# AUDIO-VISUAL RECOGNITION OF OVERLAPPED SPEECH FOR THE LRS2 DATASET

*Jianwei Yu*[1,2,*], *Shi-Xiong Zhang*[2], *Jian Wu*[3], *Shahram Ghorbani*[4], *Bo Wu*[3], *Shiyin Kang*[3], *Shansong Liu*[1], *Xunying Liu*[1], *Helen Meng*[1], *Dong Yu*[2,3]

[1]The Chinese University of Hong Kong
[2]Tencent AI Lab, Bellevue, USA, [3]Tencent AI Lab, Shenzhen, China
[4]Center of Robust Speech Systems (CRSS), University of Texas at Dallas

## ABSTRACT

Automatic recognition of overlapped speech remains a highly challenging task to date. Motivated by the bimodal nature of human speech perception, this paper investigates the use of audio-visual technologies for overlapped speech recognition. Three issues associated with the construction of audio-visual speech recognition (AVSR) systems are addressed. First, the basic architecture designs i.e. end-to-end and hybrid of AVSR systems are investigated. Second, purposefully designed modality fusion gates are used to robustly integrate the audio and visual features. Third, in contrast to a traditional pipelined architecture containing explicit speech separation and recognition components, a streamlined and integrated AVSR system optimized consistently using the lattice-free MMI (LF-MMI) discriminative criterion is also proposed. The proposed LF-MMI time-delay neural network (TDNN) system establishes the state-of-the-art for the LRS2 dataset. Experiments on overlapped speech simulated from the LRS2 dataset suggest the proposed AVSR system outperformed the audio only baseline LF-MMI DNN system by up to 29.98% absolute in word error rate (WER) reduction, and produced recognition performance comparable to a more complex pipelined system. Consistent performance improvements of 4.89% absolute in WER reduction over the baseline AVSR system using feature fusion are also obtained.

***Index Terms*—** audio-visual speech recognition, overlapped speech, speech separation, multi-modal

## 1. INTRODUCTION

Automatic speech recognition (ASR) of overlapped speech is a highly challenging task to date. The presence of interfering speakers introduces a large mismatch between clean and overlapped speech and significant performance degradation. To this end, previous research efforts were heavily focused on speech separation techniques that can convert mixed speech into speaker dependent signals.

Prior to the deep learning era, computational auditory scene analysis (CASA) [1] approaches containing perceptual cue [2] based time-frequency mixed speech decomposition and grouping stages were often used. Amid the rapid progress brought by deep learning technologies to speech recognition, they have drawn increasing research interests for overlapped speech separation and recognition. Deep neural network (DNN) based speaker turn detection and weighted finite-state transducer (WFST) two-talker decoding approaches were proposed for single channel multi-talker speech recognition in [3]. Deep clustering based separation techniques that use spectrogram embeddings were proposed in [4–6]. Permutation invariant training (PIT) [7] was also developed as a general solution to map single channel, monaural mixed speech inputs to those of individual speakers. When multi-channel microphone arrays are employed, acoustic beaming algorithms [8] or neural network based beamforming architectures [9] can be used to enhance the desired speaker's signal, while attenuating the interference from other speakers.

Motivated by the bimodal nature of human speech perception [2, 10], and the invariance of visual information to acoustic signal corruption, audio-visual speech recognition (AVSR) technologies [11–15] can also be used for overlapped speech separation [16–23] and the back-end recognition component. However, the use of visual modality in the recognition stage of system development for overlapped speech remains limited to date. To the best of our knowledge, the only previous work in this direction was reported in [24].

Three issues need to be addressed when developing such systems. First, the fundamental architecture design of an AVSR system i.e. end-to-end and hybrid needs to be investigated. This issue has been studied recently in ASR [25], while it still remains unclear for AVSR to date. Second, in order to robustly integrate the audio and visual modalities, a careful design of modelling components for modality fusion is required. Traditionally a simple audio-visual feature concatenation based fusion scheme can be used [24]. However, it provides limited flexibility in fusion when the visual inputs become less reliable and poorer in quality [26–28]. Third, state-of-the-art systems developed for overlapped speech recognition are often based on a pipelined architecture containing explicit speech separation and recognition components [18, 29, 30]. The front-end separation components are often optimized using error costs that are different from those used in the back-end recognition components. Moreover, they often require parallel training data containing clean speech as the learning targets, which are impractical to obtain for unseen and potentially mismatched domains or tasks. This can further limit such systems' wider application.

