faces are obtained from YouTube, then the audios are mixed and the speakers' face features are extracted use a convolutional neural network (CNN). The mixed speech feature and face feature of each speaker are feed into a separation network to predict the amplitude and phase of each speaker's speech spectrum. Then with inverse short-time Fourier transform (iSTFT), the predicted speech of each speaker is obtained. The target speaker's visual representation is used to extract the amplitude of target speaker's speech spectrum from the mixed speech spectrum, Then the predicted amplitude of spectrum and the phase of mixed speech are used to generate the phase of target speaker's speech. Then with iSTFT, we can generate the target speaker's clean speech[13, 14].

The cross-mode predict model to predict the amplitude of target speaker's spectrum using visual information. Then the predicted amplitude spectrum is used to compute the ideal binary mask of target speaker. With this mask, the target speaker's speech can be extracted from the mixed speech. However, this method is speaker-related, where the cross-mode predict model is dependent on the target speaker. A dataset named TCD-TIMIT is created by recording the lip videos of speakers reading the sentences in TIMIT corpus. The lip videos are recorded in both 0 degree and 30 degree. Also, a speech recognition baseline is also provided. The deep feedforward sequential memory network is used to improve the performance of audio-visual speech recognition. Besides, the per-frame dropout is used to simulate the missing of some video frames. The convolutional neural network and long short-term memory are used to process audio and video features, then the features are averaged across time axis and combined to feed to a classification layer to predict which alphabet is said by the speaker. However, the scene is quite simple and don't work on continuous speech recognition[15-21].

The paper is organized as follows. The Section I is the introduction of previous works. The methods used in this paper is discussed in Section II. The experimental results will be analyzed in Section III, followed by the conclusions in Section IV.

II. AUDIO-VISUAL BASED SPEECH SEPARATION SYSTEM

In this section, we will first discuss the framework of the audio-visual based speech separation system. Then we will introduce the SpecAugment based data augmentation technology and how to use it in the speech separation system.

A. Framework

The PIT based framework is used to do separation firstly, where the audio-based and visual-based features are process separately and then concatenated to a vector as the input of

978-1-7281-7738-0/20/$31.00 ©2020 IEEE

the LSTM/BLSTM based separation network. The two outputs of the separation network and the two target spectrums may have two possible permutations (N! possible permutations for N speaker), which can be written as follows:

$$L_1 = \left|X_1 - \hat{X}_1\right|^2 + \left|X_2 - \hat{X}_2\right|^2 \quad (1)$$

$$L_2 = \left|X_1 - \hat{X}_2\right|^2 + \left|X_2 - \hat{X}_1\right|^2 \quad (2)$$

$$L_{pit} = \min(L_1, L_2) \quad (3)$$

Where L_1 and L_2 are the losses of two permutations, X_1 and X_2 are the target spectrums of two speakers respectively, $\hat{X}_1$ and $\hat{X}_2$ are the predicted spectrums of two speakers respectively, L_{pit} is the loss for training.

In this framework, the output permutation is chosen from all the possible permutations and is invariant with the order of input visual features. Since if we change the order of input visual features, the output permutation can still keep the same or switch to another one.

However, unlike the audio-only based speech separation system, the visual information can be obtained from different camera, or can be extracted using the technology in computer vision. Thus, it is reasonable to obtain each speaker's visual information separately. Then we may align the output permutation to the order of input visual features, i.e., the first output always corresponds to the predicted speech of the person with first input visual feature. We note this system as 'FIX', since the output permutation is fixed given the input visual information order, as illustrated in Figure 2.

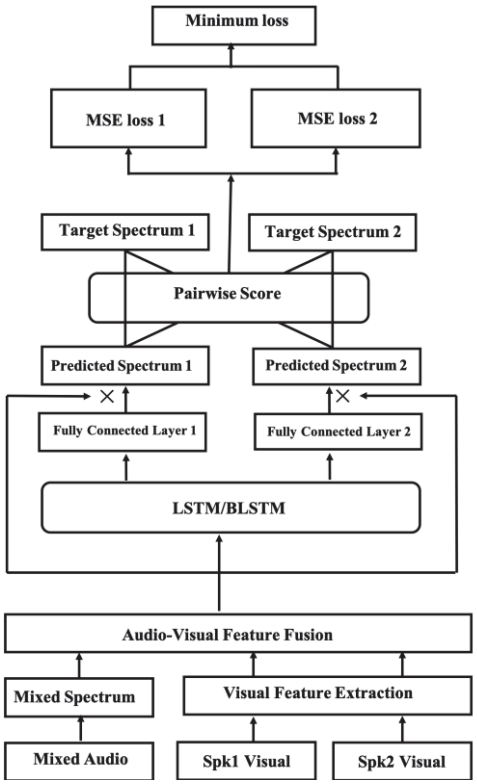

Fig. 1. The PIT based speech separation system.

