Online Audio-Visual Speech Separation with Generative Adversarial Training

PENG, ZHANG

Institute of Automation, Chinese Academy of Sciences; School of Artificial Intelligence, University of Chinese Academy of Science.

JIAMING, XU*

Institute of Automation, Chinese Academy of Sciences; University of Chinese Academy of Science.

YUNZHE, HAO

Institute of Automation, Chinese Academy of Sciences; University of Chinese Academy of Science.

BO, XU

Institute of Automation, Chinese Academy of Sciences; School of Artificial Intelligence, University of Chinese Academy of Science; Center for Excellence in Brain Science and Intelligence Technology, CAS.

Audio-visual speech separation has been demonstrated to be effective in solving the cocktail party problem. However, most of the models cannot meet online processing, which limits their application in video communication and human-robot interaction. Besides, SI-SNR, the most popular training loss function in speech separation, results in some artifacts in the separated audio, which would harm downstream applications, such as automatic speech recognition (ASR). In this paper, we propose an online audio-visual speech separation model with generative adversarial training to solve the two problems mentioned above. We build our generator (i.e., audio-visual speech separator) with causal temporal convolutional network block and propose a streaming inference strategy, which allows our model to do speech separation in an online manner. The discriminator is involved in optimizing the generator, which can reduce the negative effects of SI-SNR. Experiments on simulated 2-speaker mixtures based on challenging audio-visual dataset LRS2 show that our model outperforms the state-of-the-art audio-only model Conv-TasNet and audio-visual model advr-AVSS under the same model size. We test the running time of our model on GPU and CPU, and results show that our model meets online processing. The demo and code can be found at https://github.com/aispeech-lab/oavss.

CCS CONCEPTS • Computing methodologies~Machine learning~Machine learning approaches~Neural networks • Theory of computation~Design and analysis of algorithms~Online algorithms.

Additional Keywords and Phrases: Audio-visual speech separation, online processing, generative adversarial training, causal temporal convolutional network.

---

* Corresponding author

1 INTRODUCTION

Speech separation aims to separate individual audio from an audio mixture of multiple simultaneous talkers, which is an indispensable front-end module in intelligent speech applications, such as automatic speech recognition (ASR) [1]. Solving speech separation tasks by using only audio as input is extremely challenging and does not provide an association of the separated speech signal with the target speaker who we pay attention to. Recently, many models [2, 3, 4, 5] attempt to solve this problem by audio-visual method, which use visual features to "focus" the audio from the target speaker in a scene and improve the speech separation quality. Experimental results show that they are effective. Disappointingly, the successful separation in many of those audio-visual speech separation (AVSS) models are contingent upon the non-causal configuration of speech separation network (e.g., [2, 3, 4, 5]) or rely on a large visual front-end (e.g., [2, 4]), which means that they require future information from the input, need lots of computation and take a huge time delay. It greatly limits the deployment of such models in online applications such as video communication and so on. To our knowledge, there is no AVSS model that can meet online processing until now. Besides, how to reduce the word error rate (WER) of separated speech on the public ASR platforms (e.g., Baidu ASR, Yitu ASR), i.e., reduce the negative effects of SI-SNR, is also a key issue.

Therefore, in this paper, we propose an online audio-visual speech separation model with generative adversarial training. The generative adversarial training is adopted to reduce the negative effects of SI-SNR. We build our generator (i.e., audio-visual speech separator) with causal temporal convolutional network (TCN) block and propose a streaming inference strategy, which allows our model to do online speech separation and maintain a small model size. Discriminator can further help the generator to generate more natural speech. We conduct experiments on 2-speaker mixtures simulated from audio-visual dataset LRS2 [6] and test the running time on the GPU and CPU. Results show that our model achieves significant performance in an online manner. This study represents a major step toward the realization of the AVSS model for real-world speech processing technologies. Our contributions are as follows:

- We propose an online audio-visual speech separation model with generative adversarial training for the first time, which achieves significant performance on challenging audio-visual dataset LRS2 and outperforms the state-of-the-art audio-only model Conv-TasNet and audio-visual model advr-AVSS under the same model size;
- We propose a streaming inference strategy towards the time-domain and TCN based model, which makes our model meets online processing on the GPU and CPU;
- We show that generative adversarial training can further reduce the WER of separated speech and improve other metrics without any additional parameters of the model.

