






# Audio-Visual Multi-Channel Integration and Recognition of Overlapped Speech

Jianwei Yu , Shi-Xiong Zhang, *Member, IEEE*, Bo Wu, Shansong Liu ,  
Shoukang Hu , *Graduate Student Member, IEEE*, Mengzhe Geng, Xunying Liu , *Member, IEEE*,  
Helen Meng, *Fellow, IEEE*, and Dong Yu , *Fellow, IEEE*

**Abstract**—Automatic speech recognition (ASR) technologies have been significantly advanced in the past few decades. However, recognition of overlapped speech remains a highly challenging task to date. To this end, multi-channel microphone array data are widely used in current ASR systems. Motivated by the invariance of visual modality to acoustic signal corruption and the additional cues they provide to separate the target speaker from the interfering sound sources, this paper presents an audio-visual multi-channel based recognition system for overlapped speech. It benefits from a tight integration between a speech separation front-end and recognition back-end, both of which incorporate additional video input. A series of audio-visual multi-channel speech separation front-end components based on *TF masking*, *Filter&Sum* and *mask-based MVDR* neural channel integration approaches are developed. To reduce the error cost mismatch between the separation and the recognition components, the entire system is jointly fine-tuned using a multi-task criterion interpolation of the scale-invariant signal to noise ratio (Si-SNR) with either the connectionist temporal classification (CTC), or lattice-free maximum mutual information (LF-MMI) loss function. Experiments suggest that: the proposed audio-visual multi-channel recognition system outperforms the baseline audio-only multi-channel ASR system by up to 8.04% (31.68% relative) and 22.86% (58.51% relative) absolute WER reduction on overlapped speech constructed using either simulation or replaying of the LRS2 dataset respectively. Consistent performance improvements are also obtained using the proposed audio-visual multi-channel recognition system when using occluded video input with the lip region randomly covered up to 60%.

**Index Terms**—Overlapped speech recognition, speech separation, audio-visual, multi-channel, visual occlusion, jointly fine-tuning.

Manuscript received November 12, 2020; revised March 22, 2021; accepted May 6, 2021. Date of publication May 13, 2021; date of current version June 14, 2021. This work was supported in part by the Hong Kong Research Grants Council GRF under Grants 14200218 and 14200220, in part by Theme-based Research Scheme T45-407/19N, Innovation and Technology Fund under Grant ITS/254/19, and in part by the Shun Hing Institute of Advanced Engineering under Grant MMT-p1-19. (*Corresponding author: Prof. Xunying Liu.*)

Jianwei Yu is with Tencent AI Lab, Shenzhen 999077, China., and also with the Chinese University of Hong Kong, Hong Kong 999077 (e-mail: jwyu@se.cuhk.edu.hk).

Shansong Liu, Shoukang Hu, Mengzhe Geng, Xunying Liu, and Helen Meng are with the Department of System Engineering and Engineering Management, Chinese University of Hong Kong, Hong Kong 999077 (e-mail: sliu@se.cuhk.edu.hk; skhu@se.cuhk.edu.hk; mzgeng@se.cuhk.edu.hk; xylu@se.cuhk.edu.hk; hmmeng@se.cuhk.edu.hk).

Shi-Xiong Zhang, Bo Wu, and Dong Yu are with Tencent AI Lab, Bellevue, WA 98005 USA (e-mail: auszhang@tencent.com; lambowu@tencent.com; dyu@tencent.com).

Digital Object Identifier 10.1109/TASLP.2021.3078883

## I. INTRODUCTION

**D**ESPITE the rapid progress of automatic speech recognition (ASR) technologies in the past few decades, recognition of overlapped speech remains a highly challenging task. The presence of interfering speakers creates a large mismatch between the target speaker's clean speech and the mixed signal. This often leads to large performance degradation of current ASR systems. To this end, acoustic beamforming techniques integrating sensor data from multiple array channels are widely used. These multi-channel array signal integration approaches are normally implemented as time or frequency domain linear filters that are capable of “listening” in the target speaker's direction while minimizing the effects of noise distortions and other interfering speakers from other directions. The desired target speech signal is thus enhanced.

Microphone arrays play a key role in state-of-the-art ASR systems designed for multi-talker overlapped and far field speech [1]–[6], often following a traditional speech enhancement prior to recognition based system architecture. These systems contain two separately developed components: the speech separation and enhancement front-end module, and the speech recognition back-end. These two components are often integrated in a pipelined manner. The separation front-end module is often implemented using conventional beamforming techniques represented by either time domain delay and sum [7], [8] or frequency domain minimum variance distortionless response (MVDR) [9], [10] and the related generalized eigenvalue (GEV) [11] channel integration approaches. The former uses generalized phase correlation between sensor inputs and a Viterbi search procedure to estimate the optimal channel delays and their respective combination weights. The frequency domain beamforming approaches maximizes the signal to noise ratio (SNR) of the filtered outputs.

With the successful and wider application of deep learning based speech technologies, microphone array channel integration methods have evolved into a variety of neural network (NN) based designs in the past few years. These NN based methods can be classified into three main categories including *TF masking*, *Filter&Sum* and *mask-based MVDR* or *GEV*. In contrast to the traditional mask based single channel speech separation methods [12], [13], multi-channel information is fed into DNNs in the TF masking approaches [14], [15] to predict spectral time-frequency (TF) mask labels for a reference channel that

specify whether a particular TF spectrum point is dominated by the target speaker or interfering sources to facilitate speech separation. The neural *Filter&Sum* approaches directly estimate the beamforming filter parameters in either time domain [16]–[18] or frequency domain [19] to produce the separated outputs. The *mask-based MVDR* [4]–[6], [20]–[23] and related *mask-based GEV* [24], [25] approaches predict the TF masks using DNNs before estimating the power spectral density (PSD) matrices for the target and overlapping speakers to obtain the beamforming filter parameters. Compared with the conventional stand-alone beamforming approaches, these neural based methods allow a tighter integration with the downstream recognition back-end [5], [6], [19], [25], [26]. Large performance improvements have been reported for overlapped speech recognition tasks by using microphone array based multi-channel inputs [5], [6]. However, the current systems' performance gap between overlapped and non-overlapped speech remains large.

Inspired by the bi-modal nature of human speech perception, there has been increasing interest in incorporating visual information into the speech separation and recognition systems for far-field and overlapped speech. The advantages of these approaches are three folds: 1) visual information contains additional cues such as lip movements that differentiate the target speech from other interfering sources; 2) lip movements can provide further information over articulation to improve phonetic discrimination; 3) visual modality is usually invariant to the acoustic signal corruption in noisy or multi-talker environment. Previous research has successfully used visual information to improve single-channel overlapped speech separation [27]–[29] and recognition [30]–[32] performance. Recently there has also been increasing efforts in developing audio-visual multi-channel input based speech enhancement systems designed for speech separation [33] and de-reverberation [34]. However, currently there is a lack of holistic, full incorporation of visual information into both the front-end speech separation module and the back-end recognition component.

Performance of audio-visual overlapped speech separation and recognition systems crucially depends on the quality of the video input in terms of the complementary information being provided on top of the audio. Such sensitivity can be demonstrated by, for example, when the mouth area is obstructed by a mask (often required in the current pandemic), a microphone, or if the speaker stands far away from the camera (low-resolution video inputs). Only limited previous research was conducted to investigate the system fragility to the aforementioned video occlusion and low-resolution video input problems. In [35], the authors applied dropout to parts of the DNN acoustic model of a single channel audio-visual speech recognition system connected to the video input during model training to improve the resulting system's robustness. In [29], the comparable effect in audio-visual speech separation was investigated by using video with artificial occlusion over the mouth region. However, there has been very limited previous research investigating the performance sensitivity to video occlusion and low-resolution video inputs in the context of a complete audio-visual multi-channel recognition system of overlapped speech.

In order to address the above issues, an audio-visual multi-channel overlapped speech recognition system featuring tightly integrated separation front-end and recognition back-end is proposed in this paper. Firstly, for the speech separation front-end, a series of audio-visual microphone array channel integration methods including *TF masking*, *Filter&Sum* and *mask-based MVDR* are proposed respectively. Secondly, in order to reduce the error cost mismatch between the separation and the recognition components that are traditionally trained on different objective functions, they are jointly fine-tuned using a multi-task criterion interpolation of the scale-invariant signal to noise ratio (Si-SNR) with either the connectionist temporal classification (CTC) [36], or lattice-free maximum mutual information (LF-MMI) [37], [38] loss function. Thirdly, this paper investigates the influence of visual occlusion and low-resolution visual inputs on the proposed systems. To improve the robustness of audio-visual multi-channel speech recognition systems to visual occlusion, both angle features provided by video cameras mounted on a microphone array and multi-style training consisting of occluded video of lip region coverage up to 80% are used. In addition, the image in-painting technique [39] is also investigated to restore the occluded video inputs for the visual occlusion issue. Experiments suggest that: 1) the proposed audio-visual multi-channel recognition system outperforms the baseline audio-only multi-channel ASR systems by up to 8.04% (31.68% relative) and 22.86% (58.51% relative) absolute WER reduction on overlapped speech constructed using either simulation or replaying of the LRS2 dataset; 2) consistent performances improvements are obtained across all audio-visual multi-channel systems when multi-task criterion based joint fine-tuning is used in place of a pipelined configuration. In particular the jointly fine-tuned audio-visual multi-channel system using mask-based MVDR beamforming produced WER reductions by up to 4.2% (19.7% relative) and 5.1% absolute (25.4% relative) on the simulated and the replayed data over the pipelined system; 3) consistent performance improvements are also obtained using the proposed audio-visual multi-channel recognition system when even using occluded video input with the lip region randomly covered up to 60%.

In summary, this work makes three main contributions:

- This paper presents the first work on incorporating visual inputs in both the speech separation front-end and the recognition back-end within a bimodal and multi-channel inputs based overlapped speech recognition system. A systematic overview and comparison over three different audio-visual channel integration methods featuring a tight integration between the separation and the recognition components is given. In contrast, video information is added into only either the separation front-end [27], [28], [40], or the recognition back-end alone [31], [41], [42]. A more holistic use of video cues as investigated in this paper was not considered.
- This is the first work that uses an interpolated error cost that combines the lattice-free MMI based sequence discriminative training criterion and the scale-invariant signal to noise ratio (Si-SNR) metric to integrate the separation front-end

and recognition back-end. In contrast, the previous research focused on using cross entropy based error cost [5], [6], [43] in the overall end-to-end system fine-tuning and integration stage.

- This paper presents the first more complete attempt to investigate the effect from video input occlusion on both the separation and the recognition components as well as the final system performance on overlapped speech.

In contrast, the previous investigation on the effect from occluded video is limited to either the separation module [29] or recognition [35] component.

