INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Audio-Visual Fusion using Multiscale Temporal Convolutional Attention for Time-Domain Speech Separation

*Debang Liu*[1], *Tianqi Zhang*[1], *Mads Græsbøll Christensen*[2], *Ying Wei*[1], *Zeliang An*[1]

[1]School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2]Audio Analysis Lab, CREATE, Aalborg University, 9000 Aalborg, Denmark

debangliuPaper@163.com, zhangtq@cqupt.edu.cn, mgc@create.aau.dk

## Abstract

Audio-only speech separation methods cannot fully exploit audio-visual correlation information of speaker, which limits its separation performance. Additionally, audio-visual separation methods usually adopt traditional idea of feature splicing and linear mapping to fuse audio-visual features, this approach requires us to think more about fusion process. Therefore, in this paper, combining with the changes of speaker mouth landmarks, we propose a time-domain audio-visual temporal convolution attention speech separation method (AVTA). In AVTA, we design a multiscale temporal convolutional attention (MTCA) to better focus on contextual dependencies of time sequences. We then use sequence learning and fusion network composed of MTCA to build a separation model for speech separation task. On different datasets, AVTA achieves competitive performance, and compared to baseline methods, AVTA is better balanced in training cost, computational complexity and separation performance.

**Index Terms**: audio-visual fusion, time-domain, speech separation, temporal convolutional attention, training cost

## 1. Introduction

Similar to the "cocktail party problem" where humans can track target speech well in a noisy environment [1], the main task of speech separation is to separate target speech in mixture, which is also a basic task of signal processing [2].

In the recent decade, neural network modeling methods have been applied to various aspects of speech processing [3], and the performance of speech separation methods based on deep learning has also been improved significantly. These methods use a data-driven approach to learn better separation models, which greatly compensate for the shortcomings of traditional methods.

At present, audio-only single-channel speech separation methods are mainly based on deep neural networks to extract the time and frequency characteristics of speech signal. Specifically, frequency-domain methods usually perform short-time Fourier transform (STFT) on speech signal to obtain the spectrum, which is used as the input of the neural network to estimate time frequency (T-F) mask of the source [2,4–6]. But these methods have problems with phase reconstruction and latency in calculating the spectrogram. The time-domain methods [7] such as Conv-TasNet and DPRNN [8, 9], etc., directly model time-domain signal to separate the source. However, these methods set a smaller filter length (i.e. convolution kernel size) of the encoder to obtain a longer coded sequence and improve the separation performance, which undoubtedly makes separation network to process longer sequences, and also dramatically increases the complexity and training cost of the model.

Visual information has been demonstrated to help our understanding of speech [10], and it has already yielded many applications in speech enhancement, speech recognition and active speaker detection [11–15]. Therefore, speech separation combined with visual information is a natural idea. Audio-visual single-channel speech separation mainly learns audio-visual fusion strategy [16] or association relationship [17] to assist separation. Where, audio-visual deep clustering for speech separation(A-VDC) [17] has strong generalization ability and robustness to different number of speakers, but frequency-domain modeling and complex network structure make it more suitable for offline systems. Additionally, time-domain audio-visual separation method [18], expands Conv-TasNet [8]. But, it obtains the correlation information of the speech and lip movements of speakers by pre-training lip-reading network, which makes model unable to be trained end-to-end. Recently, the audio-visual speech separation method Visual Voice [19] learns audio-visual embeddings and dependencies from unlabeled videos and achieves better performance on multiple datasets. However, Visual Voice adds a T-F transformation step to the original complex cross-modal learning network to obtain the spectrogram and perform speech separation, which increases the modeling difficulty.