2 RELATED WORK

We briefly review related work in the areas of audio-visual speech separation, online audio-only speech separation and generative adversarial networks.

2.1 Audio-visual Speech Separation

The overview of the AVSS model is shown in Figure 1. First, the AVSS model determines the number of speakers in video and track their faces. This is usually performed by face detection [7] and face tracking algorithm [8]. Then visual features (e.g., face embedding, lip embedding) belong to the target speaker are

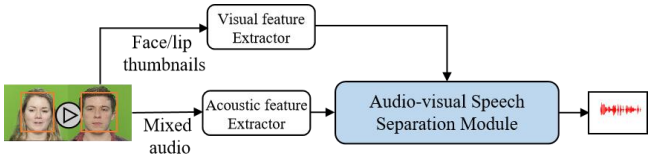

Figure 1: The overview of the audio-visual speech separation model.

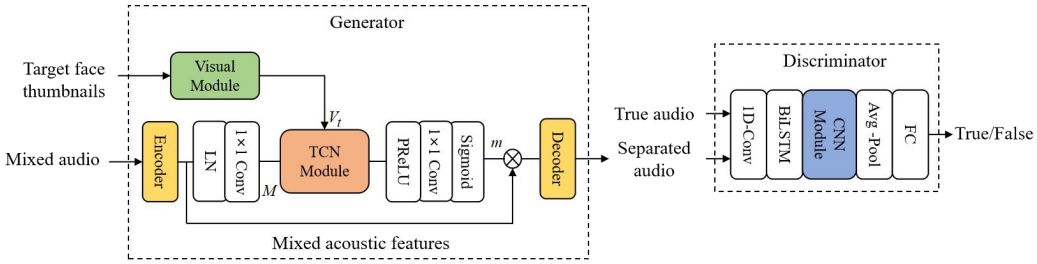

Figure 2: Our proposed online audio-visual speech separation model with generative adversarial training. ⊗: element-wise multiplication; **LN**: layer normalization.

extracted from the face or lip thumbnails, and mixed acoustic features (e.g., spectrogram) are extracted from mixed audio. Finally, the AVSS module takes in these two types of features and output target separated audio. Recent models have used deep neural networks to perform this task. Ephrat et al. [2] propose a speaker-independent audio-visual speech separation model that uses face embedding extracted by a pre-trained face recognition model as auxiliary information, which is demonstrated to be effective in real-world scenarios involving heated interviews, noisy bar, and so on. Lu et al. [3] propose an audio-visual speech separation model-based deep clustering method, which shows some robustness when visual information is partially missing. Zhang et al. [5] propose an audio-visual speech separation model with speech-related visual representation, which achieves excellent performance even on limited size datasets. The main limitation of these AVSS models is that they require future information or rely on a large visual front-end to achieve good performance, which limits their deployment of such models in online applications.

2.2 Online Audio-only Speech Separation

In the areas of audio-only speech separation, several approaches have investigated online model designs [9, 10], but either their performance is not satisfying, or the design is complicated. Han et al. [11] propose the online deep attractor network (ODANet), which can achieve a similar separation accuracy as the non-causal ODANet. Li et al. [12] propose a source-aware context network that models the speaker-independent problem as a speaker-dependent problem with a segment-wise auto-regressive network design, while it requires teacher-forcing during training and more efforts on the auto-regressive architecture design. Another example would be a time-domain separation system, the TasNet [13, 14], that directly performs separation in the waveform domain and achieves state-of-the-art performance. The main limitation of these works is that audio-only approaches build on a strategy to handle the predefined conditions (e.g., a fixed number of sources), limiting their application in the complex auditory scene (e.g., an uncertain number of sources).