The rest of the paper is organized as follows. Section II introduces three neural network based multi-channel integration methods. Section III presents various forms of audio-visual multi-channel speech separation networks. Description of the recognition back-end components and their integration with the separation front-end are given in Section IV. Experimental results are presented in Section V. Section VI draws the conclusions and discusses possible future directions.

## II. MULTI-CHANNEL SPEECH SEPARATION

### A. Multi-Channel Signal Model for Overlapped Speech

Ignoring the reverberation in the overlapped speech, the spectrum of the received speech signal  $X_r(t, f)$  recorded by a far-field microphone array composed of  $R$  channels can be modeled as:

$$X_r(t, f) = Y_r(t, f) + N_r(t, f), \quad (1)$$

where  $X_r(t, f)$ ,  $Y_r(t, f)$  and  $N_r(t, f)$  denote the short-time Fourier transform (STFT) spectra of the overlapped, target and interfering speech received by the  $r$ th microphone respectively. Without loss of generality, we select the first channel as the reference channel ( $r = 1$ ) in this paper.

### B. TF Masking

To separate the target speaker from other interfering sources, the *TF masking* approaches have been widely used in monaural speech separation tasks in the past few decades [12], [26], [44], [45]. Such approaches predict spectral TF mask labels that specify whether a particular TF spectrum point is dominated by the target speaker or other interfering sources to facilitate speech separation. Recently, several researches have shown that integrating the multi-channel information collected by a microphone array can improve the mask estimation of the reference channel and lead to better speech separation. It has been found in previous research that the complex ratio masks (CRMs) outperform both the binary masks (BMs) and real-value ratio masks (RMs) on speech separation [26], [46] and enhancement [47] tasks. For this reason, the CRM based *TF masking* approach is implemented in this work. The complex spectrum of the separated output  $Y(t, f)$  is computed as follows:

$$\begin{aligned} Y(t, f) &= M(t, f)X_r(t, f) \\ &= \mathcal{R}\{M(t, f)\}\mathcal{R}\{X_r(t, f)\} - \mathcal{I}\{M(t, f)\}\mathcal{I}\{X_r(t, f)\} \end{aligned}$$

$$+ j(\mathcal{I}\{M(t, f)\}\mathcal{R}\{X_r(t, f)\} + \mathcal{R}\{M(t, f)\}\mathcal{I}\{X_r(t, f)\}) \quad (2)$$

where  $M(t, f) \in \mathbb{C}$  is the CRM of the target speaker and  $\mathcal{R}\{\cdot\}/\mathcal{I}\{\cdot\}$  denote the real/imaginary parts of a complex number respectively. Although the *TF masking* approach can provide perceptually enhanced sounds, it has been reported that the artifacts resulting from deterministic spectral masking introduced a negative impact on downstream speech recognition system performance [2], [4], [23].

### C. Multi-Channel Integration Using Beamforming

The acoustic beamforming approaches are designed to capture sound coming from the target speaker direction while reducing interfering sounds coming from other directions. This is realized by setting the beamformer filter parameters to the target direction. A linear filter

$$\mathbf{W}(f) = [W_1(f), W_2(f), \dots, W_R(f)]^T$$

is applied to the multi-channel overlapped speech spectrum vector

$$\mathbf{X}(t, f) = [X_1(t, f), X_2(t, f), \dots, X_R(t, f)]^T$$

as follows:

$$\begin{aligned} Y(t, f) &= \mathbf{W}(f)^H \mathbf{X}(t, f) \\ &= \underbrace{\mathbf{W}(f)^H \mathbf{Y}(t, f)}_{\text{speech}} + \underbrace{\mathbf{W}(f)^H \mathbf{N}(t, f)}_{\text{noise}}, \quad (3) \end{aligned}$$

where  $(\cdot)^H$  denotes the conjugate transpose. The beamforming filter parameters in conventional beamformers are usually obtained by first estimating the steering vector, which requires the direction-of-arrival (DOA) of the target speaker before solving an optimization problem, such as MVDR beamformer. With the successful and wider application of deep learning based speech technologies, state-of-the-art neural beamforming techniques are represented by the following two approaches: 1) using NNs to directly estimate beamforming filters as in *Filter&Sum* [17]–[19]; 2) using TF masks to estimate beamforming filters as in *mask-based MVDR* or *GEV* [4], [21], [23].

### D. Filter and Sum

The neural *Filter&Sum* beamforming approaches directly estimate the beamforming filter parameters in either time domain [16]–[18] or frequency domain [19] base on deep neural networks in a fully-trainable fashion. In this work, we adopt a frequency domain *Filter&Sum* approach to produce the separated output as follows:

$$Y(t, f) = \mathbf{W}(t, f)^H \mathbf{X}(t, f) = \sum_r W_r(t, f) * X_r(t, f). \quad (4)$$

One limitation associated with the *Filter&Sum* beamformer is that the estimated filter parameters are allowed to change over very short analysis intervals, for example, between neighbouring frame windows of 25 milliseconds. In practice this is an unrealistic assumption as the speech from a target speaker tends

to remain from the same direction over a longer period of time when collected using fixed microphone arrays, before he or she moves to a different position in the room.

### E. Mask-Based MVDR

When choosing the  $r$ th channel as the reference channel, the residual signal distortion  $\xi_{r,d}(t, f)$  and the residual noise  $\xi_n(t, f)$  can be computed by Equation (5) and Equation (6) respectively:

$$\begin{aligned}\xi_{r,d}(t, f) &= Y_r(t, f) - \mathbf{W}(f)^H \mathbf{Y}(t, f) \\ &= (\mathbf{U}_r - \mathbf{W}(f))^H \mathbf{Y}(t, f),\end{aligned}\quad (5)$$

$$\xi_n(t, f) = \mathbf{W}(f)^H \mathbf{N}(t, f) \quad (6)$$

where  $\mathbf{U}_r = [0, 0, \dots, 1, \dots, 0]^T$  is a one-hot vector of which the  $r$ th entry equals to 1. The MVDR beamformer is designed to minimize the noise output while imposing a distortionless constraint on the target speech signal [10]:

$$\begin{aligned}\min_{\mathbf{w}(f)} E_t \{ |\xi_n(t, f)|^2 \} \\ \text{subject to: } E_t \{ |\xi_{r,d}(t, f)|^2 \} = 0\end{aligned}$$

The distortionless constraint in the above optimization problem is equivalent to  $\mathbf{W}(f)^H \mathbf{G}(f) = 1$ , which can be interpreted as maintaining the energy along the target direction. It can be shown that the solution of the above MVDR beamformer is:

$$\mathbf{W}(f) = \frac{\Phi_n(f)^{-1} \mathbf{G}(f)}{\mathbf{G}(f)^H \Phi_n(f)^{-1} \mathbf{G}(f)} \quad (7)$$

$$= \frac{\Phi_n(f)^{-1} \Phi_y(f)}{\text{Trace}(\Phi_n(f)^{-1} \Phi_y(f))} \mathbf{U}_r, \quad (8)$$

where  $\Phi_n(f) = E_t \{ \mathbf{N}(t, f) \mathbf{N}(t, f)^H \}$  and  $\Phi_y(f) = E_t \{ \mathbf{Y}(t, f) \mathbf{Y}(t, f)^H \}$  are the PSD matrices of the noise and target speech respectively. The MVDR filter parameter estimation in Equation (7) is expressed in terms of the noise PSD matrices and the steering vector. Alternatively it can also be re-expressed using both the target speech and noise PSD matrices as in Equation (8).

In *mask-based MVDR* approaches, the deep neural networks are used to estimate the real-value [4], [5], [23] or complex [26] TF masks of the target speech  $M^y(t, f)$  and other interfering sources  $M^n(t, f)$  respectively. The PSD matrices corresponding to each source can be calculated with the estimated TF masks shown as follows:

$$\begin{aligned}\Phi_y(f) &= \frac{\sum_{t=1}^T (M^y(t, f) * \mathbf{X}(t, f)) (M^y(t, f) * \mathbf{X}(t, f))^H}{\sum_{t=1}^T M^y(t, f) * (M^y(t, f))^H}, \\ \Phi_n(f) &= \frac{\sum_{t=1}^T (M^n(t, f) * \mathbf{X}(t, f)) (M^n(t, f) * \mathbf{X}(t, f))^H}{\sum_{t=1}^T M^n(t, f) * (M^n(t, f))^H}.\end{aligned}\quad (9)$$

The MVDR beamformer filters can then be obtained using Equation (8). Compared with both the *TF masking* and the *Filter&Sum* approaches, *mask-based MVDR* beamformers using the spatial temporal correlation in the PSD matrices to enforce

a consistent set of filter parameters to be estimated over the analysis window, in which the location of the speakers are unchanged. Hence, the processing artifacts observed in the former two approaches can be minimized. This is particularly useful when modelling the short speech segments within which the target speaker voice is recorded from the same direction using the array. Compared with both the *TF masking* and the *Filter&Sum* approach, the *mask-based MVDR* approach retains a consistent DOA estimation with a beamforming analysis window over, for example, an utterance of speech and the minimum distortion constraint in traditional MVDR beamforming. The *mask-based MVDR* approach has demonstrated state-of-the-art performance in noisy and overlapped speech recognition [4], [5], [23].

## III. AUDIO-VISUAL MULTI-CHANNEL SPEECH SEPARATION

In this section, we introduce the proposed audio-visual multi-channel speech separation networks using *TF masking*, *Filter&Sum* and *mask-based MVDR* channel integration methods.

### A. Audio Modality

In the proposed separation front-ends, three types of audio features including the complex spectrum, the inter-microphone phase differences (IPDs) [4] and the location-guided angle feature (AF) [22], [40] are adopted as the audio inputs. The detailed paradigm of the audio modality processing is illustrated in the top left corner of Fig. 1. The complex spectrum of all the microphone array channels are first computed through the STFT. Following the same recipe as in [33], the IPD feature is calculated as follows:

$$\text{IPD}^{(i,j)}(t, f) = \angle \left( \frac{X_i(t, f)}{X_j(t, f)} \right), \quad (10)$$

where  $X_i(t, f)$  represents the  $i$ -th channel's complex spectrum of the mixed signal at time frame  $t$  and frequency bin  $f$ ,  $(i, j)$  corresponding to the selected microphone pair and  $\angle(\cdot)$  outputs the angle of the input argument. The IPD feature captures the relative phase difference between microphones, which is correlated with the time difference of arrival (TDOA).