Based on the above analysis, we propose an audio-visual temporal convolution attention speech separation model (AVTA) to learn contextual and cross-modal relations of audio-visual sequences and perform separation task. In AVTA, inspired by Conv-TasNet [8] and temporal convolutional network (TCN) [20], and considering the advantages of attention mechanism in speech separation [21], we design a multiscale temporal convolutional attention network (MTCA) to focus on and learn context dependencies between sequences. In the whole deep separation framework, we first use cross-attention to get the cross-correlation information of audio-visual sequences, then use MTCA to learn the context dependence of audio-visual sequences to fuse audio-visual features. Afterwards, the fused audio-visual feature is fed into separation network to predict mask and obtain the source of each speaker. In general, the contributions of this paper are as follows: 1) We design MTCA network for dependency learning of feature sequences. 2) Combining with MTCA, we design an fusion network for audio-visual feature fusion, and construct an separation framework to improve performance. 3) Compared with baseline method, we further reduce the coded sequence length, and better balance training cost and separation performance of AVTA.

## 2. Our speech separation network

As shown in Figure 1, AVTA includes four modules: audio and visual encoder, sequence learning network, fusion and separation network, decoder. The encoder maps visual and audio


10.21437/Interspeech.2023-801

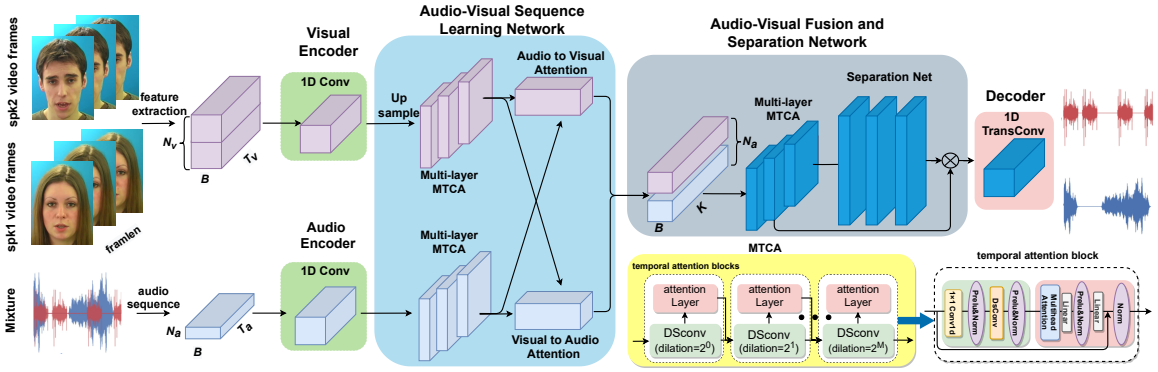


Figure 1: *The top of the picture is the AVTA network structure. The bottom right is the MTCA network structure.*

features in a high-dimensional space. Sequence learning network obtains context-dependent and cross-correlation information of audio-visual feature sequences. Fusion and separation network fuses the audio-visual features and computes the separation mask. Decoder calculates speech waveform of speaker.

### 2.1. Audio and visual encoders

The time-domain mixed speech signal can be denoted as $x \in \mathbb{R}^{B\times 1\times L_{seqa}}$, which includes speech of $C$ speakers $s_{a_1}, ..., s_{a_C} \in \mathbb{R}^{B\times 1\times L_{seqa}}$:

$$x = \sum_{i=1}^{C} s_{a_i}, i = 1, 2, ..., C, \tag{1}$$

where $B$ denotes batch size, $i$ is the $i$-th speaker, $L_{seqa}$ denotes the total number of samples of the speech signal. In the audio encoder, $x$ is intercepted into segments of length $T_a$, then these segments are combined into the input matrix $\mathbf{X} \in \mathbb{R}^{B\times T_a\times 1}$:

$$\text{AudioEncoder}\left(\mathbf{X}\right) = \mathcal{F}(\text{conv1d}(\mathbf{X}, L_a, S_a, C_{in}, C_{out})), \tag{2}$$

where conv1d$(\cdot)$ is 1D convolution operation, $L_a$ and $S_a$ are size and stride of the convolution kernel respectively. $C_{in} = 1$ and $C_{out} = N_a$ are number of input and output channels of the convolution respectively. $\mathcal{F}(\cdot)$ represents the rectified linear unit (ReLU) [22], and the output of audio encoder can be denoted by $\mathbf{E}_a \in \mathbb{R}^{B\times K_a\times N_a}$, where $K_a = \frac{2(T_a - L_a)}{L_a} + 1$ is output coded sequence length.

