# TIME-DOMAIN AUDIO-VISUAL SPEECH SEPARATION ON LOW QUALITY VIDEOS

*Yifei Wu*[1], *Chenda Li*[1], *Jinfeng Bai*[2], *Zhongqin Wu*[2], *Yanmin Qian*[1,†]

[1]MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
[2]TAL Education Group, China

## ABSTRACT

Incorporating visual information is a promising approach to improve the performance of speech separation. Many related works have been conducted and provide inspiring results. However, low quality videos appear commonly in real scenarios, which may significantly degrade the performance of normal audio-visual speech separation system. In this paper, we propose a new structure to fuse the audio and visual features, which uses the audio feature to select relevant visual features by utilizing the attention mechanism. A Conv-TasNet based model is combined with the proposed attention-based multi-modal fusion, trained with proper data augmentation and evaluated with 3 categories of low quality videos. The experimental results show that our system outperforms the baseline which simply concatenates the audio and visual features when training with normal or low quality data, and is robust to low quality video inputs at inference time.

***Index Terms***— Audio-visual, Speech Separation, Low Quality Video, Attention

## 1. INTRODUCTION

Speech separation plays an important role in addressing the "cocktail party problem" [1]. It aims to separate the clean speech for different speakers in a speech mixture of multiple speakers. The neural network based single-channel speech separation has been developed rapidly in recent years [2–10]. Most of these works focus on the audio-only blind source separation (BSS) without any additional knowledge. The label permutation problem [2] in BSS can be solved by deep clustering [2] or permutation invariant training [4, 5].

Another promising approach for single-channel speech separation is to add external information as clues to guide the separation. There are various forms of external information that can be leveraged for speech separation in real application, including the pre-enrolled speaker identity [11, 12], text or contextual [13, 14], visual clues, and brain-informed speech separation [15]. As one of the most convenient clues to collect in both training and real application, visual information of speakers has been introduced into the speech separation system in many prior works [14, 16–22], which shows the vast potential of the visual clues.

Despite the excellent ability of the visual information to aid the separation process, there are also many cases where the videos of the talkers are absent or of low quality. Such visual inputs are harmful to a well-trained audio-visual separation system, sometimes even degrades its performance below an audio-only one. Although some works [23–25] have tried to address this problem, it still remains as an unsolved problem that how we could squeeze more value from the low quality videos.

A previous work [24, 25] on speech enhancement uses a strategy that automatically switches between audio-only variational auto-encoder (VAE) and audio-visual VAE for noisy and clean video frames which is learned in an unsupervised way. It turns out to be efficient, but the information in the low quality frames is almost discarded. Another work [23] proposed a method to enhance a speech by conditioning on the talker's lip movements and optionally a speaker embedding, where the enhanced speech needs to be obtained in a second-pass manner. It gives impressive results on both enhancement and recognition, but to obtain a satisfying output, either a robust speaker model or some high quality videos should be provided.

In this paper, we extend the attention-based feature fusion method proposed in our prior work [26] to the speech separation task. The attention-based fusion method is incorporated into the Conv-TasNet [8] architecture to build a time-domain audio-visual speech separation system that is robust to low quality video inputs. The proposed system utilizes the audio feature to select the most relevant visual features within a time window to extract the necessary information for the separation process. We also present 3 common categories of low quality videos and verify the performance of our proposed model with corresponding data augmentations. For comparison, we replace the attention-based feature fusion part with a concatenation-based fusion block as the baseline. Experiments demonstrate that our method outperforms the usual concatenation method on both normal videos and low quality videos. Besides, with suitable data augmentation introduced into training, the performance of our proposed model is robust to unseen low quality videos.

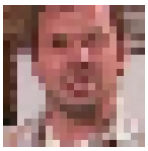
## 2. TASK DEFINITION

In this section, we firstly provide our definition of the audio-visual speech separation task and then introduce three common categories of low quality videos in the real life.

### 2.1. Audio-visual speech separation

The audio-visual speech separation task is to extract the speech signal of each talker from a speech mixture of multiple talkers. For simplicity, we always assume that there are 2 talkers talking simultaneously. Let $\mathbf{x}_a$ denote the speech mixture of the 2 talkers of length $T_a$, and $\mathbf{X}_{v1}, \mathbf{X}_{v2}$ denote the videos containing every talker's face, each of length $T_{v1}, T_{v2}$. Our aim is to build a model $f$ which takes $(\mathbf{x}_a, \mathbf{X}_{v1}, \mathbf{X}_{v2})$ as input, and outputs the clean speech signals $(\mathbf{y}_1, \mathbf{y}_2)$ of each talker.