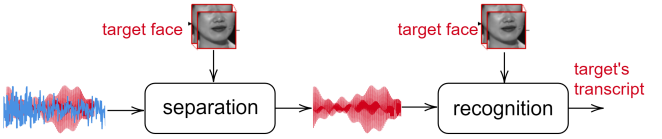
In order to address these issues, a comparison between hybrid system and end-to-end systems is first performed to find a strong basic architecture of AVSR systems. Two gated neural architectures are used to facilitate a dynamic fusion between the audio and visual modalities. A streamlined and integrated AVSR system architecture containing implicit speech enhancement and recognition components optimized consistently using the lattice-free MMI (LF-MMI) discriminative criterion is also proposed. Consistent with ASR [25], the hybrid system show better performance over published end-to-end [11, 31] systems of AVSR on LRS2 [32] dataset. Experiments on overlapped speech simulated from the LRS2 dataset suggest the proposed gated AVSR systems outperforms the audio only baseline LF-MMI time-delay neural network (TDNN) system by up to 29.9% absolute (74% relative) in WER reduction, and produced recogni-

[*] This work was done while the author was an intern at Tencent AI.

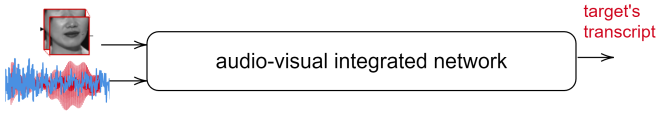
tion performance comparable to a baseline pipelined AVSR system with a more complex speech separation component. Performance improvements of 4.89% absolute (32% relative) in WER reduction over the baseline AVSR system using feature concatenation based modality fusion are also obtained.

The main contributions of this paper includes: 1) this paper is one of the beginning works to compare the performance of hybrid and end-to-end architectures of AVSR, our LF-MMI TDNN system shows state-of-the-art performance on LRS2 dataset; 2) to the best of our knowledge, this paper is the first work using a gated neural network architecture to robustly integrate audio and visual modalities for overlapped speech recognition. In contrast, the only known previous AVSR research for overlapped speech used feature concatenation based fusion [24]; 3) this paper is also the first attempt to use the LF-MMI discriminative criterion to train an integrated AVSR system for overlapped speech. In previously research reported in [33], the LF-MMI criterion was used in a jointly trained pipelined system with audio inputs only, a more complicated system architecture, training procedure and explicit requirement of parallel training data for constructing the separation component.

The rest of this paper is organized as follows. Section 2 reviews pipelined audio-visual separation and recognition baseline systems. Section 3 proposes modality fusion gates based AVSR system architectures. Experiments and results are presented in section 4. Section 5 draws the conclusion and discusses future work.




(a) Pipelined AVSR system




(b) Integrated AVSR system

**Fig. 1**. Illustration of pipelined and integrated AVSR systems.

## 2. PIPELINED AVSR FOR OVERLAPPED SPEECH

This section introduces the architecture of the pipelined AVSR system. The audio-visual separation component and recognition component are introduced in section 2.1 and Section 2.2 respectively.

### 2.1. Audio-visual speech separation

Recent research on leveraging visual modality has led to impressive results in speech separation. In these studies, various representations of the visual information such as lip appearance [17, 18] and optical flow [19, 20] are used to estimate the time-frequency (TF) mask. In this paper, the audio-visual speech separation[1] component used in the pipelined system is based on our previous work in [22]. Given an overlapped audio signal and the target speakers' mouth region of interest (ROI). The system uses the visual information to bias the separation network to directly estimate the TF mask of the target speaker. Then the spectrogram of the separated audio is obtained by multiplying the TF mask with the overlapped spectrogram.