In addition, when the geometry of the microphone array and the direction of arrival (DOA) of the target speaker  $\theta$  are given, the steering vector corresponding to the target speaker can be computed as:

$$\mathbf{G}(f) = [e^{-j \frac{2\pi f d_{11}}{c} \cos(\theta)}, e^{-j \frac{2\pi f d_{1r}}{c} \cos(\theta)}, \dots, e^{-j \frac{2\pi f d_{1R}}{c} \cos(\theta)}] \quad (11)$$

where  $d_{1r}$  is the distance between the first (reference) and  $r$ th microphone ( $d_{11} = 0$ ).  $c$  is the sound velocity.

Based on the computed steering vector, the location-guided AF feature introduced in [22], [33] are also adopted to provide discriminative information for the target speaker as follows:

$$\text{AF}(t, f) = \sum_{\{(i,j)\}} \frac{\langle \text{vec}(\frac{G_j(f)}{G_i(f)}), \text{vec}(\frac{X_i(t,f)}{X_j(t,f)}) \rangle}{\| \text{vec}(\frac{G_j(f)}{G_i(f)}) \| \cdot \| \text{vec}(\frac{X_i(t,f)}{X_j(t,f)}) \|} \quad (12)$$

where  $\| \cdot \|$  denotes the vector norm,  $\langle \cdot, \cdot \rangle$  represents the inner product and  $\{(i, j)\}$  denotes the selected microphone pairs.

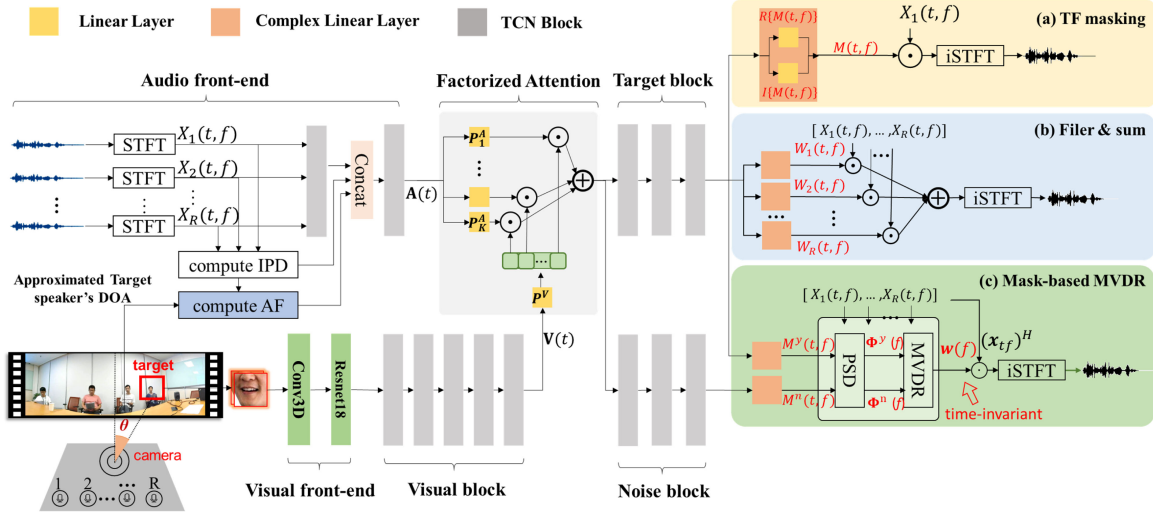


Fig. 1. Illustration of the proposed audio-visual multi-channel speech separation networks, where  $X_r(t, f)$  is the complex spectrum of each channel.  $\mathbf{V}(t)$  and  $\mathbf{A}(t)$  denote the audio and the visual embedding at frame index  $t$  respectively. The detailed paradigm of the TCN block is demonstrated in Fig. 2. (a), (b) and (c) represent three options of channel integration approaches: (a) TF masking:  $M(t, f)$  represents the complex mask of the target speaker, where  $R\{M(t, f)\}$  and  $I\{M(t, f)\}$  are the real and the imaginary part of the mask respectively; (b) Filter&Sum:  $W_r(t, f)$  denotes the beamforming filter parameters of the  $r$ th channel; (c) Mask-based MVDR:  $M^y(t, f)$  and  $M^n(t, f)$  are the complex masks of the target and the interfering sources,  $\Phi^y(f)$  and  $\Phi^n(f)$  are the corresponding PSD matrices and  $\mathbf{W}(f)$  is the time-invariant beamforming filter parameters.

$\text{vec}(\cdot)$  transforms the complex value into a 2-D vector, where the real and imaginary parts are regarded as the two vector components. The design principle of the AF is that if the TF bin  $X_i(t, f)$  is dominated by the target speaker from direction  $\theta$ , then its corresponding  $\text{AF}(t, f)$  will be close to 1, otherwise close to 0. In this work, the DOA of the target speaker can be estimated by tracking the speaker's face from a 180-degree wide-angle camera as shown in Fig. 1 (bottom left corner).

Motivated by the success of Conv-Tasnet [45] in speech separation, the temporal convolutional network (TCN) architecture, which uses a long reception field to capture more sufficient contextual information, is adopted in our separation front-ends. As shown in Fig. 2, each TCN block is stacked by 8 Dilated 1-D ConvBlock with exponentially increased dilation factors  $2^0, 2^1, \dots, 2^7$ . As shown in the Audio front-end (Fig. 1, top left corner), the complex spectrum of each microphone array channel are first concatenated and then fed into a TCN block. The outputs are concatenated with the IPD and AF features and then fed into another TCN block to compute the audio embeddings  $\mathbf{A} \in \mathbb{R}^{T \times D}$ .

## B. Visual Modality

For the visual modality, as shown in the bottom left corner of Fig. 1, the lip region of the target speaker obtained by face tracking is fed into the Visual front-end (Fig. 1, bottom left corner in green) followed by the Visual block (Fig. 1, bottom middle in gray) to compute the visual embeddings  $\mathbf{V} \in \mathbb{R}^{T \times D}$ . The network structure of the Visual front-end is similar to the one proposed in [50], which consists of a spatio-temporal convolution layer (Conv3D) and a 18-layer ResNet [51] to capture the spatio-temporal dynamics of the lip movements. The Visual block consists of 5 TCN blocks. Following the work in [31], [52],

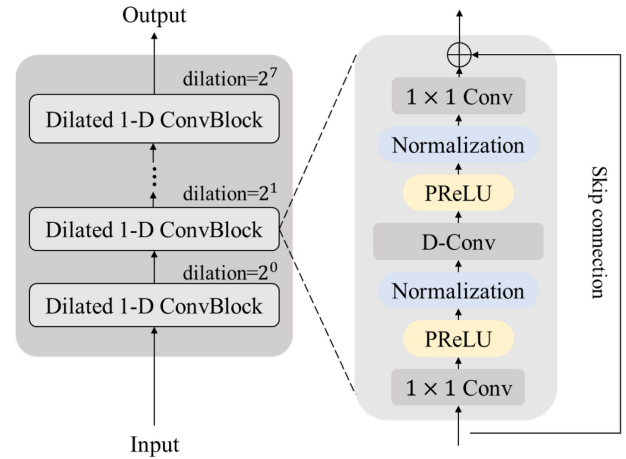


Fig. 2. Illustration of the architecture of the temporal convolutional network (TCN) Block. Each dilated 1-D ConvBlock consists of a  $1 \times 1$  convolutional layer, a depth-wise separable convolution layer (D-Conv) [48], with PReLU [49] activation function and normalization added between each two convolution layers. Skip connection is added in each dilated 1-D ConvBlock.

[53], the Visual front-end is pretrained on the lipreading task as described in [50]. The visual modality can provide not only discriminative information to facilitate phone classification, but also crucially additional cues to track and separate the target speaker from interfering sources of sound.

## C. Modality Fusion

In order to effectively integrate the audio and visual modalities, a careful design of the modality fusion scheme is required. Based on the investigation of different modality fusion methods in our previous work [40], a factorized attention-based modality

fusion method, which has been proven to outperform the most commonly used feature concatenation method [27]–[29], [54] in the audio-visual speech separation front-ends, is adopted in this work.

As shown in Fig. 1 (middle up, in light gray), the acoustic embedding  $\mathbf{A}(t)$  at frame index  $t$  is first factorized into  $K$  acoustic subspace vectors by a series of parallel linear transformations  $\mathbf{P}_k^A \in \mathbb{R}^{D \times D}$  and the visual embedding  $\mathbf{V}(t)$  is mapped into a  $K$  dimensional vector  $\mathbf{v}(t) = [v_1(t), \dots, v_K(t)]$  by projection matrix  $\mathbf{P}^V \in \mathbb{R}^{D \times K}$  in the factorized attention method as follows:

$$[\mathbf{a}_1(t), \dots, \mathbf{a}_K(t)] = [\mathbf{P}_1^A, \dots, \mathbf{P}_K^A] \mathbf{A}(t) \quad (13)$$

$$\mathbf{v}(t) = \text{Softmax}(\mathbf{P}^V \mathbf{V}(t)). \quad (14)$$

Then the fused audio-visual embedding  $\mathbf{AV}(t) \in \mathbb{R}^D$  is obtained by using the weighted sum of the acoustic subspace vectors:

$$\mathbf{AV}(t) = \sigma \left( \sum_{k=1}^K v_k(t) \mathbf{a}_k(t) \right) \quad (15)$$

where  $\sigma(\cdot)$  is the sigmoid function.

#### D. Channel Integration

As discussed in Section II, three different audio-visual multi-channel integration approaches are investigated in this work.

a) *TF masking*: The diagram of the *TF masking* approach is illustrated in Fig. 1 (top right, in light yellow). The hidden outputs of the Target block (Fig. 1, middle up in gray) are fed into a complex linear layer to estimate the complex mask of the reference channel. The structure of the complex linear layer is shown in Fig. 1 (top right in orange), which consists of two linear layers. One is used to estimate the real part  $\mathcal{R}\{M(t, f)\}$  of the complex mask, the other is used to estimate the imaginary part  $\mathcal{I}\{M(t, f)\}$ . Based on the estimated TF mask, the output complex spectrum is then computed via 2.

b) *Filter&Sum*: The diagram of the *Filter&Sum* approach is shown in Fig. 1 (right middle, in light blue). Different from the *TF masking* approach, the hidden outputs of the Target block are fed into a series of complex linear layers to estimate the time variant beamforming filter parameters  $W_r(t, f)$  corresponding to each channel frame by frame. The frequency domain beamforming outputs are then computed using 4.

c) *Mask-based MVDR*: The *mask-based MVDR* approach is demonstrated Fig. 1 (right bottom, in green). Different from the *TF masking* and the *Filter&Sum* approaches, an additional Noise block (Fig. 1, middle bottom in gray) containing 3 TCN blocks and a complex linear layer is adopted to estimate the complex TF mask  $M^n(t, f)$  for the noise signals. As discussed in [4], estimating the TF masks for both the target and noise signals can improve the speech separation performance of the *mask-based MVDR* approach. With the TF masks of the target and interference speech, the beamforming filter parameters are calculated using Equation (8) and (9) described in Section II-E. In this work, we assume that the location of the speakers are

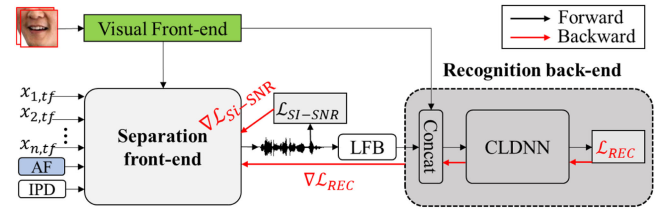


Fig. 3. Joint fine-tuning:  $\nabla \mathcal{L}_{REC}$  and  $\nabla \mathcal{L}_{Si-SNR}$  represent the gradients of speech recognition i.e. CTC, LF-MMI and speech separation Si-SNR objective functions respectively, “LFB” denotes log filter bank acoustic features.

unchanged within each utterance, which is common in meeting and restaurant environment. Therefore, in the *mask-based MVDR* approach, the beamforming filter parameters  $\mathbf{W}(f)$  are fixed with a beamforming analysis window, for example, an utterance of speech in this work.