The visual signal can be denoted as $y_i \in \mathbb{R}^{B\times N_i\times L_{seqv}}$, which is composed of the speaker mouth feature landmarks. Where feature dimension of the $i$-th speaker is $N_i = N/C$, $L_{seqv}$ is the frame length of the video stream. In the visual encoder, $y_i$ is truncated into segments $y'_i \in \mathbb{R}^{B\times N_i\times T_v}$ of length $T_v$. Then, mouth feature landmarks $y'_i$ of $C$ speakers in mixture are spliced in direction $N_i$ to form a matrix $\mathbf{Y} \in \mathbb{R}^{B\times T_v\times N}$. Afterwards, $\mathbf{Y}$ will be fed into the visual encoder:

$$\mathbf{Y} = \text{concat}\left[{y'}_1^{\top}, {y'}_2^{\top} \cdots, {y'}_C^{\top}\right], \tag{3}$$

$$\text{VisualEnoder}\left(\mathbf{Y}\right) = \mathcal{F}(\text{conv1d}(\mathbf{Y}, L_v, S_v, C_{in}, C_{out})), \tag{4}$$

when mixture contains $C = 2$ speakers, we set $N_i = 20$, the input channels of the visual encoder are $C_{in} = N = N_i C = 40$, the output channels are $C_{out} = N_v = N_a$, $L_v$ and $S_v$ are convolution kernel size and stride of the visual encoder, respectively. The output of the visual encoder is $\mathbf{E'}_v \in \mathbb{R}^{B\times K_v\times N_a}$, where the calculation process of $K_v$ is similar to $K_a$.

### 2.2. Audio-visual sequence learning network

The encoded audio and visual sequences is not synchronized in time, therefore, we use linear interpolation to upsample $\mathbf{E'}_v$ and output $\mathbf{E}_v \in \mathbb{R}^{B\times K_a\times N_a}$.

#### 2.2.1. Multiscale temporal convolutional attention network

The MTCA is composed of multiple temporal attention blocks (TABlocks). As shown in Figure 1, our TABlocks includes two parts, convolution network and attention layer. Convolutional network is similar to the TCN, which uses depthwise separable convolution DSconv$(\cdot)$ [23] to reduce the model size. This network allows AVTA to obtain different time scales features by setting different convolution kernel dilation coefficients. Then, we concatenate attention layer so that the network focus on the sequence dependence on different time scales. The audio MTCA can be defined as follows:

$$\text{DSconv}\left(\mathbf{E}_a, K_{DS}^d\right) = \mathcal{G}\left(\mathcal{G}\left(\mathbf{E}_a \circledast K_{1\times 1}\right) \circledast K_{DS}^d\right), \tag{5}$$

$$\text{AttLayer} = \text{softmax}\left(\frac{Q_a K_a^T}{\sqrt{d}}\right) V_a, \tag{6}$$

$$\text{TABlock}\left(\mathbf{E}_a, K_{DS}^d\right) = \text{AttLayer}(\text{DSconv}(\mathbf{E}_a, K_{DS}^d)), \tag{7}$$

$$\text{MTCA}\left(\mathbf{E}_a, M\right) = \underbrace{\text{TABlock}(\text{TABlock}(\mathbf{E}_a, K_{DS}^{2^m}) \cdots)}_{M}, \tag{8}$$