[1]Model details, Si-SNR evaluation and audio samples can be found in: https://yjw123456.github.io/Audio-visual-Overlapped-speech-recognition

### 2.2. Audio-visual speech recognition

For visual modality is still complementary to the separated (enhanced) audio, AVSR system is used as the recognition component in our pipelined system. In the recent studies, the AVSR systems are largely based on end-to-end architectures, such as attention-based encoder-decoder [11, 32], Connectionist-Temporal-Classfication (CTC) [11] and hybrid CTC/attention [31].

Motivated by the impressive results of hybrid system in ASR [25], in this paper, we investigate the hybrid TDNN AVSR system. The structure of the hybrid model used in the pipelined system is shown in Figure 2 (a). The mouth ROI of the target speaker is fed into the LipNet to generated the visual features. The RecogNet is a TDNN network with factored time-delay neural network (TDNN-F) [34] components, which has been shown to be effective in modeling long range temporal dependencies [34]. As we will show in section 4, the hybrid TDNN AVSR system trained with LF-MMI criterion demonstrates the state-of-the-art performance on the LRS2 dataset. We will therefore use the hybrid architecture in the following unless otherwise stated.

## 3. INTEGRATED AVSR FOR OVERLAPPED SPEECH

### 3.1. Audio-visual modality fusion

In this section, we explain the details of three different modality fusion methods used in our AVSR models: feature concatenation, visual modality driven gated fusion, and audio-visual modality driven fusion.

#### *3.1.1. Feature concatenation based fusion*

The baseline AVSR system using feature concatenation is illustrated in Figure 2 (a). The acoustic features are concatenated with the visual features extracted by the LipNet, then the concatenated features are passed to the RecogNet:

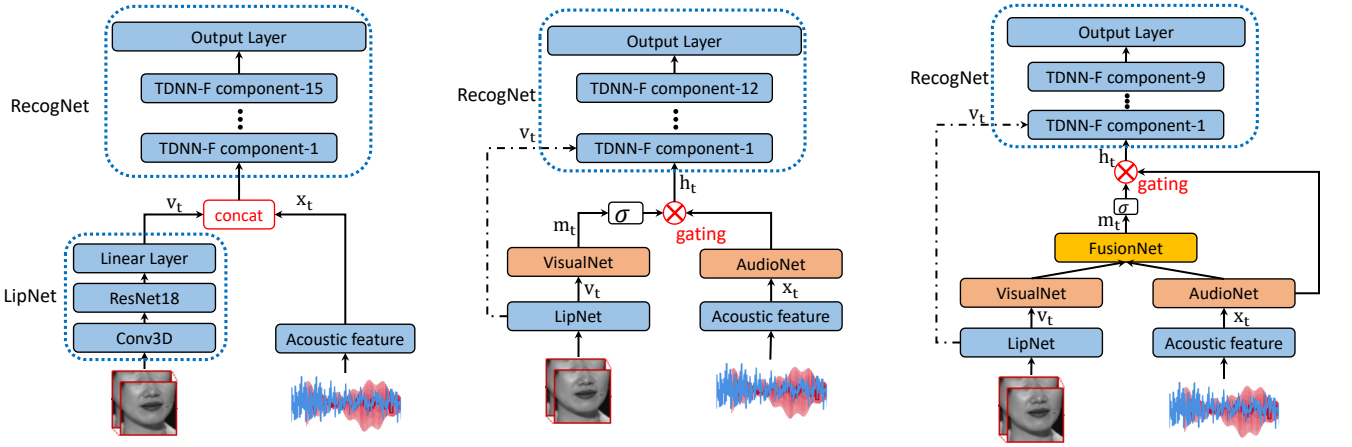
$$p(\mathbf{y_t}|\mathbf{x_t}) = \mathbf{RecogNet}([\mathbf{x_t}, \mathbf{LipNet}(\mathbf{v_t})]), \quad (1)$$

where $\mathbf{y_t}$ is the frame-level alignment of the correspond acoustic frame $\mathbf{x_t}$, $\mathbf{v_t}$ is the mouth ROI of the target speaker.