† Yanmin Qian is the corresponding author.


(a) Low Resolution


(b) Lip Concealment$_1$


(c) Lip Concealment$_2$


(d) Out Of Sync

**Fig. 1**. Common categories of low quality videos.

### 2.2. Types of low quality videos

Real applications often have to process low quality videos, which may notably degrade the system performance. From the aspect of audio-visual speech separation, there are mainly 3 common categories of low quality videos: low resolution, lip concealment and out-of-sync. The examples are illustrated in Fig.1:

**1. Low resolution** may be caused by a low-level camera or a talker's face in a far distance. In this scenario, it is difficult for a system trained with good-resolution to extract useful visual features.

**2. Lip concealment** stands for occasions in which the lips of the talker are partially or totally concealed. Since lips provide most of the information of the speech in a video [16], a video clip without lips exposed could provide little help for the system.

**3. Out-of-sync**, namely asynchronization of audio and video in time, commonly exists in live broadcasts. This makes the system trained on synchronized data hard to extract corresponding audio and visual features, thus affects the system performance.

## 3. METHOD

In this section, we propose a new model based on Conv-Tasnet to address the audio-visual speech separation task. We also introduce 3 data augmentation to address the 3 categories of low quality videos.

### 3.1. Model architecture

The architecture of our proposed model is shown in Fig.2, which mainly follows the Conv-TasNet [8] architecture while adding visual features as additional inputs. It consists of 7 parts: visual feature extractor, audio encoder, visual convolution, audio convolution, attention-based feature fusion, mixture convolution and audio decoder. The visual feature extractor will be described in Sec.4.1

The audio encoder and decoder perform the transformation between the time-domain audio signal and the embedded audio feature sequence through 1D convolution and deconvolution operations. Formally, we have

$$\begin{aligned} \text{Encoder}_a(\mathbf{x}_a) &= \text{ReLU}(\text{Conv1D}(\mathbf{x}_a), K, S) \\ \text{Decoder}_a(\mathbf{O}) &= \text{DeConv1D}(\mathbf{O}, K, S) \end{aligned} \tag{1}$$

where $\mathbf{O}$ is the feature sequence input to the audio decoder. $K$ is the kernel size and $S$ is the stride size in convolutions.

The audio convolution part and mixture convolution part each includes several 1D convolutional blocks. They are organized as $R_a$ and $R_m$ repeats where each repeat contains $X_a$ or $X_m$ stacked blocks with the convolutional dilation factors exponentially increasing from 1. The detailed structure of the block is similar to the one presented in Conv-TasNet. The visual convolution part is a stack of 1D convolutional layers accepting visual feature vector sequences. There are a ReLU activation and a batch normalization operation between every two layers.

The attention-based feature fusion part is mainly adopted from the *Query Vision* structure proposed in our previous work [26]. It exploits the attention mechanism proposed in [27] which transform the inputs into queries, keys and values to collect related features by calculating a weighted sum.

$$\text{Attention}(\mathbf{M}_1, \mathbf{M}_2) = \text{softmax}(\frac{\text{Query}(\mathbf{M}_1)\text{Key}(\mathbf{M}_2)^\top}{\sqrt{d}})\text{Value}(\mathbf{M}_2) \tag{2}$$

where $d$ is the feature dimension. This structure allows our model to focus on visual features which are more valuable and relevant to the current audio feature frame, thus helps the model to extract useful features from the low quality inputs. Denoting the output of audio convolution part by $\mathbf{F}_a$ and the outputs of visual convolution part by $\mathbf{F}_{v1}, \mathbf{F}_{v2}$, the feature fusion procedure can be represented as:

$$\begin{aligned} \mathbf{F}'_a &= \text{LayerNorm}(\mathbf{F}_a) \\ \mathbf{F}'_{vk} &= \text{LayerNorm}(\mathbf{F}_{vk}), k = 1, 2 \\ \hat{\mathbf{F}}_{vk} &= \text{Attention}(\mathbf{F}'_a, \mathbf{F}'_{vk}), k = 1, 2 \\ \mathbf{F}_m &= \text{Projection}(\text{Concat}(\mathbf{F}'_a, \hat{\mathbf{F}}_{v1}, \hat{\mathbf{F}}_{v2})) \\ \hat{\mathbf{F}}_m &= \text{FeedForward}(\text{Dropout}(\text{LayerNorm}(\mathbf{F}_m))) \end{aligned} \tag{3}$$

