TWO STAGE AUDIO-VIDEO SPEECH SEPARATION USING MULTIMODAL CONVOLUTIONAL NEURAL NETWORKS

Yang Xian[1], *Yang Sun*[1], *Wenwu Wang*[2], *Syed Mohsen Naqvi*[1]

[1]Intelligent Sensing and Communications Research Group, Newcastle University, UK
[2]Centre for Vision, Speech and Signal Processing, University of Surrey, UK

ABSTRACT

The performance of the audio-only neural networks based monaural speech separation methods is still limited, particularly when multiple-speakers are active. The very recent method [1] used the audio-video (AV) model to find the non-linear relationship between the noisy mixture and the desired speech signal. However, the over-fitting problem always happens when the AV model is trained. Hence, the separation performance is limited. To address this limitation, we propose a system with two sequentially trained AV models to separate the desired speech signal. In the proposed system, after the first AV model is trained, its output is used to calculate the training target of the second AV model, which is exploited to further improve the separation performance. The GRID audiovisual sentence corpus is used to generate the training and testing datasets. The signal to distortion ratio (SDR) and short-time objective intelligibility (STOI) proved the proposed system outperforms the state-of-the-art method.

Index Terms— speech separation, mapping relation, AV model, sequentially AV models

I. INTRODUCTION

Speech separation aims to separate the desired speech from the noisy speech mixture, which has been used as a fundamental unit in various speech-related applications such as hearing aids and robotic [2]–[5]. In the past few decades, the statistical signal processing based methods such as independent component analysis (ICA) [6], independent vector analysis [7] and computational auditory scene analysis (CASA) based methods e.g. model-based expectation maximization source separation and localization (MESSL) [8] have been proposed and proven to provide promising separation results. Recently, the deep neural networks (DNNs) claimed significant in the field of data processing, which provides a powerful tool for speech separation [9]–[13]. Therefore, many DNNs based speech separation methods are proposed. The time-frequency (T-F) features are extracted from the speech mixtures and used to train the DNNs, then the masks are estimated by trained DNNs [9], [10]. The phase and magnitude information is exploited to build the complex domain ideal ratio mask (cIRM), which improves the accuracy of ideal ratio mask [11]. Huang et al. introduced the recurrent neural network (RNN) with the joint optimization to address the monaural speech separation problem [12]. The two-stage DNNs are proposed to address the speech separation problem in noisy reverberant room environments [13]. These state-of-the-art methods still have the limitations, because the audio modality is insufficient to fully discriminate the speech sources.

Inspired by human beings who use the eyes and ears to listen and follow the voice, the researchers introduced the video signal to improve the separation results. Naqvi et al. proposed 3-D tracker based on video information by ultilizing the visual modality, which is used to calculate positions and velocities of the speech sources. A beamforming algorithm is applied to perform the speech separation task for the moving sources [14]. Then Rivet et al. provided an overview of audio-visual speech separation methods and identified the development directions of audio-visual speech separation [15]. Salmen et al. proposed a video-aided model-based source separation algorithm to address the binaural speech separation problem in reverberant room environmentS [16]. They utilized video information to provide the localization and direction of speakers, which is applied to estimate the interaural phase difference (IPD) and interaural Level difference (ILD). By using expectation-maximization (EM) algorithm the time-frequency mask is obtained. Gabbay built a mapping-based AV model to process both speech mixture and video frames. The mouth movements are used to isolate the desired clean speech signal from the noisy speech mixture [1]. Gabbay's method is used as the baseline system. However, the mapping-based method may not address the relation between the input speech mixture and training target, and the accuracy of mapping function needs to be improved.

In this paper, we propose the sequentially training AV models, the first AV model is exploited to estimate the clean speech source. Then the output of the first AV model and clean speech are used to calculate the new training target for the second AV model. And the second AV model assists the first AV model to generate clean speech signal.

The remainder of the paper is organized as follows. Section II describes the proposed audio-visual speech separation system. Then, the experimental settings and evaluation results are given in Section III. Section IV presents the conclusions and provides future work.