#### *3.1.2. Visual modality driven gated fusion*

Since visual modality is invariant to the acoustic degradation, the gated architecture is purposefully designed to extract the target speaker from overlapped speech. Compared with concatenation, gating operation is more natural and direct for doing selection. Figure 2 (b) illustrates the structure of the visual modality driven gate. First, the acoustic features and the visual features are passed to ViusalNet and AudioNet networks respectively. Then, the outputs of the AudioNet are gated by the outputs of the VisualNet $m_t$ with an element-wise multiplication:

$$\begin{aligned} \mathbf{m_t} &= \mathbf{VisualNet}(\mathbf{v_t}), \\ \mathbf{h_t} &= \mathbf{AudioNet}(\mathbf{x_t}), \otimes \sigma(\mathbf{m_t}) \\ p(\mathbf{y_t}|\mathbf{x_t}) &= \mathbf{RecogNet}(\mathbf{h_t}), \end{aligned} \quad (2)$$




(a) Feature concatenation (b) Visual modality driven gate (V⊗A) (c) Audio-visual modality driven gate (AV⊗A)

**Fig. 2**. Illustration of audio-visual fusion methods for AVSR systems: (a) feature concatenation based fusion: acoustic and video features are concatenated before fed into the RecogNet; (b) visual modality driven gated fusion; (c) audio-visual modality driven gated fusion. "⊗" denotes Hadamard product. The dashed arrow denotes concatenating the gated hidden outputs with the visual features.

where $\otimes$ denotes the Hadamard product, $\sigma(\cdot)$ is sigmoid function. The AudioNet and VisualNet have a similar architecture, each with 6 TDNN-F layers.

#### *3.1.3. Audio-visual modality driven gated fusion*

Although the visual modality is invariant to the acoustic degradation, it contains less information in terms of words or phones discrimination compared with the audio modality. This can easily be confirmed by the significant performance gap between lipreading and ASR. Therefore, just relying on the visual feature for the gating process might not the most effective approach. In this case, as shown in Figure 2 (c), a so-called FusionNet is added to the visual modality driven gate structure. The outputs of the AudioNet and the VisualNet are concatenated and sent into the FusionNet before used in the gating step:

$$\mathbf{m_t} = \mathbf{FusionNet}([\mathbf{VisualNet}(\mathbf{v_t}), \mathbf{AudioNet}(\mathbf{x_t})]), \quad (3)$$

where the FusionNet is a TDNN network containing 3 tdnn-f layers. We expect that the FusionNet can leverage both the distorted audio and visual information to provide more effective information for the gating step.

### 3.2. Integrated audio-visual system for overlapped speech

In contrast to the pipelined system with explicit separation and recognition components shown in Figure 1 (a), the integrated system as shown in Figure 1 (b) tries to implicitly do both separation and recognition in a compact model architecture using a single recognition cost function. The architectures in Figure 3 (b) and (c) can be viewed as two integrated architectures for overlapped speech recognition. More specifically, in these two models, the front-end part including the gate structure can be regarded as an implicit speech separation component, and the RecogNet in the back-end can be seen as a recognition component. The entire model is trained to optimize the LF-MMI sequence training objective function. Moreover, as the dashed arrow in Figure 2 (b) and (c) shows, the gated hidden outputs are further concatenated with the visual features before passed into the RecogNet. The motivation is that the visual features is still complementary to the gated acoustic representations.

## 4. EXPERIMENT & RESULTS

### 4.1. Experiment setup

**Dataset:** In the experiment, we created two-speaker mixtures using utterances from Oxford-BBC Lip Reading Sentences 2 (LRS2) database [32], which is one of the largest publicly available datasets for lip reading sentences in-the-wild. The database consists of mainly news and talk shows from BBC programs. It is a challenging set since it contains thousands of speakers without speaker labels and large variation in head pose. The dataset is already divided into training, validation (Train-Val) and test sets and also contains a pretraining (Pre-train) set with longer segments. In this experiment Pre-train and Train-Val parts are combined as training set.
**Data simulation:** The overlapped speech utterances are generated by first sampling one reference audio utterance and then mixing its audio with another interfering audio signals. There are always two speakers in the simulated overlapped signals. To ensure the videos of each source are available in a mixture, longer sources are truncated to be aligned with the shortest one. We mix the training data with six level SNRs (15dB, 10dB, 5dB, 0dB, -5dB and clean), both the target and interference data are sampled from the training set.
**Features:** Log Mel-filterbank acoustic features with 40 bins are used, which are extracted with a 40ms window, 10ms hop-length at a sample rate of 16kHz. As for visual inputs, the mouth ROI of LRS2 is already centered, we further crop the center 112 by 112 pixel region of all video frames and up-sample them to 100 frame per seconds using linear interpolation. Throughout this paper, we assume the target speaker's mouth region to be given.
**Model Architectures:** Our speech recognition system is developed based on the Kaldi Toolkit. Since LRS2 doesn't have a dictionary, grapheme-state units are used in our experiment. A GMM-HMM model trained on LRS2 Train-val set is used to generate frame-level alignment. The alignment of the corresponding clean target speech is used for overlapped speech. All recog-