Despite the fact that *Query Vision* works well in the audio-visual multi-talker ASR task, it turns out that the model converges very slowly due to the much longer sequence length on the time-domain setup. Thus, we adapt the local attention mechanism [28] to our model. Assume that $L_a, L_v$ are the lengths of audio and visual feature sequences, respectively. For the $i$-th query, only the scores produced by visual features with indices ranging in $[iL_v/L_a - D, iL_v/L_a + D]$ will be considered for the weighted sum. Here $D$ is artificially designed. Compared to the long-term dependency, the neighbor frames could empirically provide most of the necessary information in speech separation.

### 3.2. Data augmentation

To address the issues caused by low quality videos described in Section 2.2, we present 3 data augmentation methods to improve the model's robustness. Note that in practice we always assume that each talker's face could be located in every frame of the video even with the augmentation since frames without a face could be mainly categorized as lip concealment in our experiments.

**A. Low resolution:** Since the visual feature extractor requires inputs with a fixed resolution, videos are firstly down-sampled to a low-resolution and then up-sampled to the input resolution, both with the nearest neighbor interpolating algorithm. Results obtained by this could hopefully be a guideline for other extractors with flexible input resolution. An augmented example is shown in Fig.1(a).

**B. Lip concealment:** We mainly adopt the augmentation method proposed by [23], where the lip region in some consecutive frames of the video is concealed by a uniform noise square. See Fig.1(b) for example. While in testing, the concealment is made with emoji pictures of the same size as the noise square as Fig.1(c) to simulate a real situation.

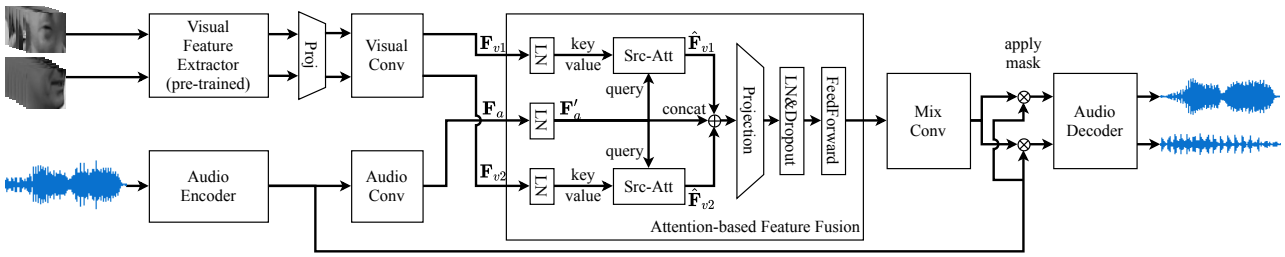



**Fig. 2**. Our proposed Audio-Visual Speech Separation model. The extracted visual features of the 2 talkers share the same parameters of the projection layer and the visual convolution part. "LN" stands for layer normalization and "Src-Att" stands for source attention. The 2 source attention modules share the same parameters. $\oplus$ represents concatenation operation. $\otimes$ represents element-wise multiplication operation.

**C. Random audio-video offset:** The extracted input visual features are advanced or delayed by several frames, thus the audio and video are out of synchronization. The advanced or delayed space is filled with the last or the first frame.

### 3.3. Loss function

The loss function is defined as the scale-invariant signal-to-noise ratio between each predicted signal and the corresponding reference signal. The permutation is assigned according to the visual input order. Formally,

$$\mathcal{L}(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \mathbf{y}_1, \mathbf{y}_2) = \frac{1}{2}(\text{SI-SNR}(\hat{\mathbf{y}}_1, \mathbf{y}_1) + \text{SI-SNR}(\hat{\mathbf{y}}_2, \mathbf{y}_2)) \quad (4)$$

## 4. EXPERIMENTS

### 4.1. Data preparation

Experiments were done on LRS2 [29] dataset consisting of videos collected from BBC television programs. All the videos and the corresponding audios are synchronized. The dataset is already divided into pretrain, train, val and test sets. There are about 183k two-talker mixtures generated by randomly selecting and summing 2 utterances from train set for training, about 1k from val set for validation and about 1k from test set for evaluation. All the mixtures are mixed with SNR in $[-10, 10]$ dB randomly. The videos are of 25fps and 160x160 resolution, and the audios are recorded at 16kHz sample rate. All videos and audios are clipped or zero-padded to 2.4s.