2.3 Generative Adversarial Networks

Generative adversarial networks (GANs) are generative model introduced by Goodfellow et al. [15], which consist of a generator (G) and a discriminator (D). The G produces samples from the data distribution $P(x)$ by transforming noise variables z into fake samples $G(z)$. The D is a classifier that aims to recognize whether the sample is from G or training data. G is trained to produce outputs that cannot be distinguished from "real" data. D is trained to do as well as possible in detecting the generator's "fake". More formally, this adversarial learning process is formulated as a two-player minimax game with the objective:

$$\min_G \max_D V(G,D) = E_{x\sim p_{data}(x)}[logD(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))]. \quad (1)$$

Regular GANs suffer from the vanishing gradient problem because of the sigmoid cross-entropy loss function adopted for the discriminator. The least-squares GANs (LSGANs) approach [16] substitutes the cross-entropy loss by the least-squares function, which can generate higher quality samples and perform more stable during the training process. The objective functions of LSGANs can be defined as follows:

$$\min_D V(D) = \frac{1}{2}E_{x\sim p_{data}(x)}\left[(D(x) - 1)^2\right] + \frac{1}{2}E_{z\sim p_{data}(z)}[(D(G(z)))^2], \quad (2)$$

$$\min_G V(G) = \frac{1}{2}E_{z\sim p_{data}(z)}[(D(G(z)) - 1)^2]. \quad (3)$$

3 OUR PROPOSED MODEL

As shown in Figure 2. We propose an online audio-visual speech separation model with generative adversarial training, which consists of a generator (G) and a discriminator (D). In our case, the G performs the audio-visual speech separation. It transforms the mixed audio and target speaker's face thumbnails into the target separated audio. The D aims to distinguish between the separated audio and true audio ones. We will introduce the generator in Section 3.1, the discriminator in Section 3.2, and the loss function in Section 3.3.

3.1 Generator

The overview of generator is shown in Figure 2. Firstly, the visual module processes the target face thumbnails to generate target deep visual features V_t, and we obtain deep mixed acoustic features M from mixed audio by the procession of the encoder, layer normalization (LN), pointwise convolution (1 × 1 Conv). Then, the deep visual features and deep mixed acoustic features are processed by the TCN module, parametric rectified linear unit (PReLU), pointwise convolution (1 × 1 Conv), and sigmoid activation function to generate the mask m. Finally, we obtain target acoustic features by doing element-wise multiplication between mixed acoustic features and mask, which are decoded as target separated audio by the decoder.

3.1.1 Encoder and Decoder.

The spectrogram-based method, i.e., time-frequency coding, and waveform-based method, i.e., time-domain coding, are commonly used methods in speech separation tasks. Time-domain coding has many advantages over time-frequency domain coding, such as trainable parameters, shorter window length, and no phase reconstruction problem [17]. Therefore, we adopt time-domain coding here, i.e., our model works directly with waveform. In our model, the encoder and decoder are modeled as one layer of 1-D convolutional neural network and one layer of the 1-D transposed convolutional neural network respectively.

3.1.2 Visual Module and TCN Module.

As shown in Figure 3(a). In our model, the visual module extracts deep visual features from raw face images.