nition systems are trained using the LF-MMI [35] criterion using leaky HMM with the cross-entropy (CE) regularization. All the models are modified based on the default setup of example script "egs/swbd/s5c/local/chain/tuning/run_tdnn_7q.sh" in the Kaldi tookit. The LipNet is pretrained on lipreading task similar to [36]. The details of the audio-visual separation model can be found in our previous work [22].

**Language Model:** The language model is a 4gram language model trained on the transcriptions of the LRS2 Pre-train set which contains more than 2 million words.

### 4.2. Hybrid vs End-to-End of AVSR

In this paper, we compare the performance of our hybrid LF-MMI TDNN system with previous end-to-end system: TM-CTC [11], TM-seq2seq [11], and hybrid CTC/Attention structure [31] on LRS2 dataset. The structure of the hybrid system used in this experiment is shown in Figure 2 (a). Results in Table 1 illustrates that our LF-MMI TDNN system significantly outperforms CE trained TDNN system and the previous state-of-the-art end-to-end models in visual-only (lipreading) , audio-only and audio-visual speech recognition tasks. The performance of CE trained TDNN system stands in the middle of different end-to-end systems, the gain of the hybrid system is mainly come from the sequence training criterion. It is worth to mention that the visual only system is supervised trained using audio frame-level alignments, which implies the potential to improve the performance of lipreading models using audio information.

**Table 1**. WERs of hybrid and end-to-end systems on visual only, audio-only and audio-visual speech recognition tasks.

| Models | V | A | A+V |
|---|---|---|---|
| TM-CTC [11] | 65.0 | 15.3 | 13.7 |
| TM-Seq2seq [11] | 49.8 | 10.5 | 9.4 |
| CTC/Attention [31] | 63.5 | 8.3 | 7.0 |
| CE TDNN | 55.02 | 10.17 | 8.95 |
| LF-MMI TDNN | **48.86** | **6.71** | **5.93** |

### 4.3. Results of pipelined AVSR

Table 2 shows the pipelined systems' word error rates (WERs) on two-speaker overlapped speech recognition task under four SNR conditions: -5dB, 0dB, 5dB and 10dB. Both the audio-only and audio-visual separation model in the pipelined system are trained using two-speaker overlapped speech simulated from LRS2 dataset. The first four rows in Table 2 shows the results of the pipelined system using clean speech trained ASR and AVSR back-end. A stronger pipelined system using ASR and AVSR trained on the mix of clean and enhanced (separated) data are also displayed in the last two rows. The separated data are obtained by feeding the overlapped training data into the audio-visual separation system. The results indicates that the use of visual information in front-end separation component, back-end recognition component or both of them can remarkably improve the performance of the pipelined system.

### 4.4. Results of integrated AVSR

In Table 3, the first two rows are the results of the ASR and AVSR systems trained on clean speech. The rest of the systems are trained on the mixture of clean and two-speaker overlapped speech. The system performance can be significantly improved up to **29.98%** absolute WER reduction by adding visual modality. We observed that the visual modality driven gated fusion (V⊗A) and audio-visual modality driven gated fusion (AV⊗A) methods significantly outperform the feature concatenation based method, which indicates efficacy of gating operation in overlap speech recognition. Compared with the pipelined systems results in Table 2, the best integrated system is slightly better than the best pipelined systems using multi-conditional trained AVSR.