A lip-reading model was pre-trained on LRW dataset following the recipe[1] in [30, 31], and its ResNet frontend acts as the visual feature extractor to extract 512-dimensional features from the mouth region of the video.

### 4.2. Training with Data Augmentation

Videos in the train set are augmented to different levels for each data augmentation type. For low resolution, videos are augmented to 80x80, 40x40 and 20x20 in resolution. For lip concealment, a noise square with side length 60px is patched on a 25%, 50% or 75% consecutive duration of the videos, starting from a random time. For random audio-video offset, a maximum 5 frames random offset is given to each video online when training.

To validate the improvement brought by the proposed data augmentations, we firstly conduct experiments with one of the three data augmentation methods respectively. Then, all the proposed augmentation methods are combined together in the training.

[1] official impl.: https://github.com/mpc001/end-to-end-lipreading

**Table 1**. Hyper-parameters of our model. The definition of each symbol is the same as in Conv-TasNet [8]. The subscript "a" stands for the audio convolution part, and "m" stands for the mixture convolution part.

| $N$ | $L$ | $B$ | $H$ | $P$ | $X_a$ | $X_m$ | $R_a$ | $R_m$ |
|---|---|---|---|---|---|---|---|---|
| 256 | 20 | 256 | 512 | 3 | 8 | 8 | 2 | 2 |

Another comparison is performed on the number of augmented visual streams $Q$. $Q \in \{0, 1, 2\}$ denotes no augmentation, augmentation for one of the visual streams, and augmentation for both of the visual streams, respectively.

### 4.3. Evaluation

The signal-to-distortion ratio (SDR) is adopted as the evaluation metric by our experiments.

To evaluate the model's performance under bad conditions, we prepare 3 low quality test sets by modifying the normal test set.

**LR10:** Videos are replaced by the 10x10 low-resolution version.

**LE75:** Each video is patched with an emoji picture of 60px side length on a 75% consecutive duration of the video, starting from a random time offset.

**RO10:** Videos are randomly shifted for $\lambda$ frames to be out of sync with the audio. Where $\lambda$ is uniformly chosen in $[-10, 10]$.

Two strategies are adopted in the bad condition evaluation. The first is to choose one visual stream and replace it with its low quality version, while the second is to replace both of the visual streams.

### 4.4. Experimental configurations

We built and evaluated our model on ESPNet-SE [32] framework. The main hyper-parameters of our model are presented in Table 1, which are defined the same as in [8]. The visual convolution part contains 5 layers with strides 1, 0.5, 1, 0.5, 1, respectively, where a stride smaller than 1 stands for a deconvolution layer. The query, key and value vectors are of 256 dims and the local attention range $D$ is 5. We also prepared a baseline model which replaces the feature fusion part of our proposed model with a 1D convolutional layer with $B \times (1 + C)$ input channels and $B$ output channels. The kernel size of the convolutional layer is $P$ and the padding length is $\lfloor P/2 \rfloor$. The models are trained until convergence with 16 as batch size using Adam optimizer. The learning rate is $10^{-3}$ and 1 epoch of warming up is used for our proposed model. For the baseline model, the learning rate is halved when there is no improvement on validation result. All the models are trained on 8 GPUs. Besides, the order of the reference signals is determined by their energy to help training.

**Table 2**. Performance comparison between the baseline model and the proposed model trained with different configurations and evaluated with normal and 3 types of unseen low quality videos. Both the SDRs evaluated with one/two low quality visual streams are provided. An audio-only setup is also presented for reference.