In all the three channel integration methods, the target speech complex spectrum extracted by each channel integration method is used to compute the target speech waveform using the inverse STFT (iSTFT) operation.

#### IV. AUDIO-VISUAL MULTI-CHANNEL SPEECH RECOGNITION

In this section, we first introduce our audio-visual speech recognition back-ends and then describe the approaches to integrate the separation and the recognition components.

##### A. Audio-Visual Speech Recognition Back-End

Extensive audio-visual speech recognition technologies have been conducted in recent years and demonstrated their efficacy in improving speech recognition performance under both clean and adverse conditions [32], [35], [41], [42], [55]–[58]. Following [52], in this work, the convolutional long short-term memory fully connected neural network (CLDNN) [59] is adopted as the recognition back-end system architecture. As shown in Fig. 3 (left, in dark gray), the log filter bank features are first calculated from the separated target speech waveform before being concatenated with the visual features extracted using the Visual front-end. The concatenated features are fed into the CLDNN network to estimate the frame level posteriors. To optimize the model parameters in the recognition back-end, two widely used training criteria i.e. CTC [36] and LF-MMI [37], [60] are investigated in this work:

1) **CTC**: The CTC approach uses a blank symbol, which can appear between the modelling units (graphemes, phonemes), to define an objective function that sums over all possible alignments of the reference transcription with the input sequence of speech frames:

$$\mathcal{L}_{CTC} = \sum_{u=1}^U \log \sum_{\pi^u, \pi_t^u \in \{\Omega \cup \epsilon\}} \prod_{t=1}^T P(\pi_t^u | \mathbf{O}^u) \quad (16)$$

where  $\mathbf{O}^u = [O_1^u, \dots, O_T^u]$  represents the input utterance of  $T$  frames and  $\Omega$  denotes the grapheme (phoneme) symbol set.  $\pi^u = [\pi_1^u, \dots, \pi_T^u]$  represents a possible alignment between  $\mathbf{O}^u$  against the CTC output token  $\pi_t^u$ , which are based on either

a grapheme (phoneme) symbol, or a special null emission “ $\epsilon$ ” token, as considered in this paper.

2) **LF-MMI**: Sequence discriminative training techniques, represented by lattice-free MMI [37], have defined state-of-the-art hybrid ASR system performance in the past few years. The MMI criterion is a discriminative objective function which aims to maximize the probability of the reference transcription while minimizing the probability of all other transcriptions:

$$\mathcal{L}_{MMI} = \sum_{u=1}^U \log \frac{P(\mathbf{O}|\mathbf{H}^u)P(\mathbf{H}^u)}{\sum_{\tilde{\mathbf{H}}^u} P(\mathbf{O}|\tilde{\mathbf{H}}^u)P(\tilde{\mathbf{H}}^u)}$$

where  $\tilde{\mathbf{H}}^u$  represents any possible transcriptions. In recent research [60], [61], it has been shown that the end-to-end LF-MMI approach can outperform CTC based approach using either phoneme or grapheme modelling units on clean speech.

### B. Integration of the Separation and Recognition Components

Traditionally, the speech separation and recognition components are developed separately and then used in a pipelined fashion [4], [21]–[23]. However, two issues arise with such approach: 1) the cost function mismatch between separation and recognition components cannot guarantee the separated outputs target to optimal recognition performance; 2) the artifacts created by separation can increase modeling confusion of the recognition component and lead to performance degradation.

According to [25], [26], [43], [62], tight integration of the two components with joint fine-tuning can address above two issues. In this work, we investigated three variants of fine-tuning methods: 1) fine-tuning the recognition system only on the enhanced signals; 2) jointly fine-tuning the separation and the recognition components using the recognition cost function; 3) jointly fine-tuning both systems using a multi-task criterion, which interpolates the recognition and Si-SNR cost functions:

$$\mathcal{L} = \mathcal{L}_{REC} + \alpha \mathcal{L}_{Si-SNR}, \quad (17)$$

where  $\alpha$  is a manually tuned weight of the Si-SNR loss and  $\mathcal{L}_{REC}$  can be either  $\mathcal{L}_{CTC}$  or  $\mathcal{L}_{LF-MMI}$  cost function. As shown in Fig. 2, the gradient of the recognition cost is propagated into the separation front-end to update the model parameters of the entire system.

## V. EXPERIMENT SETUP

In this section, we first introduce the details of the corpus adopted in this work. Second, we describe the details of generation process of the multi-channel overlapped speech by either simulation or replay. Third, we explain how we introduce visual occlusion into the video. Finally, we introduce the implementation details of the proposed systems.

### A. LRS2 Corpus

The Oxford-BBC Lip Reading Sentences 2 (LRS2) corpus [63], which is one of the largest publicly available corpora for audio-visual speech recognition, is adopted in our experiments. This corpus consists of news and talk shows from BBC program, which is a challenging task since it contains thousands

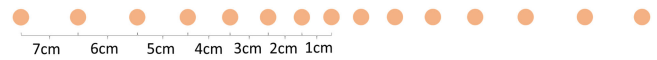


Fig. 4. The architecture of the microphone array used in the simulation and replay data recording.

of speakers with large variation in head pose. The LRS2 corpus is divided into three subsets, i.e. Pre-train, Train-val and Test set. In our experiments, the Pre-train and Train-Val subsets are combined for system training. More statistic details of the LRS2 corpus can be found in [41].

### B. Simulated Overlapped Speech

Since there is no publicly available audio-visual multi-channel overlapped speech corpus, we simulated the multi-channel overlapped speech in our experiments based on the LRS2 corpus. Details of the simulation process is described in Algorithm 1. A 15-channel symmetric linear array with non-even inter-channel spacing is used in the simulation process, as shown in Fig. 4. Reverberation is also added in the simulated data by convolving the single channel signals with the Room Impulse Responses (RIRs) generated by the image-source method [64]. The room size is randomly selected ranging from  $4 \times 4 \times 2.5$  m<sup>3</sup> to  $10 \times 8 \times 6$  m<sup>3</sup> (length  $\times$  width  $\times$  height) and the reverberation time T60 is sampled from a range of 0.05 to 0.7s. The average overlapping ratio of the simulated utterances is around 85% and SIR is around 0dB. The simulated data is divided into three subsets with 14.2k, 4.6k and 1.2k utterances respectively for training (200h), validation (2h) and evaluation (0.5h).

---

#### Algorithm 1: Data simulation process of multi-channel overlapped speech

---

**Input:** single channel non-overlapped LRS2 speech for utterance in LRS2 **do**

- 1) Randomly select an interfering utterance from another speaker in LRS2 corpus
- 2) Sample a SIR uniformly from (-6,0,6) dB
- 3) Randomly generate microphone array and speakers' position (Distance between speakers and array is 1-5m)
- 4) Scale the target and interfering sources with the sampled SIR
- 5) Generate mixed speech per channel with overlapping ratio randomly from 60% to 100%

**end for**

**Output:** multi-channel overlapped speech

---

### C. Replayed Overlapped Speech

To further evaluate the performance of the proposed systems in realistic environment, a replayed test set with 1.2k (0.5h) utterances recorded in a  $10 \times 5 \times 3$  m<sup>3</sup> meeting room is also used in our experiments. As shown in Fig. 5, two loudspeakers are used to replay different utterances of the LRS2 test set simultaneously to generate overlapped speech. The structure of the microphone array used during recording is the same

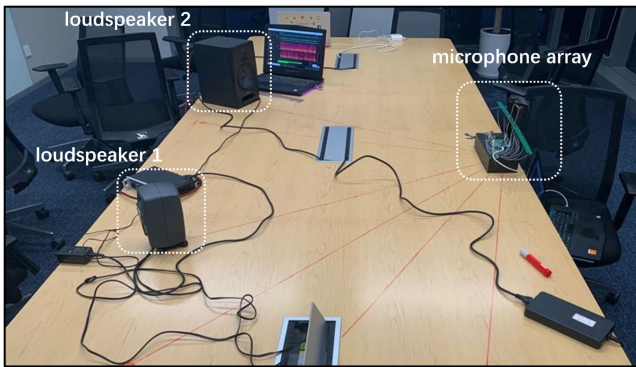


Fig. 5. Replayed recording of overlapped LRS2 test set.

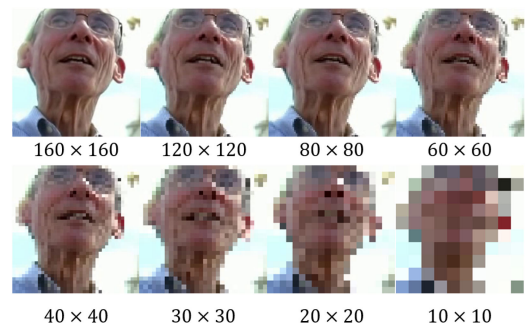
as that used in simulation. The target and interfering speakers are located at following directions related to the mounted camera, i.e.  $(15^\circ, 30^\circ)$ ,  $(45^\circ, 30^\circ)$ ,  $(75^\circ, 30^\circ)$ ,  $(105^\circ, 30^\circ)$ ,  $(30^\circ, 60^\circ)$ ,  $(90^\circ, 60^\circ)$ ,  $(120^\circ, 60^\circ)$  and  $(150^\circ, 60^\circ)$ , where the distance between the loudspeakers and microphones ranges from 1m to 1.5m. In the replayed data, the approximated DOA of the target speaker is obtained by the  $180^\circ$  camera mounted on top of the microphone array. The average overlapping ratio of the replayed overlapped speech is around 80% and SIR is around 1.5dB.