where $K_{1\times 1}$ is coefficient matrix of 1D point convolution kernel, $K_{DS}^d$ represents kernel size of DSconv$(\cdot)$ [23] operation with dilation coefficient $d$, and their stride size is 1. $\mathcal{G}(\cdot)$ represents parametric rectified linear unit (PReLU) [24] and global layer normalization (gLN) [8]. AttLayer$(\cdot)$ is attention layer, and the query, key and value of audio attention layer are $Q_a$, $K_a$, $V_a$, respectively, which come from the linear projection of the input matrix. TABlock$(\cdot)$ represents temporal attention block, the stacking of multiple TABlock$(\cdot)$ forms a MTCA $(\cdot)$, $m = 0, 1, 2..., M-1$ represents the number of iterations of TABlock$(\cdot)$. The output audio and visual feature matrix of multi-layer MTCA $(\cdot)$ is $\mathbf{F}_a, \mathbf{F}_v \in \mathbb{R}^{B\times K_a\times N_a}$.

#### 2.2.2. Feature sequence learning network

This network includes MTCA and the cross-attention layer. The cross-attention layer contains audio-to-video and video-to-audio attention layer. This cross-attention layer is the similar to Eq. (6), the query in the audio-to-video attention layer is $Q_v$, the key and value are $K_a$ and $V_a$, respectively [15], and vice versa to video-to-audio attention layer. Such a structure makes audio and visual information no longer isolated, but interrelated, which helps us to fuse the audio and visual information.

### 2.3. Audio-visual fusion and separation network

The fusion process of audio and visual feature $\mathbf{F}_a$, $\mathbf{F}_v$ can be expressed by the following formula, :

$$\mathbf{U}_a = \text{norm}(\text{PointConv}(\mathbf{F}_a, N_a, N_a^{\text{fusion}})), \tag{9}$$

$$\mathbf{U}_v = \text{norm}(\text{PointConv}(\mathbf{F}_v, N_a, N_v^{\text{fusion}})), \tag{10}$$

$$\mathbf{V}_{av} = \text{MTCA}(\text{concat}\left[\mathbf{U}_a, \mathbf{U}_v\right]), \tag{11}$$

where PointConv$(\cdot)$ is 1D point convolution, its input dimension are $N_a$, and output dimensions are $N_a^{\text{fusion}}$ and $N_v^{\text{fusion}}$ respectively. The 1D point convolution PointConv$(\cdot)$ is used to adjust the weight distribution of each audio-visual feature, and make condition $N_a^{\text{fusion}} + N_v^{\text{fusion}} = N_a$ satisfy. The norm$(\cdot)$ is layer normalization, which ensures the scale invariance of the data during the fusion of $\mathbf{F}_a$ and $\mathbf{F}_v$. Finally, we will splice $\mathbf{U}_a, \mathbf{U}_v$ to get $\mathbf{U}_{av} \in \mathbb{R}^{B \times K_a \times N_a}$. The $\mathbf{U}_{\text{av}}$ will then be fed into the MTCA to model the contextual information of the fusion feature sequence again, and obtains the final audio-visual fusion feature $\mathbf{V}_{av} \in \mathbb{R}^{B \times K_a \times N_a}$.

The separation network is mainly composed of DPRNN [9], which takes the fused audio-visual feature sequence $\mathbf{V}_{av}$ as input and obtains the prediction mask $\hat{\mathbf{M}} \in \mathbb{R}^{C \times B \times K_a \times N_a}$.

### 2.4. Speaker waveform prediction

We combine the predicted mask $\hat{\mathbf{M}}$ to calculate the speech waveform by the following formula:

$$\hat{s}_{a_i} = \text{TransConv1d}\left(\mathbf{V}_{av} \odot \hat{\mathbf{M}}_i, C_{in}, C_{out}\right), i = 1, ..., C. \quad (12)$$

where TransConv1d$(\cdot)$ is Decoder, which represents a 1D transposed convolution operation, $\hat{\mathbf{M}}_i \in \mathbb{R}^{B \times K_a \times N_a}$ denotes the prediction mask of the $i$-th speaker, $C_{in} = N_a$, $C_{out} = 1$ , and $\odot$ denotes the element-wise multiplication, i.e. matrix elements are multiplied correspondingly. $\hat{s}_{a_i} \in \mathbb{R}^{B \times L_{seqa} \times 1}$ represents the reconstructed $i$-th speaker time-domain speech signal.