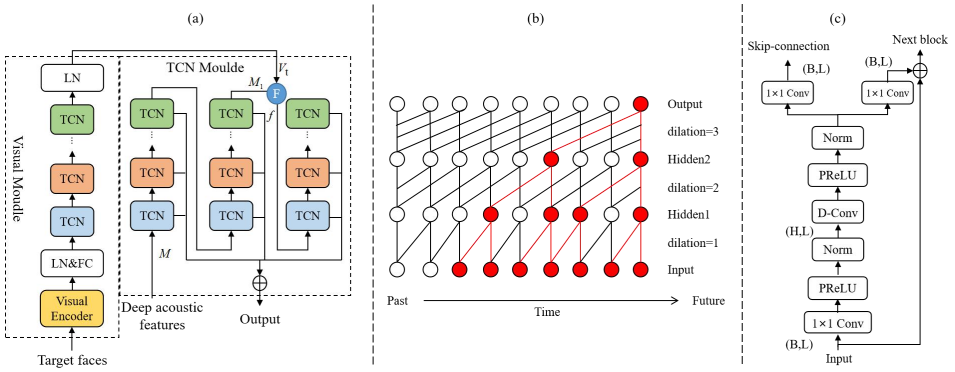

Figure 3: Overall of our proposed method. (a): Details of visual module and TCN module. **F**: fusion stage. (b): The example of receptive field of several stacked causal temporal dilated convolutional blocks (kernel size is 2). (c): Details of TCN block. **Norm**: normalization; **D-Conv**: depth-wise convolution.

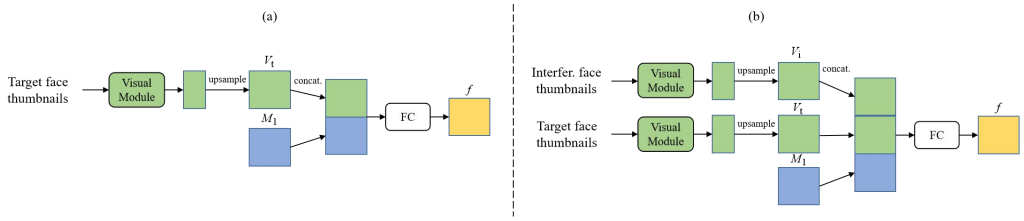

Figure 4: Two multimodal fusion methods. (a) Illustration of concatenating fusion method. (b): Illustration of deep concatenating fusion method.

The visual encoder is a pre-trained model, which is suitable for online AVSS model because of its small model size. We train it to extract speech-related visual features, following [5]. TCN module is used to process time-series data and perform the multimodal fusion. TCN block is the basic component in the visual module and TCN module, so we introduce it firstly. Specifically, the TCN block consists of several components connected in a series, e.g., pointwise convolution (1×1 Conv), PReLU, normalization, depth-wise convolution, PReLU, normalization, pointwise convolution (1×1 Conv), as shown in Figure 3(c). Besides, it has two output paths: a residual path and a skip-connection path. The output of the residual path serves as the input of the next block, and the skip-connection paths of all blocks are summed up as the output of TCN module. When performing error backpropagation, skip-connection can make the error directly reach each TCN block, which is beneficial to optimizing deep neural networks. Similar to the separation module in Conv-TasNet [14] and SpEx [18], both the TCN module and visual module consist of stacked TCN blocks. There are three groups of TCN blocks in the TCN module and one group of TCN blocks in the visual module. Each group consists of four TCN blocks. For the online setting, the causal TCN blocks ensure that only past information to be exploited. Besides, stacking causal TCN blocks can help to increase the receptive field, and the larger the receptive field, the more a network can look into the past. In our model, the dilation factors of four TCN blocks in each group are set as 1, 4, 16 and 64 respectively. Figure 3(b) illustrates the example of the receptive field of several stacked causal TCN blocks. Another key point is how to fuse the target deep visual features V_t and the TCN block's output M_1, which will be introduced in Section 3.1.3.