| $Q$ | Data Augmentation | Model | SDR(dB) | | | |
|---|---|---|---|---|---|---|
| | | | Normal | LR10 | LE75 | RO10 |
| 0 | None | Audio-only | 12.47 | - | - | - |
| | | Baseline | 13.45 | 12.54/9.67 | 12.89/11.59 | 10.54/6.10 |
| | | Proposed | **14.66** | **14.09/11.55** | **14.06/12.57** | **11.89/6.94** |
| 1 | Low Resolution | Baseline | 13.64 | 12.97/11.02 | 13.28/12.50 | 11.28/7.27 |
| | | Proposed | **14.86** | **14.53/13.29** | **14.47/13.65** | **12.69/8.23** |
| | Lip Concealment | Baseline | 13.75 | 13.53/12.35 | 13.56/12.95 | 11.40/7.52 |
| | | Proposed | **14.77** | **14.48/13.30** | **14.48/13.92** | **12.63/8.36** |
| | Max. 5 Frames Async. | Baseline | 13.08 | 12.85/11.80 | 12.77/12.04 | 12.76/**10.26** |
| | | Proposed | **14.17** | **13.91/13.11** | **13.87/13.35** | **12.89**/10.10 |
| | All | Baseline | 12.87 | 12.74/12.28 | 12.73/12.36 | 12.14/9.92 |
| | | Proposed | **14.34** | **14.16/13.64** | **14.20/13.90** | **13.03/10.53** |
| 2 | Low Resolution | Baseline | 13.27 | 13.24/12.73 | 13.12/12.72 | 10.88/7.40 |
| | | Proposed | **14.81** | **14.56/13.72** | **14.48/13.93** | **12.74/8.87** |
| | Lip Concealment | Baseline | 13.59 | 13.44/12.85 | 13.42/13.07 | 11.57/8.71 |
| | | Proposed | **14.67** | **14.45/13.75** | **14.48/14.11** | **12.80/9.47** |
| | Max. 5 Frames Async. | Baseline | 13.10 | 12.88/12.41 | 12.67/12.33 | 12.31/11.60 |
| | | Proposed | **13.53** | **13.33/13.06** | **13.26/13.01** | **12.81/11.86** |
| | All | Baseline | 12.51 | 12.36/11.98 | 12.35/12.13 | 10.33/7.61 |
| | | Proposed | **14.00** | **13.86/13.42** | **13.85/13.55** | **12.80/10.33** |

### 4.5. Results

The baseline model and the proposed model are trained with $Q = 0$ (without data augmentation), 1 and 2, evaluated on the normal test set and the 3 low quality test sets respectively. The results are listed in Table 2. An audio-only model is also trained as reference by removing the video-related modules from the baseline model (so it is literally Conv-TasNet). It shows that our proposed model outperforms the baseline model with any type of data augmentation and any number of augmented visual streams in training on almost all 4 test sets. Also, it demonstrates that low quality videos may degrade the audio-visual system's performance below the audio-only one.

It could be observed that with each augmentation method, the baseline and the proposed model both give better performance on almost all the test sets, which demonstrates the effectiveness of those methods. The exception is that random offset augmentation degrades both models' performance on the normal set, the LR10 set and the LE75 set with 1 low quality visual stream in inference. One possible explanation is that this augmentation encourages the model to focus more on the audio modality, while the useful information in the high quality visual modality is ignored to some degree. The RO10 set also contributes the lowest SDRs among all the test sets for every configuration even with random offset augmentation, indicating that it is the hardest among the three low quality categories to be addressed.

While both the system benefits from data augmentation when $Q = 1$ and only one method is used, it also does harm to the baseline model evaluated on the normal test set when trained with $Q = 2$. It suggests that this setup could reflect the model's robustness to low quality training data. Experiments done with each augmentation show that in this condition, our proposed model still provides fair or even better results than the ordinary situation. The last two lines of the table show the results of both models trained with all the augmentation methods, where in most cases both visual streams are of low quality. We observe that with not much reduction on the normal test set, our proposed model achieves fair results on LR10 and LE75 set, which are close to the one on the normal set. Besides, the proposed model also gives acceptable performance on the RO10 test set, showing its ability in utilizing the out-of-sync low quality videos.

## 5. CONCLUSIONS

In this paper, we explore the attention-based multi-modal fusion method to build a robust time-domain audio-visual speech separation system. To further improve the performance of our proposed system on low quality video inputs, we introduce 3 types of data augmentation, including low resolution, lip concealment and random offset. The evaluation results on the simulated dataset derived from LRS2 demonstrate that our proposed model outperforms the concatenation-based baseline on all the 3 types of low quality video inputs, and is robust to low quality training dataset.

## 6. ACKNOWLEDGEMENTS

This work was supported by the National Key R&D Program of China (Grant No. 2020AAA0104500) and the China NSFC projects (No. 62122050 and No. 62071288).

## 7. REFERENCES

[1] E Colin 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] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for

segmentation and separation," in *Proc. IEEE ICASSP*. 2016, pp. 31–35, IEEE.

[3] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R. Hershey, "Single-Channel Multi-Speaker Separation using Deep Clustering," in *Proc. ISCA Interspeech*, 2016, pp. 545–549.

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

[5] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. ASLP.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[6] Yi Luo and Nima Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE ICASSP*, 2018, pp. 696–700.