**Table 2**. WERs on audio-only and audio-visual pipelined systems, 'mult' means using both clean and separated as multi-conditional training data

| Separation | | Recognition | | | WER | | | | |
|---|---|---|---|---|---|---|---|---|---|
| A | V | Data | A | V | 10dB | 5dB | 0dB | -5dB | AVE |
| ✓ | ✗ | clean | ✓ | ✗ | 21.26 | 29.35 | 40.79 | 55.43 | 36.71 |
| ✓ | ✗ | | ✓ | ✓ | 14.38 | 19.87 | 27.54 | 39.06 | 25.21 |
| ✓ | ✓ | | ✓ | ✗ | 12.74 | 14.94 | 21.52 | 32.73 | 20.48 |
| ✓ | ✓ | | ✓ | ✓ | 9.70 | 11.10 | 15.36 | 22.98 | 14.79 |
| ✓ | ✓ | mult | ✓ | ✗ | 10.43 | 14.94 | 15.88 | 21.81 | 15.06 |
| ✓ | ✓ | | ✓ | ✓ | 8.15 | 9.22 | 11.44 | 14.86 | **10.92** |

**Table 3**. WERs on integrated audio-visual overlapped speech recognition. 'concat' denotes feature concatenation fusion, "+concat" denotes concatenating the visual feature with the gated output. "mult*" denotes using clean and overlapped speech with different SNR level as training data

| Data | Modality | | Fusion | WER | | | | |
|---|---|---|---|---|---|---|---|---|
| | A | V | | 10dB | 5dB | 0dB | -5dB | AVE |
| clean | ✓ | ✗ | - | 19.01 | 37.94 | 66.78 | 84.17 | 51.98 |
| | ✓ | ✓ | concat | 10.05 | 20.81 | 40.33 | 63.87 | 33.77 |
| mult* | ✓ | ✗ | - | 11.37 | 21.67 | 52.45 | 75.68 | 40.29 |
| | ✓ | ✓ | concat | 8.69 | 11.29 | 16.25 | 24.58 | 15.20 |
| mult* | ✓ | ✓ | V⊗A | 9.05 | 10.47 | 13.60 | 17.17 | 12.57 |
| | ✓ | ✓ | +concat | 8.23 | 9.80 | 12.52 | 15.78 | 11.49 |
| | ✓ | ✓ | AV⊗A | 7.87 | 9.16 | 11.42 | 14.76 | 10.80 |
| | ✓ | ✓ | +concat | **7.55** | **8.77** | **10.71** | **14.22** | **10.31** |

## 5. CONCLUSION & FUTURE WORK

This study first investigates the performance of LF-MMI trained hybrid AVSR and end-to-end AVSR systems on LRS2 dataset, a new state-of-the-art result is established by our LF-MMI TDNN model, then proposes two gated fusion methods purposely for overlapped speech recognition and compares the proposed methods with traditional pipelined systems. Experiments show: 1) the hybrid AVSR system outperforms end-to-end systems on LRS2 dataset; 2) the effectiveness of the gated fusion method; 3) the integrated system have comparable result with more complex pipelined system. In the future, this work will be extended to: 1) true cocktail party environment with noise, interference speech and reverberation; 2) a multi-channel system; 3) challenging situations, such as both the visual and audio information are degraded. Comparison between integrated system and jointly/multi-task training systems will also be investigated in the future.

## 6. ACKNOWLEDGEMENTS

This research is supported by Hong Kong Research Grants Council General Research Fund No.14200218 and Shun Hing Institute of Advanced Engineering Project No.MMT-p1-19.

## 7. REFERENCES


[1] J Rouat, "Computational auditory scene analysis: Principles, algorithms, and applications (wang, d. and brown, gj, eds.; 2006)[book review]," *IEEE Transactions on Neural Networks*, vol. 19, no. 1, pp. 199–199, 2008.

[2] Max Wertheimer, "Laws of organization in perceptual forms.," 1938.

[3] Chao Weng, Dong Yu, Michael L Seltzer, and Jasha Droppo, "Deep neural networks for single-channel multi-talker speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 10, pp. 1670–1679, 2015.

[4] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[5] Zhuo Chen, Yi Luo, and Nima Mesgarani, "Deep attractor network for single-microphone speaker separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 246–250.