II. ALGORITHM DESCRIPTION

II-A. The Proposed Method

In monaural speech separation, the noisy speech mixture can be written as:

$$y(m) = s(m) + n(m) \tag{1}$$

where $y(m)$ denotes the noisy speech mixture at discrete time m, $s(m)$ and $n(m)$ represent the speech source signal and noise at time m, respectively. By using the fast Fourier Transform (FFT), the spectrogram of noisy speech mixture at time t and frequency f is obtained as:

$$Y(t,f) = S(t,f) + N(t,f) \tag{2}$$

where $S(t,f)$ and $N(t,f)$ are the spectra of clean speech signal and noise, respectively.

In [1], the AV model is trained to find the mapping relation between the clean speech signal and noisy speech mixture. The cost function is written as:

$$Loss = \sum_t \sum_f \left(|\hat{S}(t,f)| - |S(t,f)| \right)^2 \tag{3}$$

where $|\hat{S}(t,f)|$ is the spectrogram of estimated clean speech.

Based on this method, we build a system with the sequentially trained AV models to further improve the separation performance. The system is impossible to estimate the training targets perfectly according to the no free lunch theory [17]. To estimate the clean speech, we divide the clean speech into two components, the extracted component $\hat{S}(t,f)$ and unextracted component $S_1(t,f)$. The spectrogram of clean speech is expressed as:

$$|S(t,f)| = |\hat{S}(t,f)| + |S_1(t,f)| \tag{4}$$

$\hat{S}(t,f)$ is obtained by using the first AV model. Therefore, the unextracted component can be obtained by using the spectrogram of clean speech signal minus output of the first AV model. In the second AV model, the unextracted components is estimated. The loss function of the second AV model is written as:

$$Loss_{proposed} = \sum_t \sum_f \left(|\hat{S}_1(t,f)| - |S_1(t,f)| \right)^2 \tag{5}$$

The final estimated clean speech signal is the summation of outputs of the first AV model and the second AV model. The final estimated clean speech is expressed as:

$$|\hat{S}(t,f)|_{final} = |\hat{S}(t,f)| + |\hat{S}_1(t,f)| \tag{6}$$

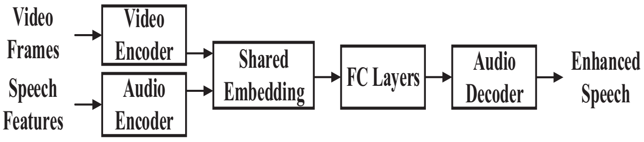

Fig. 1: The block diagram of AV module

II-B. AV model

The AV model is shown in Fig. 1. This model includes video encoder, audio encoder, shared embedding, fully connected layers and audio decoder. The AV model has two inputs: the speech features and video frames. Both the video encoder and audio encoder have the dual tower CNN. The video and audio inputs are encoded into the shared embedding. The details architectures for audio and video encoders are shown in Tables. I & II.

Table I: The architecture of audio encoder.

Layers	Filter	Kernel	Stride
1	64	5×5	2×2
2	64	4×4	1×1
3	128	4×4	2×2
4	128	2×2	2×1
5	128	2×2	2×1

In the shared embedding module, the encoded audio features and encoded video features are concatenated into the AV features. Then, the AV features are processed by the fully connected layers. In the end, the AV features are fed into the audio decoder which consists of 5 transposed convolution layers. The spectrogram of separated speech signals is generated by the audio decoder.

Table II: The architecture of video encoder.

Layers	Filter	Kernel	Stride
1	128	5×5	2×2
2	128	5×5	2×2
3	256	3×3	2×2
4	256	3×3	2×2
5	512	3×3	2×2
6	512	3×3	2×2

II-C. System Architecture

The block diagram of the proposed system is shown in Fig. 3. At the training stage, the features of clean speech signal and speech mixture are extracted, which calculate the mel-spectrogram of the audio signal. The phase information of speech mixture is kept to reconstruct the separated clean speech signal. The mel-spectrogram of the clean speech signal is sliced to the pieces of length 200 milliseconds which corresponded with the video frames. The video signal is processed by pre-processing module. In the pre-processing module, the video signal is normalized and resampled to 25 fps. Besides, it is divided into segments of 5 frames. Then, the video frames are cropped into the mouth-center frame by using 20 landmarks of human face. Both the audio feature and video feature are fed into the AV model 1. After the AV model 1 is trained, the new training target (unextracted component $S_1(t, f)$ is calculated according to the audio features and output of AV model 1 by using (4). The new training target is used to extract the remaining component of clean speech in speech mixture. The new training target and video feature are used to train the AV model 2.