3.1.3 Multimodal Fusion.

There are two multimodal fusion methods in our model. The first one is concatenate fusion (CF), which is suitable for scenarios, such as video calling, where only the target speaker's visual cues are available. The second one is deep concatenate fusion (DCF), which is suitable for scenarios, such as video conferencing, where all speakers' (e.g., target speaker and interference speaker) visual cues are available. Prior to the fusion, the visual features, including target deep visual features V_t and interference deep visual features V_i, are up-sampled so that the visual features and acoustic features M_1 have same resolution on the time dimension. Then we perform CF or DCF on visual features and acoustic features. Finally, a linear layer is used to reduce the dimension of fusion features. Specifically, for CF, the process mentioned above is shown in Figure 4(a) and can be formulated as follows:

$$f = P([M_1;upsample(V_t)]), \tag{4}$$

where $[a;b]$ represents the operation of concatenating a and b over channel dimension. P represents a linear layer and f represents fusion features. For DCF, the process is shown in Figure 4(b) and can be formulated as follows:

$$f = P([M_1;upsample(V_t);upsample(V_i)]), \tag{5}$$

where $[a;b;c]$ represents the operation of concatenating a, b, and c over channel dimension.

3.1.4 Normalization.

In our model, the choice of the normalization method depends on whether requires online processing or not. For offline (or non-causal) model, we use global layer normalization (gLN), which is proved to outperform other normalization methods, such as batch normalization and so on. In gLN, the feature is normalized over both the channel and the time dimensions, as follows:

$$\begin{cases} gLN(F) = \frac{F-E[F]}{\sqrt{Var[F]+\varepsilon}} \odot \gamma + \beta \\ E[F] = \frac{1}{NT}\sum_{NT} F \\ Var[F] = \frac{1}{NT}\sum_{NT}(F-E[F])^2 \end{cases}, \tag{6}$$

where $F \in R^{N\times T}$ is the feature. $\gamma, \beta \in R^{N\times 1}$ are trainable parameters. ε is a small constant for numerical stability. In online model, gLN cannot be applied since it need future information of the features, so we use cumulative layer normalization (cLN) operation to perform step-wise normalization, cLN is defined as follows:

$$\begin{cases} cLN[f_k] = \frac{f_k - E[f_{t\le k}]}{\sqrt{Var[f_{t\le k}]+\varepsilon}} \odot \gamma + \beta \\ E[f_{t\le k}] = \frac{1}{Nk}\sum_{Nk} f_{t\le k} \\ Var[f_{t\le k}] = \frac{1}{Nk}\sum_{Nk}(f_{t\le k} - E[f_{t\le k}])^2 \end{cases}, \tag{7}$$

where $f_k \in R^{N\times 1}$ is the $k\text{-}th$ frame of the entire feature F, $f_{t\le k} \in R^{N\times k}$ corresponds to the features of k frames $[f_1,\ldots,f_k]$.

3.2 Discriminator

In our model, the discriminator takes in true audio or separated audio and output a label, e.g., true or false, as shown in Figure 2. It consists of several components, e.g., 1-D convolution, bidirectional long-short term memory (BiLSTM), CNN module, average pooling, linear layer. The 1-D convolution extracts acoustic features from the raw waveform. One layer of BiLSTM captures the time dependency on acoustic features. CNN

module learns features that benefit the robust classification task. Average pooling reduces input's width and height and generates important features. Two linear layers can be viewed as the classifier. For the CNN module, six consecutive convolutional blocks are utilized, each of which consists of a 2-D convolutional layer (2-D Conv), spectral normalization (SN) [19], and PReLU. SN is utilized herein to stabilize the training process of the discriminator. The kernel size and the stride of 2-D Conv are set to (3, 2) and (2, 2) along the width and height respectively. We set the number of channels about six consecutive convolutional blocks as 16, 16, 32, 32, 64 and 64 respectively. Besides, we set the number of units of two linear layers as 16 and 1 respectively.