[6] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.

[7] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[8] Dimitris A Pados and George N Karystinos, "An iterative algorithm for the computation of the mvdr filter," *IEEE Transactions On signal processing*, vol. 49, no. 2, pp. 290–300, 2001.

[9] Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, Xiong Xiao, and Fil Alleva, "Recognizing overlapped speech in meetings: A multi-channel separation approach using neural networks," *arXiv preprint arXiv:1810.03655*, 2018.

[10] Harry McGurk and John MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746, 1976.

[11] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[12] Shiliang Zhang, Ming Lei, Bin Ma, and Lei Xie, "Robust audio-visual speech recognition using bimodal dfsmn with multi-condition training and dropout regularization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 6570–6574.

[13] Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel, "Deep multimodal learning for audio-visual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, pp. 2130–2134.

[14] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[15] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 9299–9306.

[16] Faheem Khan and Ben Milner, "Speaker separation using visually-derived binary masks," in *Auditory-Visual Speech Processing (AVSP) 2013*, 2013.

[17] 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," *arXiv preprint arXiv:1804.03619*, 2018.

[18] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "The conversation: Deep audio-visual speech enhancement," *arXiv preprint arXiv:1804.04121*, 2018.

[19] Rui Lu, Zhiyao Duan, and Changshui Zhang, "Listen and look: Audio–visual matching assisted speech source separation," *IEEE Signal Processing Letters*, vol. 25, no. 9, pp. 1315–1319, 2018.

[20] Rui Lu, Zhiyao Duan, and Changshui 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.

[21] Faheem Ullah Khan, Ben P Milner, and Thomas Le Cornu, "Using visual speech information in masking methods for audio speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1742–1754, 2018.

[22] Jian Wu, Yong Xu, Shi-Xiong Zhang, Lian-Wu Chen, Meng Yu, Lei Xie, and Dong Yu, "Time domain audio visual speech separation," *arXiv preprint arXiv:1904.03760*, 2019.

[23] Faheem Khan and Ben Milner, "Speaker separation using visual speech features and single-channel audio.," in *INTERSPEECH*, 2013, pp. 3264–3268.

[24] Guan-Lin Chao, William Chan, and Ian Lane, "Speaker-targeted audio-visual models for speech recognition in cocktail-party environments," *arXiv preprint arXiv:1906.05962*, 2019.

[25] Christoph Lüscher, Eugen Beck, Kazuki Irie, Markus Kitza, Wilfried Michel, Albert Zeyer, Ralf Schlüter, and Hermann Ney, "Rwth asr systems for librispeech: Hybrid vs attention," *Interspeech, Graz, Austria*, 2019.

[26] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[27] Shansong Liu, Shoukang Hu, Yi Wang, Jianwei Yu, Rongfeng Su, Xunying Liu, and Helen Meng, "Exploiting visual features using bayesian gated neural networks for disordered speech recognition," *Proc. Interspeech 2019*, pp. 4120–4124, 2019.

[28] Fei Tao and Carlos Busso, "Gating neural network for large vocabulary audiovisual speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 7, pp. 1286–1298, 2018.

[29] Jun Du, Qing Wang, Tian Gao, Yong Xu, Li-Rong Dai, and Chin-Hui Lee, "Robust speech recognition with speech enhanced deep neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[30] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," *arXiv preprint arXiv:1907.04975*, 2019.

[31] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic, "Audio-visual speech recognition with a hybrid ctc/attention architecture," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 513–520.

[32] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3444–3453.

[33] Zhehuai Chen, Jasha Droppo, Jinyu Li, Wayne Xiong, Zhehuai Chen, Jasha Droppo, Jinyu Li, and Wayne Xiong, "Progressive joint modeling in unsupervised single-channel overlapped speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 1, pp. 184–196, 2018.

[34] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks.," in *Interspeech*, 2018, pp. 3743–3747.

[35] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi.," in *Interspeech*, 2016, pp. 2751–2755.

[36] T. Afouras, J. S. Chung, and A. Zisserman, "Deep lip reading: a comparison of models and an online application," in *INTERSPEECH*, 2018.