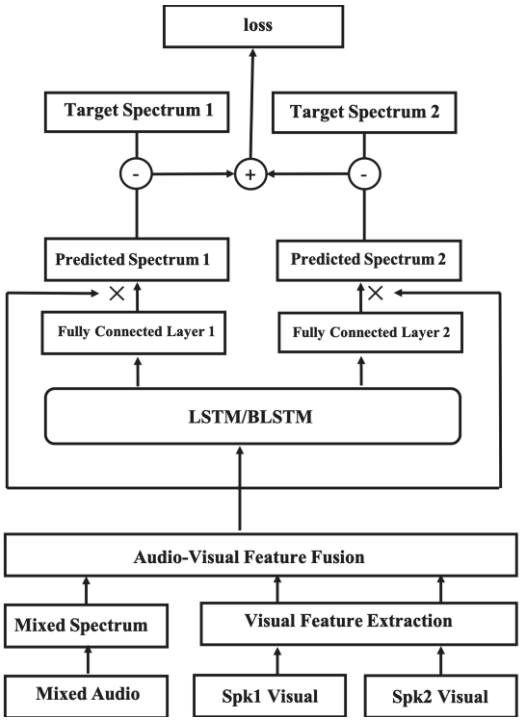

Fig. 2. The FIX based speech separation system.

In the FIX system, the loss is computed as follows:

$$L_{fix} = \left|X_1 - \hat{X}_1\right|^2 + \left|X_2 - \hat{X}_2\right|^2 \quad (4)$$

Where X_i and $\hat{X}_i$ are the target and predicted speech spectrums correspond to the same speaker with i-th visual information.

B. SpecAugment

SpecAugment is proposed in [22], which is a powerful data augmentation technology for end-to-end speech recognition. It is composed of three parts, named time warping, frequency masking and time masking respectively. Time warping will slightly warp the left part and right part spectrum, while frequency masking and time masking respectively will set some random areas across frequency axis and time axis to zeros or mean value of the feature. Though it is widely used in speech recognition system, it hasn't been explored in speech separation task.

Besides, in audio-visual based speech separation system, the SpecAugment technology can be adapted to the visual features. In this paper, since the feature is already extracted by previous work and is not the original picture or video, we cannot to the technology similar to frequency masking on video features. Besides, time warping will destroy the time alignment between input feature and target spectrum, so it is not used in both audio spectrum and video features.

Also, note that time masking on video features is not the same as the per-frame dropout in [17]. Because the first one is to randomly drop some continuous frames, while the second one is to randomly drop every frame.

In this paper, we denote the time masking and frequency masking on audio spectrum as A-TM and A-FM, while the time masking on visual feature as V-TM.

III. Experimental Results

A. Dataset

The dataset often used in single-channel multi-talker speech separation is the WSJ0-2mix dataset [5-9]. However, this dataset doesn't contain visual features. Thus, we choose to use TCD-TIMIT dataset and mix the speech of different speakers with a random signal-to-noise of 0~5 dB to simulate a training, validation and test set, with 20k, 2k and 2k sentences respectively. The dataset is about 30 hours, and the speakers in training, validation and test do not overlap.

The video feature has already been extracted in the TCD-TIMIT corpus, which is processed to have the same frame rate as speech recognition, i.e., 10ms per frame. Thus, for the spectrum of mixed speech and target speech, we choose to do STFT with a frame length 20ms and shift 10ms, and the spectrum has a dimension of 161. The video feature has a dimension of 40.

During training, the training set is doubled by exchanging the order of input features of two speakers while keep the same target permutation for PIT. However, for FIX, it is doubled by both exchanging the order of input features and target spectrums.

B. Network Structure

The network is implemented using TensorFlow [23]. The input to the network is the amplitude of the mixed spectrum and two speaker's video features, which are concatenated and has a total dimension of 241. We use 3-layer LSTM and BLSTM for the separation network, among which LSTM is used for fast development. For the LSTM, the cell size is 640, while for BLSTM, the cell size is 512 among each direction. The two outputs have a dimension of 161, which corresponds to the phase-sensitive mask of the target spectrum, and is multiplied by the mixed spectrum to predict the amplitude of the target speech. Then with the phase of the mixed speech and iSTFT, the predicted target speech is generated.