3.3 Loss Function

The G, D networks are jointly optimized by the generative adversarial training algorithm. Mathematically, we define mixed audio, separated audio, true audio, random sampled clean audio and visual input as a_m, a_e, a_t, a_{rs} and v. D tries to classify audio as true or separated. With the LSGANs approach, the loss function can be formulated as follows:

$$\min_D L(D) = E_{a_{rs}\sim p_{data}(a_{rs})}\left[(D(a_{rs}) - l_a)^2\right] + E_{a_m\sim p_{data}(a_m),v\sim p_{data}(v)}[(D(G(a_m, v)) - l_b)^2], \tag{8}$$

where l_a and l_b are sampled from $(0.9, 1.1)$ and $(0, 0.2)$ respectively. The generator G is trained to separate target audio that cannot be distinguished from "true" audio by the discriminator D, so D can correct the output of G towards the realistic distribution. The separation loss of G is denoted as the negative scale-invariant source-noise ratio (SI-SNR), as follows:

$$\min_G L_{ss}(G) = -10\log 10\frac{\|\alpha\cdot a_t\|^2}{\|a_e-\alpha\cdot a_t\|^2}, \tag{9}$$

where a_e and a_t are normalized to zero mean, $\alpha = a_e^T a_t/\|a_t\|^2$. The adversarial loss of G, with the LSGANs approach, can be defined as follows:

$$\min_G L_{gan}(G) = E_{a_m\sim p_{data}(a_m),v\sim p_{data}(v)}[(D(G(a_m, v)) - 1)^2]. \tag{10}$$

Therefore, the final loss of G is as follows:

$$\min_G L(G) = L_{gan}(G) + L_{ss}(G). \tag{11}$$

4 EXPERIMENTS

4.1 Dataset

We conduct experiments on 2-speaker mixtures created from LRS2 dataset [6], which consists of thousands of spoken sentences from BBC television with their corresponding transcriptions. The training, validation, and test sets are generated according to the broadcast date. The method of generating 2-speaker mixtures as follows: we randomly select two audios from the dataset, then mix them at a random signal-noise-ratio (SNR) between -5 dB and 5 dB. The corresponding two videos are concatenated to simulate a cocktail party scene, as shown in Figure 1. Finally, we simulated 40k, 5k, and 3k utterances for training, validation and test sets respectively.

4.2 Implementation Details

In the generator, the kernel size and stride of encoder and decoder are set as 32 and 16 respectively. The number of input channels and the number of output channels in first pointwise convolution are set as 512 and 128 respectively. The number of input channels and the number of output channels in second pointwise convolution are set as 128 and 512 respectively. In the visual module, the dimension of speech-related visual

features is 256, and the number of units of FC is 64. In the TCN block, the H is set as 512, and the kernel size of depth-wise convolution is 3. In the discriminator, the kernel size, stride, number of input channels, and number of output channels are set as 400, 160, 1 and 256 respectively. The number of hidden channels in BiLSTM is 64. For each video clip, we resample the video to 25 FPS and convert it to video frames firstly. Then we use a face detector [20] and a face tracking algorithm to find the faces in each frame and resize the face image to 256 × 256. Besides, the audio is resampled to 16 kHz. Our model is trained with five seconds of audio/video samples using Adam [21] optimizer for 100 epochs with early stopping when there is no improvement in validation loss for consecutive 10 epochs. The initial learning rates for generator and discriminator are set as 1e-3 and 2e-4 respectively.

The baselines are the state-of-the-art audio-only model Conv-TasNet [14], the state-of-the-art audio-visual model advr-AVSS [5] and raw mixed audio. To make a fair comparison, we modify the model size of these two models to the same size as our model and obey the setting in their papers.

4.3 Online Streaming Inference Strategy

In order to test our model in online conditions, we propose a streaming inference method. Specifically, the chunk length of audio and video are both set as T_c, and the stride is T_c too. Considering the encoder should encode complete audio information, we concatenate 1 ms past and 1 ms future audio to the start and end of the input audio chunk respectively. The decoder will output many chunks of separated audio. To obtain complete separated audio, we perform an overlap-add operation between each chunk of separated audio, and the overlap is 1 ms. The causal temporal dilated convolutional network needs history hidden states to compute the current time step state. Therefore, we use a buffer to store the history of hidden states, which needs to update partially after the current chunk computation is completed. The strategy mentioned above can make our model online inference without performance drop.