#### D. Visual Occlusion and Low-Resolution

As discussed in Section I, the performance of audio-visual speech separation and recognition systems crucially depends on the quality of the video input fed into these systems. In the experiments of this paper, an ablation study is conducted to assess the impact on system performance due to two forms of video input quality degradation often found in real world applications, before a set of techniques designed to improve robustness against such degradation are evaluated and their performance analysed. First, low-resolution visual inputs were generated by gradually reducing the original video resolution from  $160 \times 160$  pixels to  $120 \times 120$ ,  $80 \times 80$ ,  $60 \times 60$ ,  $40 \times 40$ ,  $30 \times 30$ ,  $20 \times 20$ , and eventually down to  $10 \times 10$  pixels, as shown in Fig. 6(a). Second, the video frames were occluded by applying randomly sized and positioned square patches to the lip region of each speaker. The size of the occluded lip regions randomly varies from  $45 \times 45$  to  $60 \times 60$  pixels. For each utterance, occlusion is applied to a randomly sampled window of consecutive video frames. The ratio between uncovered and occluded frames is randomly sampled from the following settings: 20%, 40%, 60% and 80%. As such video occlusion is applied to a contiguous area of the lip region, it is regarded as a more naturalistic form of video quality degradation in comparison with the frame level drop-out approach [35], [40].

#### E. Implementation Details

**Features:** 1) In the separation front-ends (Fig. 1, top left corner), the 257-dimensional complex spectrum of each channel is extracted using a 512-point FFT with 32ms hanning window and 16ms frame rate. In our implementation, the STFT operation is implemented as a convolution layer to enable on-the-fly



(a) Low-resolution visual inputs



(b) Visual occlusion

Fig. 6. Examples of (a) low-resolution visual input by gradually reducing the original video resolution from  $160 \times 160$  pixels to  $10 \times 10$  pixels; (b) occluded visual input with randomly sized ( $45 \times 45$  to  $60 \times 60$  pixels) and positioned square patches applied to the lip region.

computation. The AF and IPD features are computed using 9 microphone pairs (1,15), (2, 14), (3, 13), (1, 7), (12, 4), (11, 5), (12, 8), (7, 10) and (8, 9). These microphone pairs are selected to sample different spacing between microphones following [15], [33]. 2) The 40-dimensional log filter bank features extracted using a 40ms window and 10ms frame rate are adopted as the input feature of the recognition back-end. Similar to the STFT operation, the log filter bank extractor is also implemented as a layer in the network to enable on-line extraction. 3) For the visual front-end, the original  $160 \times 160$  video frames in LRS2 are centrally cropped by a  $112 \times 112$  window and then up-sampled to align with the audio frames via linear interpolation.

**Separation front-end:** In the separation front-ends (Fig. 1, middle, in gray), for each TCN block (Fig. 2), the number of channels in the  $1 \times 1$  Conv layer is set to be 256 for every Dilated 1-D ConvBlock. As for the D-Conv layer, the kernel size is set to be 3 with 512 channels. The Visual front-end (Fig. 1, bottom left corner in green) uses the same hyper-parameter settings as described in [50]. Following [33], the number of the acoustic subspace  $K$  is set to be 10 with  $\mathbf{P}^V \in \mathbb{R}^{256 \times 10}$  and  $\mathbf{P}_k^A \in \mathbb{R}^{256 \times 256}$  in the factorized attention layer. The output dimension of the complex linear layer is set to be 257.

**Recognition back-end:** In our experiments, the CTC and LF-MMI based recognition back-ends use the same neural network structure, which consists of four 2-dimensional convolutional layers with channel sizes (64, 64, 128, 128) and kernel size  $3 \times 3$  followed by four 1280 hidden units BLSTM layers and a softmax layer. Context-free grapheme units are used as the output layer targets in both the CTC and LF-MMI based models. The end-to-end LF-MMI criterion is implemented following the recipe<sup>1</sup> in [38]. The language model (LM) used in recognition is

<sup>1</sup><https://github.com/pytorch/examples/>



TABLE I

PERFORMANCE OF SINGLE CHANNEL ASR AND AVSR SYSTEMS ON ECHO FREE AND REVERBERANT SPEECH WITH OR WITHOUT OVERLAPPING. “SIMU” AND “REPLAY” DENOTES THE SIMULATED AND THE REPLAYED TEST DATA. † DENOTES A STATISTICALLY SIGNIFICANT IMPROVEMENT IS OBTAINED OVER THE CORRESPONDING ASR BASELINE

| Sys | Data                          | Criterion              | +visual | WER (%) |        |
|-----|-------------------------------|------------------------|---------|---------|--------|
|     |                               |                        |         | simu    | replay |
| 1   | Echo free<br>non-overlapped   | $\mathcal{L}_{CTC}$    | ✗       | 11.04   |        |
| 2   |                               |                        | ✓       | 9.77†   |        |
| 3   |                               | $\mathcal{L}_{LF-MMI}$ | ✗       | 9.44    |        |
| 4   |                               |                        | ✓       | 8.55†   |        |
| 5   | Reverberant<br>non-overlapped | $\mathcal{L}_{CTC}$    | ✗       | 15.33   |        |
| 6   |                               |                        | ✓       | 13.93†  |        |
| 7   |                               | $\mathcal{L}_{LF-MMI}$ | ✗       | 14.36   |        |
| 8   |                               |                        | ✓       | 11.61†  |        |
| 9   | raw channel 1<br>overlapped   | $\mathcal{L}_{CTC}$    | ✗       | 75.34   | 80.55  |
| 10  |                               |                        | ✓       | 32.06†  | 31.93† |
| 11  |                               | $\mathcal{L}_{LF-MMI}$ | ✗       | 65.44   | 71.03  |
| 12  |                               |                        | ✓       | 28.92†  | 28.89† |

a 4-gram LM constructed on 2.33M words of the LRS2 Train-val and Pre-train data transcripts.

All of our models are trained using 4 NVIDIA Tesla P40 GPU cards. For all results presented in this paper, matched pairs sentence-segment word error (MAPSSWE) based statistical significance test was performed at a significance level  $\alpha=0.05$ .

## VI. EXPERIMENTAL RESULTS

In this section, we describe the experiment results. First, to investigate the effectiveness of visual features extracted from the video frames, we compare the audio-only and audio-visual speech recognition systems without explicit speech separation components on non-overlapped and overlapped speech. Second, to tightly integrate the separation front-end and recognition back-end, we investigate the performance of three different integration methods in the proposed systems. We use the original LRS2 utterances as the echo free non-overlapped speech. The reverberant non-overlapped speech is simulated from the original LRS2 data using image-source method. Third, we systematically investigate the impact of the visual features on the proposed system to confirm the strength and importance of the visual information. Finally, we investigate the impact of visual occlusion on the proposed systems.

### A. Speech Recognition Without Separation Front-End

Table I presents the WER results of the CTC and LF-MMI based ASR and AVSR systems without using microphone array and explicit speech separation components on echo free and reverberant speech with or without speech overlapping.

Several trends can be observed from Table I:

- 1) For both the CTC and LF-MMI based systems, using visual information can significantly improve the recognition performance over the audio-only systems by up to 1.27% (sys.1 vs. sys.2) and 2.75% (sys.7 vs. sys.8) absolute WER reduction on echo free and reverberant non-overlapped speech. Especially, the audio-visual recognition system largely outperforms the audio-only system by

TABLE II

SI-SNR RESULTS OF *TF-MASKING*, *FILTER&SUM* AND *MASK-BASED MVDR* SEPARATION FRONT-ENDS

| Sys | AF | +visual | TF masing | Filter&Sum | MVDR |
|-----|----|---------|-----------|------------|------|
| 1   | ✓  | ✗       | 9.40      | 10.87      | 8.73 |
| 2   | ✗  | ✓       | 9.77      | 11.02      | 8.84 |
| 3   | ✓  | ✓       | 10.16     | 11.60      | 9.03 |

up to 33.28% and 48.62% (sys.9 vs. sys.10) absolute WER reduction on simulated and replayed overlapped speech respectively, which proves the effectiveness of the extracted visual features on overlapped speech recognition.

- 2) In our experiments, both the reverberation and the interfering speech are introduced into the simulated and the replayed multi-channel overlapped speech. Compared with the large performance degradation over 50% absolute WER increase caused by speech overlapping (sys.5 vs. sys.9, sys.7 vs. sys.11), the reverberation only introduces around 4% absolute WER degradation against the echo free speech (sys.1 vs. sys.5, sys.3 vs. sys.7). This indicates that overlapping speech (sys.9-12) is the more dominant contributing factor leading to large performance degradation against clean speech based recognition systems (sys.1-4) than reverberation (sys.5-8) on the LRS2 data considered in this paper.
- 3) The LF-MMI based systems outperform the CTC based systems on both non-overlapped and overlapped speech with and without visual modality in our experiments.

Based on the second observation, we focus on solving the speech overlapping issue in this work. Since we are not aiming at dereverberation in our overlapped speech recognition systems, the WER results on the reverberant non-overlapped speech (sys.5-8) can be defined as the upper bound for all subsequent experiments.

### B. Performance of Audio-Visual Speech Separation Front-Ends

The Si-SNR results of the *TF-masking*, *filter&sum* and *mask-based MVDR* separation front-ends are shown in Table II. Several trends can be observed in Table II:

- 1) Using visual features in the separation front-ends can improve the the Si-SNR performance (sys.1 vs. sys.3).
- 2) Separation front-ends using only visual features have comparable results with separation front-ends using only AFs.
- 3) Compared with the *TF-masking* and *filter&sum* separation front-ends, the *mask-based MVDR* separation front-ends show relatively lower Si-SNR. One possible explanation is that the *mask-based MVDR* benefits from the distortionless constraint which is not adopted in the other two approaches.