During training, we use Adam as the optimizer with an initial learning rate of 5e-5, which will be decayed by 0.7 once the loss on validation set is increased after epoch. The dropout is applied to the output of LSTM/BLSTM layers to prevent the model from getting overfitted. Since LSTM is much fast to train, we will use it to verify the effectiveness of the methods proposed in this paper.

C. Audio-only Baseline

The baseline using audio-only information for speech separation is important to see the importance of visual information in the further. Table I shows the audio-only baselines using PIT training strategy and LSTM/BLSTM as separation networks， where the average signal to distortion ratio (SDR) is reported on the test set. As illustrated in the table, BLSTM outperforms LSTM significantly by an improvement of 1.84 dB.

TABLE I. The baseline using audio-only information

Baseline	Average SDR (dB)	
	LSTM	BLSTM
Separation Network	6.64	8.48

D. Audio Spectrum Masking

The effects of time masking and frequency masking on audio spectrum (A-TM and A-FM) are explored in the audio-only system. We only masks once for A-TM and A-FM respectively, but modifies the max size of masking to see the effects, where the actual masking size is a uniform random number between 0 and the max size. The results are presented on Table II.

TABLE II. The performance of time masking and frequency masking on spectrum

Separation Network	A-TM mask size	A-FM mask size	Average SDR (dB)
LSTM	N/A	N/A	6.64
	32	N/A	6.83
	64	N/A	6.91
	96	N/A	6.79
	N/A	16	6.79
	N/A	32	6.86
	N/A	48	6.66
	64	32	6.92

Firstly, the time masking on spectrum can slightly improve the performance of separation, with a max improvement of 0.27 dB. Secondly, the frequency masking on spectrum can also improve the SDR, with a max value of 0.22 dB. Finally, we combine the time masking and frequency masking, and get a 0.28 dB improvement.

From the above results, we can see that the masking-based data augmentation technology can improve the separation results. But when they are combined together, the total improvement is only 0.01 dB better than the system using only time masking. May be a more suitable combination can be found by fine-tuning the A-TM and A-FM size.

E. Audio-Visual Baseline

We then explored to use the visual information to improve the separation performance. We explored the performance of LSTM and BLSTM using both PIT and FIX training strategy. The results are concluded in Table III.

TABLE III. PIT and FIX training strategy when using audio-visual inputs

Separation Network	Training strategy	Average SDR (dB)
LSTM	PIT	6.87
	FIX	8.83
BLSTM	PIT	10.738
	FIX	10.739

As illustrated in the table, for the LSTM based system, the FIX strategy is much better than the PIT strategy, while for BLSTM, both strategies performs almost same. We maintain that the modeling capacity of BLSTM is much stronger than LSTM, which makes BLSTM tends to learn the correspondence between the order of input video features and output permutation. While for LSTM, when training using PIT strategy, the modeling capacity is too weak to learn the correspondence. Also, since we double the dataset by exchanging the order of input features of two speakers while keep the same target permutation, the model may learn that the visual information is not useful, since the

SDR of 6.87 dB is only a little better than the audio-only baseline's 6.64 dB compared to the results of BLSTM.

To confirm this, we plot the curve of training losses L_1, L_2 and L_{pit} in PIT strategy, as well as L_{fix} in FIX strategy, which is equal to L_1, as shown in Figure 3. As illustrated in the figure, when using FIX strategy, the LSTM learns the correspondence between the order of input video features and output permutation. While no matter using FIX or PIT strategy, the BLSTM always learned the correspondence, as L_{fix} is almost the same as the L_{pit}.

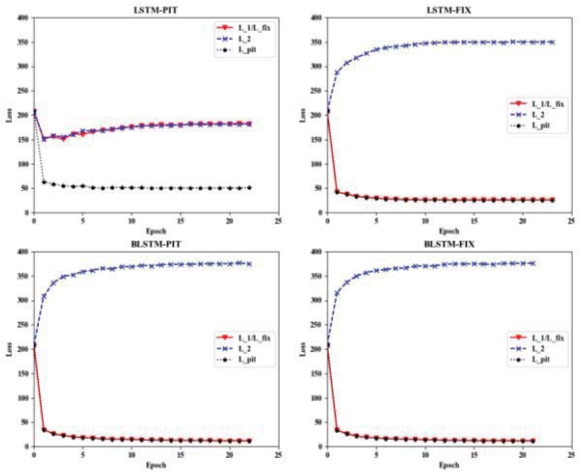

Fig. 3. Training losses in PIT and FIX strategy

F. Video Feature Masking

Now we explored the influence of V-TM's max size, and the results are presented in Table IV. As illustrated in the table, the results is not stable enough, but in most cases, a slight gain can be obtained from the V-TM, where the max value is 0.26 dB.