## 3. Experiments

### 3.1. Dataset

In this paper, we use the GRID [25] and VoxCeleb2 (Vox2) [26] datasets to verify the effectiveness of AVTA. GRID dataset includes 34 speakers, each speaker has 1000 frontal face recordings, each recording has a duration of 3 seconds. Vox2 dataset contains over 1 million utterances, and all face tracks for each speaker are extracted from YouTube videos, with 5994 speakers in training set and 118 speakers in test set. All video sampling rate in this paper is 25 frames per second (FPS), the audio sampling rate is 8 kHz, where the length of segments $T_a, T_v$ are 2 seconds. We use S3FD [27] to extract mouth feature landmarks. In order to ensure the randomness of the mixture in GRID dataset, we randomly order the speakers to mix them. The training sets for GRID, Vox2 datasets are 7.8 and 248.7 hours, respectively, the validation sets are 1.1 and 5.0 hours respectively, the test sets are 0.3 and 5.0 hours, respectively.

### 3.2. Experiment configurations

In AVTA, we set the convolutional stride of audio encoder to half the kernel size, and $C_{in}$ is set to 1 to accommodate the 1D time-domain signal. The number of input channels of the visual encoder is 40, the size of the convolution kernel is 3, and the stride is 1. The size of the convolution kernel in TABlock is set to 3, the stride is 1, other parameter settings are shown in Section 4. We use 6-layer DPRNN as separation network [9]. In the training parameters, we set the epoch to 100, the initial learning rate to 1.5e-4, the batch size to 8, and use the Adam [28] optimizer. In the process of training, if the performance of 8 consecutive epochs does not improve, we will reduce the learning rate by half, and when 10 epochs do not show better performance, we manually stop the training. AVTA uses uPIT [29] to solve the source permutation problem, and uses automatic mixed-precision to reduce training time on Vox2 dataset. All experiments are conducted on the NVIDIA RTX 3090 GPU.

## 4. RESULTS

For all results, MTCA Layer is the number of MTCA network layers, $M$ is the number of TABlocks in each MTCA, $N_a$ is the feature dimension of audio coded sequences, KS and Stride represent the size and stride of the 1D convolution kernel in the audio encoder, respectively, $K_a$ is the coded output sequences length of the audio encoder. SI-SNRi and SDRi [8, 30] are used to evaluate separation performance, and their units are dB. MACs represents multiply-accumulate operations. Training time and GPU memory are from model evaluation results under a batch size of 8, and their units are ms and GB respectively. Param represents the model size.

### 4.1. Optimize network parameters

In Table 1, we use the GRID dataset to get 2-speaker mixture to evaluate the effect of different network parameters on the speech separation performance, where KS is set to 40.

Table 1: *Effect of different configurations in AVTA.*

| $M$ | MTCA Layer | $Na$ | SI-SNRi | SDRi | MACs (G/s) | Training Time | GPU Memory | Param |
|---|---|---|---|---|---|---|---|---|
| 2 | 2 | 128 | 12.73 | 13.90 | 2.54 | 125 | 7.49 | 1.8M |
| 4 | 2 | 128 | 12.88 | 14.08 | 3.06 | 163 | 9.90 | 2.5M |
| 8 | 2 | 128 | 13.36 | 14.56 | 4.11 | 231 | 14.78 | 3.8M |
| 2 | 4 | 128 | 12.92 | 14.13 | 3.06 | 165 | 9.92 | 2.5M |
| 2 | 8 | 128 | 13.39 | 14.62 | 4.11 | 235 | 14.77 | 3.8M |
| 4 | 4 | 128 | 13.41 | 14.63 | 4.11 | 229 | 14.63 | 3.8M |
| 4 | 8 | 128 | 13.62 | 14.82 | 6.21 | 323 | 24.28 | 6.5M |
| 8 | 4 | 128 | 13.53 | 14.74 | 6.21 | 323 | 24.22 | 6.5M |
| 4 | 4 | 256 | 14.51 | 15.72 | 11.09 | 280 | 16.67 | 12.6M |