In the testing stage, the feature of speech mixture and video signal are given to trained AV model 1 and AV model 2. The AV model 1 can generate the main part of clean speech signal, then the unextracted component is estimated by AV model 2. Finally, the recovery module is used to reconstruct the speech signal by using outputs of two trained models as (6).

We plot a set of spectrograms in Fig. 2. It can be observed both the baseline method and proposed method can successfully separate the clean speech signal from the speech mixture. Since more information of clean speech is provided by the proposed system, the spectrogram of proposed system is more similar to the spectrogram of clean speech signal when compared with the spectrogram of baseline system.

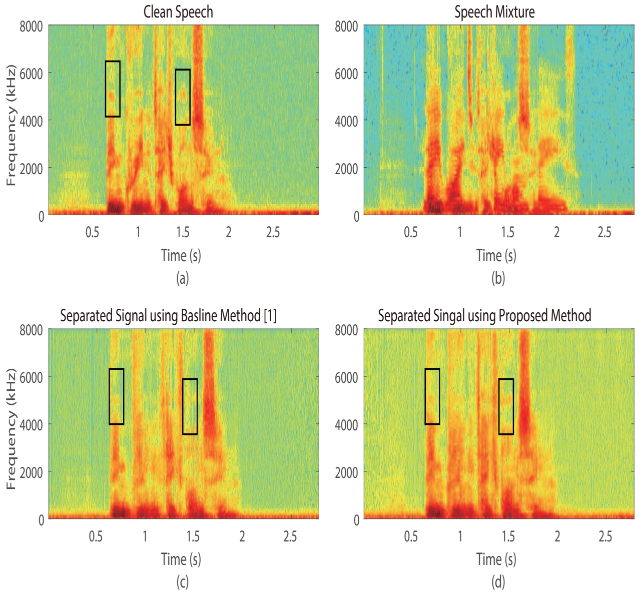

Fig. 2: Spectroagrams of different signals: (a) spectrogram of clean speech; (b) spectrogram of speech mixture; (c) spectrogram of separated speech signal by basline method [1]; (d) spectrogram of separated speech signal by proposed method. The speech mixture is generated by Interference speaker 3 noise at 3dB SNR level.

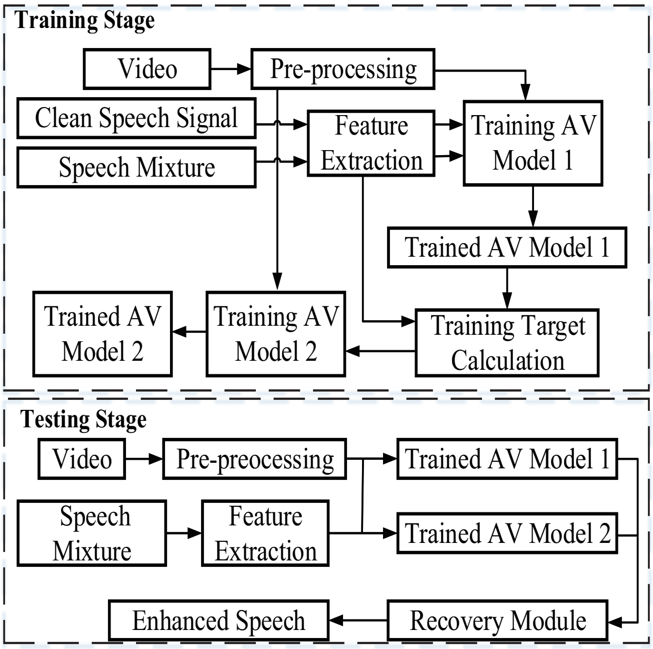

Fig. 3: The block diagram of the proposed sequentially AV models. The AV model 1 is trained firstly, then the AV model 2 is trained to further improve the separation performance of AV model 1.

III. EXPERIMENTAL EVALUATIONS

III-A. Datasets

We use clean speech signals and interference speech signals are selected from the GRID audiovisual sentence corpus [18]. The GRID corpus contains audio and video of 34 speakers, and everyone spoke 1000 sentences. Five speakers from GRID corpus are selected, three male speakers, two female speakers. One male speaker is used as target speaker, which is separated from speech mixture. The speeches of two female speakers and two male speakers are used as the interference noises, which is mixed with the target speech to generate the speech mixture. The clean speech signals are mixed with interference signals at three signal-to-noise ratio (SNR) levels (3dB, 0dB, -3dB).