Fig. 7 shows example spectra of target clean, overlapped, audio-only separated, and audio-visual separated speech segments obtained using the *TF masking* based speech separation front-end. The spectrum portions circled using yellow dotted lines in (c) and (d) represent the interfering speaker’s speech,

TABLE III  
PERFORMANCE OF DIFFERENT FINE-TUNING METHODS CONDUCTED ON AUDIO-VISUAL MULTI-CHANNEL SPEECH RECOGNITION SYSTEMS. † AND ‡ DENOTES A STATISTICALLY SIGNIFICANT IMPROVEMENT IS OBTAINED OVER THE PIPELINED CTC (SYS.2) AND LF-MMI (SYS.6) SYSTEMS

| Sys | Fine-tuning  |      |       | TF masking  |                         |                         | Filter&Sum  |                         |                         | MVDR        |                         |                         |
|-----|--|------|-------|-------------|-------------------------|-------------------------|-------------|-------------------------|-------------------------|-------------|-------------------------|-------------------------|
|     | Criterion  | Sep. | Recg. | Si-SNR simu | WER simu                | WER replay              | Si-SNR simu | WER simu                | WER replay              | Si-SNR simu | WER simu                | WER replay              |
| 1   | Not Applied  |      |       | 10.16       | 26.1                    | 28.0                    | 11.60       | 23.5                    | 30.9                    | 9.03        | 26.1                    | 25.8                    |
| 2   | $\mathcal{L}_{CTC}$                                  | ✗    | ✓     | 10.16       | 22.9                    | 23.2                    | 11.60       | 19.2                    | 24.1                    | 9.03        | 19.3                    | 17.3                    |
| 3   | $\mathcal{L}_{CTC}$                                  | ✓    | ✓     | 8.15        | 19.3 <sup>†</sup>       | 18.0 <sup>†</sup>       | 6.04        | 17.2 <sup>†</sup>       | 19.9 <sup>†</sup>       | 4.14        | 18.6 <sup>†</sup>       | 16.9 <sup>†</sup>       |
| 4   | $\mathcal{L}_{CTC} + \alpha \mathcal{L}_{Si-SNR}$    | ✓    | ✓     | 8.50        | <b>18.6<sup>†</sup></b> | <b>18.0<sup>†</sup></b> | 9.17        | <b>16.1<sup>†</sup></b> | <b>19.2<sup>†</sup></b> | 7.72        | <b>18.4<sup>†</sup></b> | <b>16.9<sup>†</sup></b> |
| 5   | Not Applied  |      |       | 10.16       | 23.8                    | 26.2                    | 11.60       | 21.1                    | 28.1                    | 9.03        | 23.1                    | 22.8                    |
| 6   | $\mathcal{L}_{LF-MMI}$                               | ✗    | ✓     | 10.16       | 20.7                    | 21.4                    | 11.60       | 18.2                    | 25.1                    | 9.03        | 20.3                    | 20.1                    |
| 7   | $\mathcal{L}_{LF-MMI}$                               | ✓    | ✓     | 8.03        | 17.7 <sup>‡</sup>       | 18.7 <sup>‡</sup>       | 9.20        | 16.9 <sup>‡</sup>       | 22.4 <sup>‡</sup>       | 5.65        | 16.3 <sup>‡</sup>       | 15.5 <sup>‡</sup>       |
| 8   | $\mathcal{L}_{LF-MMI} + \alpha \mathcal{L}_{Si-SNR}$ | ✓    | ✓     | 8.89        | <b>17.7<sup>‡</sup></b> | <b>18.3<sup>‡</sup></b> | 10.73       | <b>16.6<sup>‡</sup></b> | <b>21.6<sup>‡</sup></b> | 8.40        | <b>16.1<sup>‡</sup></b> | <b>15.0<sup>‡</sup></b> |

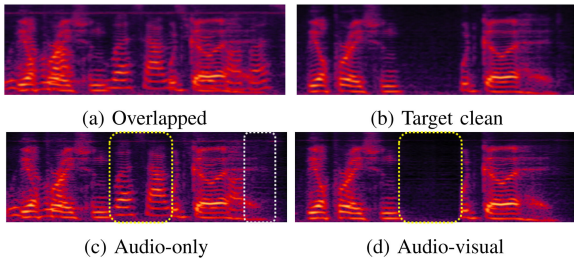


Fig. 7. Example spectra of overlapped, target clean, audio-only separated, and audio-visual separated speech segments obtained using the  $TF$  masking based speech separation front-end of Fig. 1(a). The spectrum portions circled using yellow dotted lines in (c) and (d) represent the interfering speaker's speech, which is almost removed in (d) after applying audio-visual speech separation.

which is largely removed in (d) after applying audio-visual speech separation.<sup>2</sup>

### C. Performance of Different Fine-Tuning Methods

The WER results of the audio-visual multi-channel system using different fine-tuning approaches aiming for integrating the separation front-end and recognition back-end are shown in Table II. In these experiments, the visual features are used in both the separation and the recognition components, while the AF features are adopted in the separation front-end only. Before integration, both the separation and the recognition components are trained separately. In the pipelined systems (sys.2, sys.5), the recognition back-ends are fine-tuned using the separation outputs, while the separation front-ends are kept unchanged. In the jointly fine-tuned systems, both the separation and the recognition components are fine-tuned using the recognition cost function (sys.2, sys.5) or multi-task criterion (sys.3, sys.6). For the CTC based system,  $\alpha$  is set as 0.1 for the  $TF$  masking approach and 1 for the  $Filter\&Sum$  and  $mask$ -based  $MVDR$  approaches. For the LF-MMI system,  $\alpha$  is set as 0.01 (larger  $\alpha$  will lead to performance degradation) for all the three channel integration approaches.

Several trends can be observed from Table III:

- 1) Compared with the audio-visual speech recognition systems without any separation front-ends in Table I (sys.10,

<sup>2</sup>More examples of audio-visual multi-channel speech separation can be found in: <https://yjuw123456.github.io/Audio-visual-Multi-channel-Integration-and-Recognition-of-Overlapped-Speech/>

- 2) The jointly fine-tuned systems consistently outperform the pipelined systems for all the three channel integration methods (sys.2 vs. sys.3, sys.6 vs. sys.7), which confirms our arguments in Section III-B.
- 3) Compared with the jointly fine-tuned systems using only the recognition cost, systems using multi-task criterion only provide marginal recognition improvements (sys.3 vs. sys.4, sys.7 vs. sys.8).
- 4) Different from the trend in Table I, the LF-MMI based jointly fine-tuned systems do not always outperform the CTC based systems, especially on the replay test set.
- 5) Jointly fine-tuning the separation and the recognition components using only the speech recognition cost degrades the speech separation performance in terms of Si-SNR (sys.2 vs. sys.3, sys.6 vs. sys.7) by up to 4.9dB. However, jointly fine-tuning these two components by multi-task criterion only degrades the Si-SNR performance by up to 1.27dB (sys.2 vs. sys.4, sys.6 vs. sys.8).

Considering the average performance contrast between the CTC and LF-MMI costs fine-tuned systems over three beamforming methods on both the simulated and the replayed data in Table II (sys.3 vs. sys.6), we adopt jointly fine-tuned systems using only the CTC cost function in all subsequent experiments.

### D. Performance of Audio-Visual Multi-Channel AVSR Systems

In this section, we systematically investigate the performance improvements attributed to the visual modality in three types of audio-visual multi-channel overlapped speech recognition systems featuring  $TF$  masking,  $Filter\&Sum$  and  $mask$ -based  $MVDR$  neural beamformers. The visual modality's impact on system performance is further analysed in a more advanced AVSR system configuration when it is used in combination with the angle features described previously in Section III-A. To compare the performance between the conventional channel integration methods with the NN based methods, the traditional frequency domain delay and sum ( $Delay\&Sum$ ) beamformer is also adopted in this experiment. The steering vectors used in such beamformer are computed based on the ground truth

TABLE IV

PERFORMANCE OF AUDIO-ONLY AND AUDIO-VISUAL OVERLAPPED SPEECH RECOGNITION SYSTEMS USING VARIOUS CHANNEL INTEGRATION METHODS.

THE SEPARATION AND THE RECOGNITION COMPONENTS ARE JOINTLY FINE-TUNED USING THE CTC LOSS. "AF" DENOTES ANGLE FEATURE. †, ‡ AND \* DENOTES A STATISTICALLY SIGNIFICANT IMPROVEMENT IS OBTAINED OVER THE *TF MASKING* (SYS.5), *FILTER&SUM* (SYS.10) AND *MASK-BASED MVDR* (SYS.15) AUDIO-ONLY BASELINE SYSTEMS

| Sys | Separation      |    |         | Recognition | WER(%)        |               |
|-----|-----------------|----|---------|-------------|---------------|---------------|
|     | method          | AF | +visual | +visual     | simu          | replay        |
| 1   | raw channel 1   |    |         | ✗           | 75.36         | 80.55         |
| 2   | raw channel 1   |    |         | ✓           | 32.06         | 31.93         |
| 3   | Delay&Sum       | ✓  | -       | ✗           | 49.25         | 44.34         |
| 4   |                 | ✓  | -       | ✓           | 25.81         | 24.46         |
| 5   | TF masking      | ✓  | ✗       | ✗           | 33.12         | 46.75         |
| 6   |                 | ✗  | ✓       | ✗           | 24.64†        | 26.49†        |
| 7   |                 | ✓  | ✓       | ✗           | 23.17†        | 23.59†        |
| 8   |                 | ✗  | ✓       | ✓           | 21.32†        | 21.52†        |
| 9   |                 | ✓  | ✓       | ✓           | <b>19.25†</b> | <b>18.03†</b> |
| 10  | Filter&Sum      | ✓  | ✗       | ✗           | 30.24         | 43.83         |
| 11  |                 | ✗  | ✓       | ✗           | 23.09‡        | 24.67‡        |
| 12  |                 | ✓  | ✓       | ✗           | 21.77‡        | 24.66‡        |
| 13  |                 | ✗  | ✓       | ✓           | 21.02‡        | 20.02‡        |
| 14  |                 | ✓  | ✓       | ✓           | <b>17.21‡</b> | <b>19.87‡</b> |
| 15  | Mask-based MVDR | ✓  | ✗       | ✗           | 25.38         | 39.07         |
| 16  |                 | ✗  | ✓       | ✗           | 23.96*        | 23.48*        |
| 17  |                 | ✓  | ✓       | ✗           | 23.41*        | 21.17*        |
| 18  |                 | ✗  | ✓       | ✓           | <b>17.34*</b> | <b>16.21*</b> |
| 19  |                 | ✓  | ✓       | ✓           | 18.57*        | 16.85*        |

DOA for the simulated data and the approximated DOA for the replayed data.

From Table IV, several trends can be observed:

- (1) Adding visual features can significantly improve the recognition performance on both the simulated and the replayed overlapped speech by up to 13.87% and 28.72% (sys.5 vs. sys.9), 13.03% and 23.96% (sys.10 vs. sys.14), 8.04% and 22.86% (sys.15 vs. sys.18) absolute WER reduction for the *TF masking*, *Filter&Sum* and *mask-based MVDR* approaches respectively.
- (2) When we only use the visual, but not the angle features in the proposed audio-visual multi-channel AVSR systems (sys.8, sys.13, sys.18), similar recognition performance is retained on both simulated and replayed data for all the three channel integration methods (sys.8 vs. sys.9, sys.13 vs. sys.14, sys.18 vs. sys.19).
- (3) Using visual information in both the separation and the recognition back-ends performs better than using visual information only in the separation front-ends. (sys.7 vs. sys.9, sys.12 vs. sys.14, sys.17 vs. sys.19)
- (4) When we only use the angle features, a large performance gap between the simulated and the replayed data can be observed (sys.5,10,15). Since we use the ground truth DOA for the simulated data and approximated DOA for the replayed data, this phenomenon indicates that these three systems (sys.5,10,15) are sensitive to the precision of the DOA estimation. However, by adding visual features (sys.6-9, sys.11-14, sys.16-19), such performance gap is narrowed down greatly, which further confirms the efficacy of the visual modality.