We find that increasing the MTCA Layer, or increasing TABlocks $M$, can improve the separation performance of the network. However, when the $M$ and MTCA layer is increased to 4, 8 or 8, 4, their performance improvement is reduced. The potential reason is that under the current configuration, the network temporal receptive field reachs the limit of the speech time sequences, resulting in limited performance. When feature dimension $N_a$ increases, the separation performance of the model will be improved, but it will lead to a sharp increase in training time, GPU memory, model complexity and size. In summary, we choose the configuration where $M$ is 4, MTCA layer is 4, and $N_a$ is 128 to better balance model training cost and separation performance.

Table 2: *Performance comparison of different network structures modeling of AVTA on GRID dataset.*

| Method | KS | Stride | $K_a$ | SI-SNRi | SDRi |
|---|---|---|---|---|---|
| AVTA | 40 | 20 | 799 | 13.41 | 14.63 |
| AVTA-32 | 32 | 16 | 999 | 14.09 | 15.31 |
| AVTA-CatLiner | 40 | 20 | 799 | 12.31 | 13.52 |
| AVTA-Add | 40 | 20 | 799 | 12.27 | 13.48 |
| AVTA-NoCross | 40 | 20 | 799 | 13.25 | 14.44 |
| AVTA-NoDS | 40 | 20 | 799 | 12.39 | 13.58 |
| AVTA-NoVisual | 40 | 20 | 799 | 13.27 | 14.47 |

Table 2 compares the separation performance of different network structures modeling of AVTA. In the table, AVTA is the method proposed in this paper, the fusion network of AVTA-CatLiner is a traditional feature splicing and linear mapping structure. The fusion network of AVTA-Add is a simple structure of adding audio and visual features, AVTA-NoCross removes the audio-visual cross-attention layer, AVTA-NoDS removes the convolutional network in MTCA, AVTA-NoVisual

Table 3: *Training cost and model complexity evaluation on GRID dataset.*

| Method | KS | $K_a$ | MACs (G/s) | Training Time | GPU Memory | Param |
|---|---|---|---|---|---|---|
| Conv-TasNet [8] | 16 | 1999 | 9.96 | 210 | 10.11 | 5.1M |
| Conv-TasNet1 | 10 | 3199 | 15.94 | 306 | 14.93 | 5.1M |
| DPRNN [9] | 2 | 15999 | 84.89 | 500 | 28.20 | 2.6M |
| AVTA | 40 | 799 | 4.11 | 229 | 14.63 | 3.8M |
| AVTA-32 | 32 | 999 | 5.14 | 292 | 20.50 | 3.8M |
| AVTA-CatLiner | 40 | 799 | 2.03 | 104 | 5.07 | 1.2M |
| AVTA-Add | 40 | 799 | 2.03 | 105 | 5.05 | 1.2M |
| AVTA-NoCross | 40 | 799 | 3.90 | 216 | 13.63 | 3.6M |
| AVTA-NoDS | 40 | 799 | 3.48 | 197 | 13.03 | 3.0M |
| AVTA-NoVisual | 40 | 799 | 4.01 | 219 | 14.19 | 3.7M |