In total, the training data contains 8902 speech mixtures and the validation data includes 3192 noisy speech mixtures. 8902 speech mixtures are exploited to test the proposed system.

The SDR [19] and STOI [20] are used to evaluate the separation performance. The STOI ranges from 0 to 1. The higher value indicates a better separation performance for SDR and STOI.

III-B. Experimental Results

Tables. III & IV and Fig. 2 provide the separation performance of the baseline system [1] and the proposed system with four interference speech in terms of STOI and SDR improvement.

Table III: Separation performance comparison in terms of SDR improvement with different methods, SNR levels. Two male speakers (Interference 1 and Interference 2) are used in experiments. Each result is the average value of 742 experiments. **Bold** number indicates the best performance.

Measuere	SDR Improvement (dB)					
SNR Level	-3dB		0dB		3dB	
Noise Methods	Interference 1	Interference 2	Interference 1	Interference 2	Interference 1	Interference 2
Gabbay [1]	5.64	5.80	3.75	4.02	1.36	1.75
Proposed	**6.05**	**6.42**	**4.37**	**4.58**	**1.62**	**2.86**

Table IV: Separation performance comparison in terms of SDR improvement with different methods, SNR levels. Two female speakers (Interference 3 and Interference 4) are used in experiments. Each result is the average value of 742 experiments. **Bold** number indicates the best performance.

Measuere	SDR Improvement (dB)					
SNR Level	-3dB		0dB		3dB	
Noise Methods	Interference 3	Interference 4	Interference 3	Interference 4	Interference 3	Interference 4
Gabbay [1]	6.43	3.89	3.99	2.15	1.59	0.10
Proposed	**6.89**	**4.42**	**4.60**	**3.24**	**2.55**	**1.05**

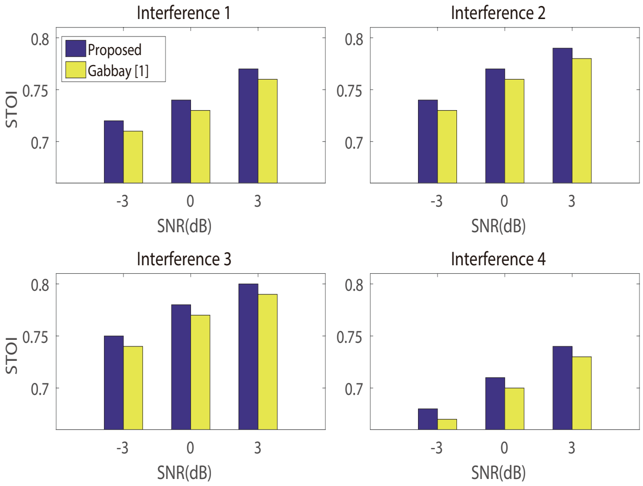

Fig. 4: Averaged STOI of baseline method [1] and the proposed method for different interference speakers (Interference 1, Interference 2, Interference 3, Interference 4). Each result is the average value of 742 experiments.

In terms of SDR improvement, the baseline system and the proposed system can effectively address the speech separation problem. However, the proposed system outperforms the baseline system at four interference scenarios. The main advantage of the proposed system over the benchmark system is the proposed system can provide more SDR improvement in high SNR level. For example, at 3dB SNR level with Interference speaker 2, the SDR improvement of the proposed system over the benchmark system around 1.5dB, which proves the proposed system can provide more information of the clean speech signal over the baseline system at high SNR level. Nevertheless, the highest SDR improvement is obtained by the proposed system at -3dB SNR level for Interference speaker 3. Both the baseline system and the proposed method perform well at low SNR level, which indicates the AV model can successfully remove the interference speech component from the speech mixture at low SNR level. Moreover, the proposed method obtains the highest SDR improvement for Interference 3, when compared with other three interference speeches. In terms of STOI, both the baseline system and the proposed system successfully separated the desired speech component from the speech mixture. From Fig. 4, it is clear that the STOI of proposed system is better than the baseline system. The proposed system generates a similar STOI performance for interference speaker 1, 2 and 3. For Interference 4, both systems obtain the lowest STOI scores.