- (5) The NN based separation front-ends (sys.5-19) outperform the conventional *Delay&Sum* beamformer (sys.3-4), which confirms the strength of the NN based channel integration methods. In addition, compared with the *TF masking* (sys.5-9) and *Filter&Sum* (sys.10-14) approaches, the *mask-based MVDR* systems (sys.15-19) show better performance on the replayed data set.

### E. Impact of Low-Resolution Visual Inputs

In this section, we further investigate the robustness of the proposed *TF masking* and *mask-based MVDR* multi-channel AVSR systems in Table IV (sys.6-9 and sys.16-19) when lower resolution video inputs are used, as previously described in Section V-D. Fig. 8 shows the relationship between WER and visual input resolution for the *TF masking* and the *mask-based MVDR* based AVSR systems. Several trends can be observed from Fig. 8:

- 1) Although low-resolution visual inputs can cause performance degradation, the proposed systems consistently outperform the baseline audio-only systems even when the video resolution is aggressively reduced to as low as  $40 \times 40$  pixels down from the full resolution of  $160 \times 160$ .
- 2) The *mask-based MVDR* ASR and AVSR systems are more robust to low resolution visual inputs than the *TF masking* based comparable ASR and AVSR systems.

### F. Impact of Visual Occlusion

In this section, we further investigate the robustness of the proposed AVSR multi-channel recognition systems when the video data quality is degraded by visual occlusion, as previously described in Section V-D.

As shown in Table V, three methods are considered to improve the system performance with occluded visual inputs:

1) **Using angle features:** With the DOA information contained by angle features, the negative influence of visual input occlusion can be alleviated. In the *TF masking* systems, using AF consistently improves the system robustness to visual occlusions (sys.4 vs. sys.5, sys.8 vs. sys.9, sys.12 vs. sys.13). For the *mask-based MVDR* systems, angle features show their effectiveness when the occlusion rate is larger than 60% (sys.17 vs. sys.18, sys.25 vs. sys.26).

2) **Using multi-style occluded data:** In order to improve the generalization to video occlusion at various percentage settings described in Section V-D, a 400-hour multi-style audio-visual training data set containing a 200h subset with video occlusion applied and the remaining half based on the original 200h data without occlusion was used to fine-tune the *TF masking* and the *mask-based MVDR* multi-channel AVSR systems in Table IV (sys.6-9 and sys.16-19) using the CTC cost function. A general trend can be found in Fig. 9 (systems fine-tuned using multi-style occluded data shown on red lines) and Table V: the multi-style occluded data fine-tuned audio-visual multi-channel recognition systems (sys.13 and sys.26 in Table V) consistently outperform the baseline audio-only systems (Sys.1, 14 in Table V) even when using occluded video input with the lip region randomly covered up to 80% for the *TF masking* (sys.1 vs. sys.13) and 60%

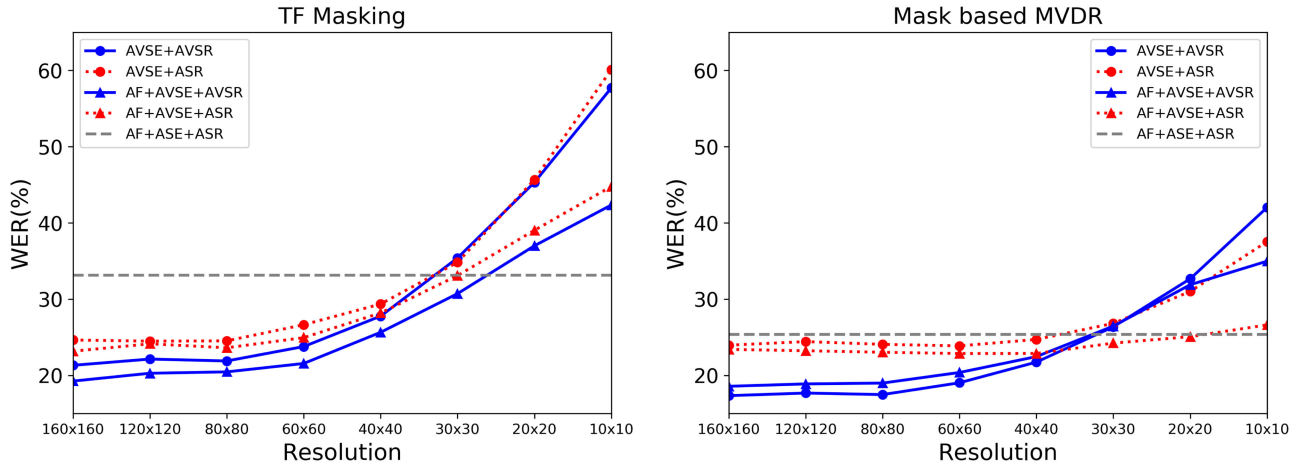


Fig. 8. WER(%) of *TF masking* and *mask-based MVDR* based AVSR systems of Table IV using different visual inputs resolutions ranging from  $160 \times 160$  to  $10 \times 10$ . “AVSE” and “AVSR” denote using visual modality in the separation front-end and recognition back-end respectively, “AF+” stands for optionally using angle features.

TABLE V

WER(%) OF CTC BASED *TF MASKING* AND *MASK-BASED MVDR* BASED AVSR SYSTEMS WHEN EVALUATED ON DATA WITH VISUAL OCCLUSION RANGING FROM 0% UP TO 80% COVERAGE OF THE LIP REGION. † AND ‡ DENOTES A STATISTICALLY SIGNIFICANT IMPROVEMENT IS OBTAINED OVER THE *TF MASKING* (SYS.1) AND *MASK-BASED MVDR* (SYS.14) BASED AUDIO-ONLY SYSTEMS

| Sys | Separation      |    |         | Recognition<br>+visual | Training set |      | Test set    | WER(%)                   |                          |                          |                          |                          |  |
|-----|-----------------|----|---------|------------------------|--------------|------|-------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--|
|     | method          | AF | +visual |                        | no-occ       | occ  |             | 0%                       | 20%                      | 40%                      | 60%                      | 80%                      |  |
| 1   |                 | ✓  | ✗       | ✗                      | 200h         | ✗    | occ         |                          |                          | 33.12                    |                          |                          |  |
| 2   |                 | ✗  | ✓       | ✗                      |              |      |             | 24.64 <sup>†</sup>       | 30.02 <sup>†</sup>       | 34.91                    | 39.88                    | 44.52                    |  |
| 3   |                 | ✓  | ✓       | ✗                      | 200h         | ✗    | occ         | 23.17 <sup>†</sup>       | 27.88 <sup>†</sup>       | 32.48                    | 36.56                    | 39.77                    |  |
| 4   |                 | ✗  | ✓       | ✓                      |              |      |             | 21.32 <sup>†</sup>       | 28.30 <sup>†</sup>       | 34.86                    | 40.42                    | 44.83                    |  |
| 5   |                 | ✓  | ✓       | ✓                      |              |      |             | <b>19.25<sup>†</sup></b> | <b>24.82<sup>†</sup></b> | <b>30.00<sup>†</sup></b> | <b>33.39</b>             | <b>37.90</b>             |  |
| 6   |                 | ✗  | ✓       | ✗                      |              |      |             | 24.64 <sup>†</sup>       | 31.06 <sup>†</sup>       | 34.43                    | 36.33                    | 39.52                    |  |
| 7   | TF masking      | ✓  | ✓       | ✗                      | 200h         | ✗    | in-painting | 23.17 <sup>†</sup>       | 29.63 <sup>†</sup>       | 31.13 <sup>†</sup>       | 33.84                    | 35.22                    |  |
| 8   |                 | ✗  | ✓       | ✓                      |              |      |             | 21.32 <sup>†</sup>       | 29.80 <sup>†</sup>       | 33.94                    | 38.03                    | 41.91                    |  |
| 9   |                 | ✓  | ✓       | ✓                      |              |      |             | <b>19.25<sup>†</sup></b> | <b>26.22<sup>†</sup></b> | <b>29.72<sup>†</sup></b> | <b>32.17<sup>†</sup></b> | <b>34.28</b>             |  |
| 10  |                 | ✗  | ✓       | ✗                      |              |      |             | 24.43 <sup>†</sup>       | 27.57 <sup>†</sup>       | 29.8 <sup>†</sup>        | 32.39                    | 34.44                    |  |
| 11  |                 | ✓  | ✓       | ✗                      | 200h         | 200h | occ         | 23.39 <sup>†</sup>       | 25.5 <sup>†</sup>        | 27.54 <sup>†</sup>       | 29.44 <sup>†</sup>       | 30.87 <sup>†</sup>       |  |
| 12  |                 | ✗  | ✓       | ✓                      |              |      |             | 20.75 <sup>†</sup>       | 24.46 <sup>†</sup>       | 28.53 <sup>†</sup>       | 32.76                    | 35.12                    |  |
| 13  |                 | ✓  | ✓       | ✓                      |              |      |             | <b>18.57<sup>†</sup></b> | <b>21.02<sup>†</sup></b> | <b>23.78<sup>†</sup></b> | <b>26.04<sup>†</sup></b> | <b>28.21<sup>†</sup></b> |  |
| 14  |                 | ✓  | ✗       | ✗                      | 200h         | ✗    | occ         |                          |                          | 25.38                    |                          |                          |  |
| 15  |                 | ✗  | ✓       | ✗                      |              |      |             | 23.96 <sup>‡</sup>       | 24.55 <sup>‡</sup>       | 26.70                    | 28.59                    | 31.20                    |  |
| 16  |                 | ✓  | ✓       | ✗                      | 200h         | ✗    | occ         | 23.41 <sup>‡</sup>       | 23.81 <sup>‡</sup>       | 23.99 <sup>‡</sup>       | 24.55 <sup>‡</sup>       | 25.36                    |  |
| 17  |                 | ✗  | ✓       | ✓                      |              |      |             | <b>17.34<sup>‡</sup></b> | <b>21.92<sup>‡</sup></b> | <b>25.03</b>             | 29.62                    | 34.01                    |  |
| 18  |                 | ✓  | ✓       | ✓                      |              |      |             | 18.57 <sup>‡</sup>       | 22.66 <sup>‡</sup>       | 25.66                    | <b>28.77</b>             | <b>31.61</b>             |  |
| 19  |                 | ✗  | ✓       | ✗                      |              |      |             | 23.96 <sup>‡</sup>       | 25.03                    | 25.50                    | 26.79                    | <b>27.35</b>             |  |
| 20  | Mask-based MVDR | ✓  | ✓       | ✗                      | 200h         | ✗    | in-painting | 23.41 <sup>‡</sup>       | 24.28 <sup>‡</sup>       | <b>24.47<sup>‡</sup></b> | <b>24.41<sup>‡</sup></b> | <b>24.47<sup>‡</sup></b> |  |
| 21  |                 | ✗  | ✓       | ✓                      |              |      |             | <b>17.34<sup>‡</sup></b> | <b>22.11<sup>‡</sup></b> | 24.64 <sup>‡</sup>       | 27.72                    | 29.54                    |  |
| 22  |                 | ✓  | ✓       | ✓                      |              |      |             | 18.57 <sup>‡</sup>       | 22.64 <sup>‡</sup>       | 25.47                    | 27.77                    | 29.13                    |  |
| 23  |                 | ✗  | ✓       | ✗                      |              |      |             | 23.62 <sup>‡</sup>       | 24.17 <sup>‡</sup>       | 25.63                    | 25.53                    | 27.16                    |  |
| 24  |                 | ✓  | ✓       | ✗                      | 200h         | 200h | occ         | 22.85 <sup>‡</sup>       | 23.35 <sup>‡</sup>       | 23.62 <sup>‡</sup>       | 24.16 <sup>‡</sup>       | <b>24.27<sup>‡</sup></b> |  |
| 25  |                 | ✗  | ✓       | ✓                      |              |      |             | <b>16.20<sup>‡</sup></b> | <b>19.10<sup>‡</sup></b> | <b>21.82<sup>‡</sup></b> | 23.75 <sup>‡</sup>       | 26.74                    |  |
| 26  |                 | ✓  | ✓       | ✓                      |              |      |             | 18.08 <sup>‡</sup>       | 20.27 <sup>‡</sup>       | 22.27 <sup>‡</sup>       | <b>23.60<sup>‡</sup></b> | 25.80                    |  |