removes the visual part of the network. We find that fusion network of AVTA can improve the separation performance better than other fusion methods. Compared with AVTA-NoCross, AVTA adds audio-visual cross-correlation information before fusion, which helps separation. The AVTA-NoVisual removes visual information from the AVTA network, leading to a drop in separation performance, it illustrates the effectiveness of visual information in audio-visual separation network of AVTA. Combining with Table 3, we find that all AVTA with different network structures are efficient enough, and the separation performance of AVTA is better. In addition, AVTA-32 sets a smaller KS. Obviously, its coded sequences $K_a$ is longer, and its performance is improved, but longer coded sequences directly leads to a sharp increase in the complexity and training cost of the model. In order to adjust such an imbalance, we set AVTA audio convolution kernel size KS to 40.

### 4.2. Visualization

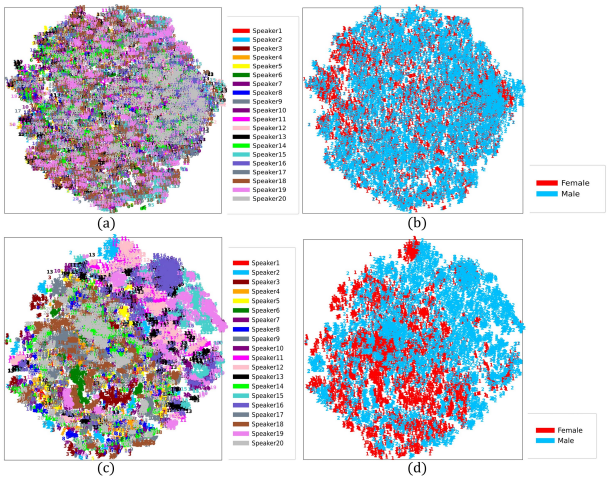


Figure 2: *We use t-SNE to reduce the dimensionality of the encoder output and fused features respectively, and visualize the spatial distribution. We selected a total of 20 random speakers in the Vox2 test set, including 10 males and 10 females.*

In visualization, Figure 2(a), 2(b) represent the encoder output features, Figure 2(c), 2(d) represent the audio-visual fusion output features, in Figure 2(a), 2(c), different colors (or speaker id) correspond to different speakers. In Figure 2(b), 2(d), different colors (or speaker id) correspond to different genders of speakers. We find that the clustering trend is more obvious in Figure 2(c), 2(d) compared to Figure 2(a), 2(b). The underlying reason is that through the process of feature sequence learning and fusion, the audio and visual information is correlated with each other, and the fused features of the same speaker and the same gender are clustered in space. Therefore, AVTA sequence learning and fusion network facilitates the separation process.

### 4.3. Method comparison

Table 4: *Performance comparison of different methods under GRID and Vox2 datasets.*

| Dataset | Method | KS | $K_a$ | SI-SNRi | SDRi |
|---|---|---|---|---|---|
| GRID | Conv-TasNet [8] | 16 | 1999 | 12.25 | 13.45 |
| | Conv-TasNet1 | 10 | 3199 | 12.53 | 13.73 |
| | DPRNN [9] | 2 | 15999 | 13.38 | 14.59 |
| | AVTA | **40** | **799** | **13.41** | **14.63** |
| Vox2 | Conv-TasNet [8] | 16 | 1999 | 10.74 | 11.59 |
| | Conv-TasNet1 | 10 | 3199 | 10.97 | 11.83 |
| | DPRNN [9] | 2 | 15999 | 12.10 | 12.95 |
| | AVTA | **40** | **799** | **12.15** | **13.01** |