In summary, the proposed system has better separation performance over the baseline system in terms of SDR and STOI. The experimental results prove that the proposed system can provide more information of desired speech signal. In the higher SNR level, the proposed system can provide significant SDR improvements over the baseline system, which again confirms the unextracted component of the desired speech signal can be estimated by the proposed system.

IV. CONCLUSIONS AND FUTURE WORK

We proposed speech separation systems with the sequentially trained AV models to further improve the separation performance of AV speech separation method. The first AV model was used to estimate the spectrogram of clean speech signal. Then unextracted component was calculated by using the clean speech minus estimated speech of AV model 1, which was exploited to train the AV model 2. The unextracted information of clean speech was estimated by the second AV model. Then the output of two trained AV models was used to generate the desired speech signal, which keeps system to extract more information of the desired speech signal from speech mixture. The experimental results confirmed separation performance of the proposed system outperforms the state-of-art method.

For future research, the system will be modified to solve the speaker independent speech separation problem, which improves the robustness of the system will be further improved.

V. REFERENCES

[1] A. Gabbay, A. Shamir, and S. Peleg, “Visual speech enhancement,” *arXiv preprint arXiv:1711.08789*, 2017.

[2] J. Harris, B. Rivet, S. M. Naqvi, J. A. Chambers, and C. Jutten, “Real-time independent vector analysis with student’s t source prior for convolutive speech mixtures,” in *Proc. of ICASSP*, 2015.

[3] Z. Y. Zohny, S. M. Naqvi, and J. A. Chambers, “Enhancing MESSL algorithm with robust clustering based on student’s t-distribution,” *Electronics Letters*, vol. 50, pp. 552–554, 2014.

[4] Y. Liang, J. Harris, S. M. Naqvi, G. Chen, and J. A. Chambers, “Independent vector analysis with a generalized multivariate Gaussian source prior for frequency domain blind source separation,” *Signal Processing*, vol. 105, pp. 175–184, 2014.

[5] Y. X. Wang, A. Narayanan, and D. L. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[6] A. Hyvarinen and E. Oja, *Independent Component analysis*. Wiley Press, 2001.

[7] Y. Liang, G. Chen, S. M. Naqvi, and J. A. Chambers, “Independent vector analysis with multivariate student’s t-distribution source prior for speech separation,” *Electronics Letters*, vol. 49, no. 16, pp. 1035–1036, 2013.

[8] Y. Sun, Y. Xian, P. Feng, J. A. Chambers, and S. M. Naqvi, “Estimation of the number of sources in measured speech mixtures with collapsed Gibbs sampling,” *Proc. of SSPD*, 2017.

[9] A. Narayanan and D. L. Wang, “Ideal ratio mask estimation using deep neural network for robust speech separation,” *Proc. of ICASSP*, 2013.

[10] Y. X. Wang, A. Narayanan, and D. L. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[11] D. S. Williamson, Y. X. Wang, and D. L. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 17, pp. 483–492, 2016.

[12] P. S. Huang, M. Kim, M. H. Johnson, and P. Smaragdis, “Joint optimization of masks and deep recurrent neural networks for monaural source separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[13] Y. Sun, W. Wang, J. A. Chambers, and S. M. Naqvi, “Two-stage monaural source separation in reverberant room environments using deep neural networks,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, no. 1, pp. 125–139, 2019.

[14] S. M. Naqvi, M. Yu, and J. A. Chambers, “A multimodal approach to blind source separation of moving sources,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 895–910, 2010.

[15] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, “Audiovisual speech source separation: An overview of key methodologies,” *IEEE Signal Processing Magazine*, vol. 19, no. 7, pp. 2125 – 2136, 2011.

[16] M. S. Salman, S. M. Naqvi, A. Rehman, W. Wang, and J. A. Chambers, “Video-aided model-based source separation in real reverberant rooms,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1900–1912, 2013.

[17] D. Wolpert and W. Macready, “No free lunch theorems for optimization,” *TIEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.

[18] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audiovisual corpus for speech perception and automatic speech recognition,” *J. Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[19] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Transanctions on Audio Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 31, no. 3, pp. 125–134, 2014.