[7] Yuzhou Liu and DeLiang Wang, "Divide and conquer: A deep casa approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. ASLP.*, vol. 27, no. 12, pp. 2092–2102, 2019.

[8] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[9] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE ICASSP*, 2020, pp. 46–50.

[10] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong, "Attention Is All You Need In Speech Separation," in *Proc. IEEE ICASSP*, 2021, pp. 21–25.

[11] Marc Delcroix, Katerina Zmolikova, Keisuke Kinoshita, Atsunori Ogawa, and Tomohiro Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Proc. IEEE ICASSP*, 2018, pp. 5554–5558.

[12] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John Hershey, Rif A Saurous, Ron J Weiss, Ye Jia, and Ignacio Lopez Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. ISCA Interspeech*, 2019, pp. 2728–2732.

[13] Keisuke Kinoshita, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani, "Text-informed speech enhancement with deep neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] Chenda Li and Yanmin Qian, "Listen, Watch and Understand at the Cocktail Party: Audio-Visual-Contextual Speech Separation," in *Proc. ISCA Interspeech*, 2020, pp. 1426–1430.

[15] Enea Ceolini, Jens Hjortkjær, Daniel D. E. Wong, James O'Sullivan, Vinay S. Raghavan, Jose Herrero, Ashesh D. Mehta, Shih-Chii Liu, and Nima Mesgarani, "Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception," *NeuroImage*, vol. 223, pp. 117282, Dec. 2020.

[16] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael 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, pp. 1–11.

[17] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "The Conversation: Deep Audio-Visual Speech Enhancement," in *Proc. ISCA Interspeech*, 2018, pp. 3244–3248.

[18] Jian Wu, Yong Xu, Shi-Xiong Zhang, Lian-Wu Chen, Meng Yu, Lei Xie, and Dong Yu, "Time domain audio visual speech separation," in *Proc. IEEE ASRU*, 2019, pp. 667–673.

[19] Chenda Li and Yanmin Qian, "Deep Audio-Visual Speech Separation with Attention Mechanism," in *Proc. IEEE ICASSP*, 2020, pp. 7314–7318.

[20] Karthik Ramesh, Chao Xing, Wupeng Wang, Dong Wang, and Xiao Chen, "Vset: A Multimodal Transformer for Visual Speech Enhancement," in *Proc. IEEE ICASSP*, 2021, pp. 6658–6662.

[21] Jiyoung Lee, Soo-Whan Chung, Sunok Kim, Hong-Goo Kang, and Kwanghoon Sohn, "Looking into your speech: Learning cross-modal affinity for audio-visual speech separation," in *Proc. IEEE CVPR*, 2021, pp. 1336–1345.

[22] Ruohan Gao and Kristen Grauman, "Visualvoice: Audio-visual speech separation with cross-modal consistency," in *Proc. IEEE CVPR*, 2021, pp. 15495–15505.

[23] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," in *Proc. ISCA Interspeech*, 2019, pp. 4295–4299.

[24] Mostafa Sadeghi and Xavier Alameda-Pineda, "Robust unsupervised audio-visual speech enhancement using a mixture of variational autoencoders," in *Proc. IEEE ICASSP*, 2020, pp. 7534–7538.

[25] Mostafa Sadeghi and Xavier Alameda-Pineda, "Switching variational auto-encoders for noise-agnostic audio-visual speech enhancement," in *Proc. IEEE ICASSP*, 2021, pp. 6663–6667.

[26] Yifei Wu, Chenda Li, Song Yang, Zhongqin Wu, and Yanmin Qian, "Audio-Visual Multi-Talker Speech Recognition in a Cocktail Party," in *Proc. ISCA Interspeech*, 2021, pp. 3021–3025.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[28] Minh-Thang Luong, Hieu Pham, and Christopher D Manning, "Effective approaches to attention-based neural machine translation," in *Proc. EMNLP*, 2015, pp. 1412–1421.

[29] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," in *Proc. IEEE CVPR*, 2017, pp. 3444–3453.

[30] Themos Stafylakis and Georgios Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," in *Proc. ISCA Interspeech*, 2017, pp. 3652–3656.

[31] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-End Audiovisual Speech Recognition," in *Proc. IEEE ICASSP*, 2018, pp. 6548–6552.

[32] Chenda Li, Jing Shi, Wangyou Zhang, et al., "ESPNet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration," in *Proc. IEEE SLT*, 2021, pp. 785–792.