In Tables 3 and 4, we find that the baseline methods Conv-TasNet and DPRNN [8, 9] can improve performance by setting smaller audio encoder convolution kernels, but coded sequence will also become longer, leading to an increase in model complexity and training cost. Under the optimal configuration of original paper [8, 9], the model size of DPRNN is slightly smaller than AVTA, but MACs are about 20 times of AVTA, training time and GPU memory are about 2 times of AVTA, the separation performance is slightly lower than AVTA, and the length of the coded sequence has already increased to the limit. Although Conv-TasNet has slightly less training time and GPU memory, its separation performance, MACs, and model size are far worse than AVTA. Conv-TasNet1 tries to reduce KS to improve performance, but at this time the MACs, training time and GPU memory, separation performance and model size are all worse than AVTA, continuing to reduce KS will only further increase training cost and computational complexity, resulting in a more unbalanced model. To summarize, AVTA optimizes fusion network, reduces length of coded sequences, training cost, and achieves competitive separation performance compared to the baseline methods. In addition, AVTA achieves a better trade-off in model computational complexity, training cost and separation performance.

## 5. Conclusion

In this paper, we design a MTCA network for time sequences, build an audio-visual feature sequence learning and fusion network based on MTCA, and finally design a novel audio-visual speech separation framework AVTA. In this framework, we utilize multi-layer MTCA to model the contextual information, audio and visual cross-modal relationships to fuse audio-visual features and separate speech of speakers. Finally, we conduct comparative experiments on challenging datasets, the results show that AVTA achieves excellent separation performance on 2 datasets, and the model has a better trade-off between training cost and separation performance. For future works, we will continue to focus on the cross-modal learning and modeling process of multimodal features, design more efficient separation models, and further optimize the network structure.

## 6. Acknowledgements

This paper is supported by the National Natural Science Foundation of China (No. 61671095, 61702065, 61701067, 61771085, 62201113), and the Natural Science Foundation of Chongqing (No.cstc2021jcyj-msxmX0836).

# 7. References


[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 1759–1763, 2014.

[4] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2016-May, pp. 31–35, 2016.

[5] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-Independent Speech Separation with Deep Attractor Network," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[6] D. Yu, M. Kolbæk, Z. Tan, and J. H. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–245, 2017.

[7] Y. Luo and N. Mesgarani, "TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April, pp. 696–700, 2018.

[8] Y. Luo and N. Mesgarani, "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, 2019.

[9] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 46–50, 2020.

[10] E. Zion Golumbic, G. B. Cogan, C. E. Schroeder, and D. Poeppel, "Visual input enhances selective speech envelope tracking in auditory cortex at a "Cocktail Party"," *Journal of Neuroscience*, vol. 33, no. 4, pp. 1417–1426, 2013.

[11] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, pp. 117–128, 2018.

[12] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg, "Seeing Through Noise: Visually Driven Speaker Separation And Enhancement," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 3051–3055, 2018.

[13] R. Tsunoda, R. Aihara, R. Takashima, T. Takiguchi, and Y. Imai, "Speaker-targeted audio-visual speech recognition using a hybrid ctc/attention model with interference loss," *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 251–255, 2022.

[14] T. Afouras, J. S. Chung, and A. Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-September, pp. 4295–4299, 2019.

[15] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking?: Exploring long-term temporal features for audio-visual active speaker detection," *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[16] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, 2018.

[17] R. Lu, Z. Duan, and C. Zhang, "Audio-Visual Deep Clustering for Speech Separation," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 27, no. 11, pp. 1697–1712, 2019.

[18] J. Wu, Y. Xu, S.-X. Zhang, L. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 667–673, 2019.

[19] R. Gao and K. Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15490–15500, 2021.

[20] C. S. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. Hager, "Temporal convolutional networks for action segmentation and detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1003–1012, 2017.

[21] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25, 2021.

[22] A. F. Agarap, "Deep learning using rectified linear units (relu)," *ArXiv*, vol. abs/1803.08375, 2018.

[23] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807, 2017.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015.

[25] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[26] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *INTERSPEECH*, 2018.

[27] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Li, "S3fd: Single shot scale-invariant face detector," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 192–201, 2017.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[29] M. Kolbæk, D. Yu, Z. H. Tan, and J. Jensen, "Multitalker 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.

[30] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1462–1469, 2006.