4.4 Evaluation Metrics

We evaluate the quality of the separated audio using several measures, such as signal-to-distortion ratio (SDR) [22], perceptual evaluation of speech quality (PESQ) [23], and the short-time intelligibility measure (STOI) [24]. To further assess the intelligibility of the separated audio, we use the Baidu automatic speech recognition (ASR) system to compute the word error rate (WER) between the separated audio and the ground truth target audio. The higher SDR, PESQ, STOI represents better. The lower WER represents better.

5 RESULTS AND ANALYSIS

The experimental results of our models and baseline models are summarized in Table 1. By analyzing the results, we can draw the following conclusions. In Table 1, we can see that our model with DCF and GAT outperforms the audio-only model Conv-TasNet and the audio-visual model advr-AVSS on all evaluation metrics (i.e., SDR, PESQ, STOI, WER). Besides, the deep concatenate fusion (DCF) method outperforms the concatenate fusion (CF) method. The reason may be that the DCF method utilizes additional visual information, so the model learns a better mask representation by the DCF method that enhances the target speaker and suppresses the interference speaker. It is obvious that our model with generative adversarial training (GAT) outperforms our model without it. In this paper, we prove that not only GAT is suitable for the time-domain, audio-visual speech separation method, but also GAT can bring steady improvement of speech

Table 1: The performance of our models and baseline models on the test set. CF represents concatenate fusion; DCF represents deep concatenate fusion; GAT represents generative adversarial training.

Models	Online	SDR (dB)	PESQ	STOI (%)	WER (%)
Raw mixed audio	-	0.15	1.78	69	72.2
Conv-TasNet [14]	No	9.60	2.58	87	25.3
advr-AVSS [5]	No	9.94	2.64	87	22.7
Ours (CF)	No	10.01	2.65	87	22.5
Ours (DCF)	No	10.30	2.68	**88**	20.3
Ours (CF+GAT)	No	10.24	2.67	87	20.1
Ours (DCF+GAT)	No	**10.48**	**2.70**	**88**	**19.2**
Conv-TasNet [14]	Yes	6.25	2.22	80	39.0
Ours (CF)	Yes	4.95	2.13	78	39.3
Ours (DCF)	Yes	6.27	2.28	82	32.6
Ours (CF+GAT)	Yes	5.48	2.21	79	36.1
Ours (DCF+GAT)	Yes	**6.76**	**2.33**	**83**	**30.8**

Table 2: The running time of our model on the GPU and CPU under different chunk lengths.

Hardware Platform	Chunk length (T_c)						
	40 ms	80 ms	120 ms	160 ms	200 ms	240 ms	280 ms
CPU	166 ms	167 ms	173 ms	175 ms	178 ms	182 ms	187 ms
GPU	24 ms	24 ms	24 ms	25 ms	25 ms	25 ms	25 ms

quality without any additional parameters. In addition, the results show that GAT is an effective method to reduce the WER of separated speech. We test the running time of our model on the GPU (NVIDIA GeForce GTX 1080Ti) and CPU (Intel i7) under the online streaming inference setting. The definition of online is that the system's response time is less than the length of the time window that is to be processed, so we can draw two conclusions from Table 2. When $T_c \geq 200\,ms$, our model meets online processing on CPU. When $T_c \geq 40\,ms$, our model meets online processing on GPU. The reasonable delays of video communication and human-robot interaction are less than 200 ms and 1 s respectively [25], so our model meets the delay requirements of these scenarios. Besides, the number of our model's parameters is about 11.6 M. To our knowledge, this is the smallest audio-visual speech separation model.

6 CONCLUSIONS

In this paper, we propose an online audio-visual speech separation model for the first time and integrate it with generative adversarial training. Results show that our models achieve significant improvements compared with Conv-TasNet and advr-AVSS. Besides, we propose a streaming inference method towards the time-domain, TCN-based model, and results show that our model meets the online processing on the GPU and CPU. Our model has great potential for application in online scenarios, such as video communication, human-robot interaction and so on.