TABLE IV. PERFORMANCE OF V-TM WITH DIFFERENT MASKING SIZE

Separation Network	Training strategy	V-TM mask size	Average SDR (dB)
LSTM	FIX	N/A	8.83
		16	8.91
		32	8.83
		48	8.91
		64	9.09

G. Final Results

Now A-TM and A-FM, as well as V-TM, are applied to the BLSTM model using FIX strategy. To keep the training stability, in each batch, we only apply A-TM and A-FM, or just V-TM, with the best parameter in Table II and Table IV. The final results are shown in Table V. As illustrated in the table, with the data augmentation, we can improve the performance by 0.045 dB, and the result is much better than the audio-only system.

TABLE V. THE FINAL RESULTS OF BLSTM USING FIX STRATEGY

Separation Network	A-TM mask size	A-TM mask size	V-TM mask size	Average SDR (dB)
BLSTM	N/A	N/A	N/A	10.739
	64	N/A	N/A	10.762
	64	64	32	10.784

IV. CONCLUSIONS

In this paper, we proposed to improve the audio-visual based speech separation by two methods, named FIX strategy and masking based data augmentation technologies. The first one use the order of input visual features of two speakers to determine the output permutation and thus prevent the use of PIT and also improve the training stability, since the LSTM failed when training using PIT strategy. The another one is used to slightly improve the performance of BLSTM based audio-visual speech separation system, and improve more for the LSTM based speech separation systems.

In the further, the speech separation based on time domain can be explored with the proposed methods, like Time-domain Audio Separation Network (TasNet) [24].

REFERENCES

[1] Seide, Frank, Gang Li, and Dong Yu. "Conversational speech transcription using context-dependent deep neural networks." Twelfth annual conference of the international speech communication association. 2011.

[2] Saon, George, et al. "English Conversational Telephone Speech Recognition by Humans and Machines." Proc. Interspeech 2017 (2017): 132-136.

[3] Rao, Kanishka, Haşim Sak, and Rohit Prabhavalkar. "Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer." 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2017.

[4] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in Proc. ICASSP. IEEE, 2016, pp. 31–35.

[5] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in Proc. Inter- speech, 2016, pp. 545–549.

[6] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in Proc. ICASSP. IEEE, 2017, pp. 246–250.

[7] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in Proc. ICASSP. IEEE, 2017, pp. 241–245.

[8] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multi-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 10, pp. 1901–1913, 2017.

[9] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," in Proc. Interspeech, 2017, pp. 2456–2430.

[10] Y. Qian, X. Chang, and D. Yu, "Single-channel multi-talker speech recognition with permutation invariant training," Speech Communication, vol. 104, pp. 1–11, 2018.

[11] X. Chang, Y. Qian, K. Yu, and S. Watanabe, "End-to-end monaural multi-speaker asr system without pretraining," in Proc. ICASSP, 2019.

[12] Ephrat, Ariel, et al. "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation." ACM Transactions on Graphics (TOG) 37.4 (2018): 1-11.

[13] Afouras, Triantafyllos, Joon Son Chung, and Andrew Zisserman. "The Conversation: Deep Audio-Visual Speech Enhancement." Proc. Interspeech 2018 (2018): 3244-3248.

[14] Gabbay, Aviv, et al. "Seeing through noise: Visually driven speaker separation and enhancement." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[15] Harte, Naomi, and Eoin Gillen. "TCD-TIMIT: An audio-visual corpus of continuous speech." IEEE Transactions on Multimedia 17.5 (2015): 603-615.

[16] Zhang, Shiliang, et al. "Robust audio-visual speech recognition using bimodal DFSMN with multi-condition training and dropout regularization." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[17] Zhang, Shiliang, et al. "Deep-fsmn for large vocabulary continuous speech recognition." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[18] Cheng, Gaofeng, et al. "An Exploration of Dropout with LSTMs." Interspeech. 2017

[19] Feng, Weijiang, et al. "Audio visual speech recognition with multimodal recurrent neural networks." 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, 2017.

[20] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[21] Park, Daniel S., et al. "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition." Proc. Interspeech 2019 (2019): 2613-2617.

[22] Abadi, Martín, et al. "Tensorflow: A system for large-scale machine learning." 12th USENIX symposium on operating systems design and implementation (OSDI 16). 2016.

[23] Luo, Yi, and Nima Mesgarani. "Tasnet: time-domain audio separation network for real-time, single-channel speech separation." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.