ACKNOWLEDGMENTS

This work was supported by the Major Project for New Generation of AI (Grant No. 2018AAA0100400), and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB32070000).

REFERENCES

[1] Keisuke Kinoshita, Tsubasa Ochiai, Marc Delcroix and Tomohiro Nakatani. 2020. Improving noise robust automatic speech recognition with single channel time-domain enhancement network. In ICASSP, pp. 7009–7013.

[2] Ariel Ephrat et al. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. ACM Transactions on Graphics, 2018, 37(4CD):112.1-112.11.

[3] Rui Lu, Zhiyao Duan, and Changshui Zhang. 2019. Audio–Visual Deep Clustering for Speech Separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 11, pp. 1697-1712.

[4] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. The conversation: Deep audio-visual speech enhancement. In Interspeech.

[5] Peng Zhang, Jiaming Xu, Jing Shi, Yunzhe Hao and Bo Xu. 2020. Audio-visual Speech Separation with Adversarially Disentangled Visual Representation. arXiv preprint arXiv:2011.14334.

[6] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals and Andrew Zisserman. 2018. Deep audiovisual speech recognition. IEEE transactions on pattern analysis and machine intelligence.

[7] Davis E King. 2009. Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research, vol. 10, pp. 1755–1758.

[8] Bruce D. Lucas and Takeo Kanade. 1981. An iterative image registration technique with an application to stereo vision. In IJCAI.

[9] Masahiro Sunohara, Chiho Haruta and Nobutaka Ono. 2017. Low-latency real-time blind source separation for hearing aids based on time-domain implementation of online independent vector analysis with truncation of non-causal components. In ICASSP. IEEE, pp. 216–220.

[10] Yuxuan Wang, Deliang Wang, and Ke Hu. 2017. Real-time method for implementing deep neural network based speech separation. uS Patent App. 14/536,114.

[11] Cong Han, Yi Luo, Nima Mesgarani. 2019. Online Deep Attractor Network for Real-time Single-channel Speech Separation. In ICASSP.

[12] Zeng-Xi Li, Yan Song, Li-Rong Dai, and Ian McLoughlin. 2018. Source-aware context network for single-channel multi-speaker speech separation. In ICASSP. IEEE, pp. 681–685.

[13] Yi Luo and Nima Mesgarani. 2018. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 696–700.

[14] Yi Luo and Nima Mesgarani. 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM transactions on audio, speech, and language processing, vol. 27, no. 8, pp. 1256-1266.

[15] Ian Goodfellow et al. 2014. Generative adversarial nets. in International Conference on Neural Information Processing Systems, pp. 2672–2680.

[16] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. arXiv preprint arXiv:1611.0407.

[17] Yunzhe Hao, Jiaming Xu, Jing Shi, Peng Zhang, Lei Qin and Bo Xu. 2020. A unified of framework for low-latency speaker extraction in cocktail party environments. In Interspeech, pp. 1431-1435.

[18] Chenglin Xu, Wei Rao, Eng Siong Chng, and Haizhou Li. 2020. Spex: Multiscale time domain speaker extraction network. arXiv preprint arXiv:2004.08326.

[19] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957.

[20] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503.

[21] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. in Proc. Int. Conf. Lear. Represent., 2014.

[22] Emmanuel Vincent, Rémi Gribonval and Cédric Févotte. 2006. Performance measurement in blind audio source separation. IEEE transactions on audio, speech, and language processing, vol. 14, no. 4, pp. 1462-1469.

[23] Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In ICASSP, pp. 749–752.

[24] Cees H. Taal, Richard C. Hendriks, Richard Heusdens and Jesper Jensen. 2011. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125–2136.

[25] Caroline E Harriott and Julie A. Adams. 2017. Towards reaction and response time metrics for real-world human-robot interaction. 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 799-804.