for the mask-based MVDR (sys.14 vs. sys.26) multi-channel systems.

3) **In-painting:** A visual in-painting neural network following [39] was trained using the occluded Train-val set of LRS2 to in-paint the occluded visual image. Fig. 10 shows some examples of the occluded images before and after being restored using the in-painting approach. If the input image is occluded, the in-painting network can restore the occluded region with

some distortion. In contrast, if the video is not occluded, the in-painting network will keep the image almost unchanged. From Fig. 9 (line in green vs. line in blue) and Table V (sys.2-5 vs. sys.6-9, sys.15-18 vs. sys.19-22) the following can be observed: using in-painting neural network can improve both the *TF-masking* and the *mask-based MVDR* multi-channel AVSR systems’ robustness to visual occlusion when the occlusion rate is 60% or above.

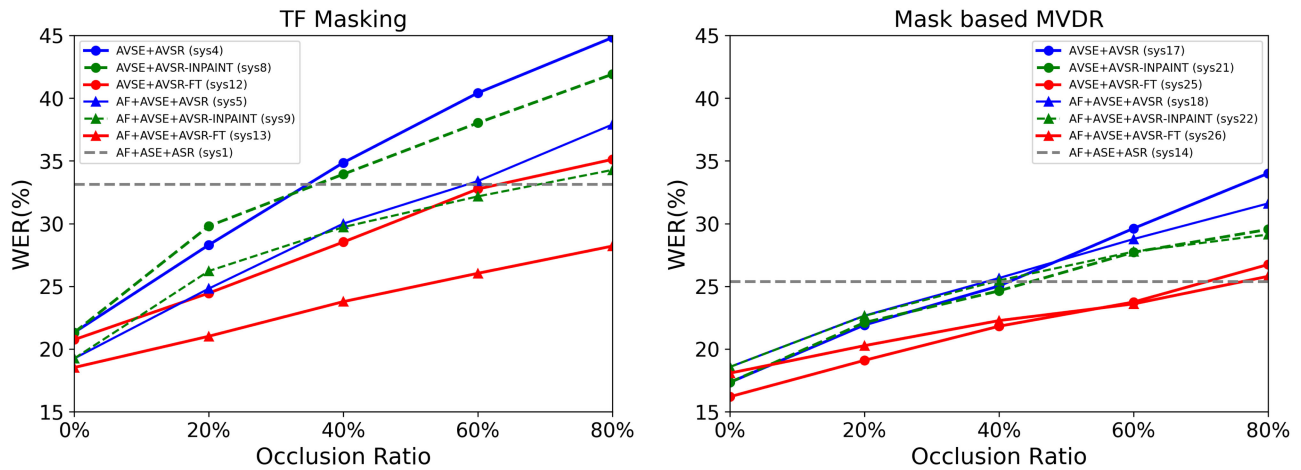


Fig. 9. WER(%) of the CTC based *TF masking* and *mask-based MVDR* based AVSR systems of Table V when evaluated on data with visual occlusion ranging from 0% up to 80% coverage of the lip region. “AVSE” and “AVSR” denote using visual modality in the separation front-end and recognition back-end respectively, “AF+” stands for optionally using angle features, “+FT” denotes fine-tuning the system on multi-style data mixed with original and occluded video inputs, “+INPAINT” denotes using the in-painting network to restore the occluded video.



Fig. 10. Examples of (a) original video snapshots; (b) randomly occluded video; and (c) restored video obtained using the in-painting network.

Several additional trends can be observed from Table V and Fig. 9:

- (1) As shown in Fig. 8, the *mask-based MVDR* based systems are more robust to visual occlusion than the *TF-masking* based systems. One possible explanation is that the *mask-based MVDR* filter estimation exploits audio-video information across the entire speech segment and thus more robust to the partial, if not complete, occlusion being applied to the video data. This is different the other beamforming methods where no explicit constraint on using such longer span spatial-temporal contexts is enforced.
- (2) The multi-style occluded data fine-tuning method outperforms the in-painting method (line in red vs. line in green, Fig. 9). One possible explanation is that the in-painting network only use the visual information from the current occluded image frame to explicitly recover the occluded image, while during multi-style occluded data fine-tuning, both the speech separation front-end and the recognition back-end will learn the systematic variability among the occluded videos of the same audio contents but with different percentage of occlusion. This allows the resulting *TF-masking* or *mask-based MVDR*

AVSR systems to implicitly build connection between the original and occluded data of the same audio and thus improves their robustness against video occlusion.

## VII. CONCLUSION

In this work, we propose an audio-visual multi-channel based recognition system for overlapped speech. A series of audio-visual multi-channel speech separation front-ends based on *TF masking*, *Filter&Sum*, and *mask-based MVDR* are developed. Jointly fine-tuning approaches are studied to integrate the separation and the recognition components. The impact of visual occlusion is also investigated. Experiments suggest that the proposed system significantly outperforms the baseline audio-only multi-channel ASR system on overlapped speech constructed using either simulation or replaying of the LRS2 dataset, which demonstrate the advantages of the visual information for overlapped speech recognition. In the future, this work will be extended to: 1) further integrating an audio-video de-reverberation component; 2) multi-input multi-output AVSR systems facilitating speech separation and recognition for multi-talkers’ speech, 3) more advanced visual occlusion restoration methods to address visual occlusion issue.

## ACKNOWLEDGMENT

The authors would like to thank Yiwen Shao and Yiming Wang for the deep discussion about the LF-MMI implementation details.

## REFERENCES

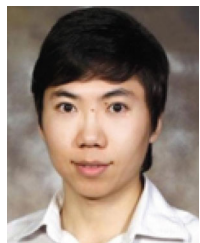
- [1] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines,” in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 504–511.
- [2] T. Yoshioka *et al.*, “The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices,” in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 436–443.

- [3] M. Harper, "The automatic speech recognition in reverberant environments (aspire) challenge," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 547–554.
- [4] T. Yoshioka, H. Erdogan, Z. Chen, X. Xiao, and F. Alleva, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," in *Proc. INTERSPEECH*, 2018, pp. 3038–3042.
- [5] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "MIMO-speech: End-to-end multi-channel multi-speaker speech recognition," *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 237–244.
- [6] X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, "End-to-end multi-speaker speech recognition with transformer," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6134–6138.
- [7] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.
- [8] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2011–2022, Sep. 2007.
- [9] D. A. Pados and G. N. Karystinos, "An iterative algorithm for the computation of the MVDR filter," *IEEE Trans. Signal, Process.*, vol. 49, no. 2, pp. 290–300, Feb. 2001.
- [10] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 260–276, Feb. 2010.
- [11] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1529–1539, Jul. 2007.
- [12] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.
- [13] M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.
- [14] F. Bahmaninezhad *et al.*, "A comprehensive study of speech separation: Spectrogram vs waveform separation," in *Proc. INTERSPEECH*, 2019, pp. 4574–4578.
- [15] L. Chen, M. Yu, D. Su, and D. Yu, "Multi-band pit and model integration for improved multi-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 705–709.
- [16] T. N. Sainath *et al.*, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 965–979, May 2017.
- [17] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S.-C. Liu, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 260–267.
- [18] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal, Process.*, 2020, pp. 6394–6398.
- [19] X. Xiao *et al.*, "Deep beamforming networks for multi-channel speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 5745–5749.
- [20] Z.-Q. Wang and D. Wang, "On spatial features for supervised speech separation and its application to beamforming and robust ASR," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5709–5713.
- [21] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. INTERSPEECH*, 2016, pp. 1981–1985.
- [22] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 558–565.
- [23] T. Yoshioka *et al.*, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5739–5743.
- [24] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 196–200.
- [25] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 5325–5329.
- [26] Y. Xu *et al.*, "Joint training of complex ratio mask based beamformer and acoustic model for noise robust ASR," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6745–6749.
- [27] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. INTERSPEECH*, 2018, pp. 3244–3248.
- [28] J. Wu *et al.*, "Time domain audio visual speech separation," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 667–673.
- [29] T. Afouras, J. S. Chung, and A. Zisserman, "My lips are concealed: Audio-visual speech enhancement through obstructions," in *Proc. INTERSPEECH*, 2019, pp. 4295–4299.
- [30] G.-L. Chao, W. Chan, and I. Lane, "Speaker-targeted audio-visual models for speech recognition in cocktail-party environments," in *Proc. INTERSPEECH*, 2016, pp. 2120–2124.
- [31] J. Yu *et al.*, "Audio-visual recognition of overlapped speech for the LRS2 dataset," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6984–6988.
- [32] T. Makino *et al.*, "Recurrent neural network transducer for audio-visual speech recognition," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 905–912.
- [33] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multimodal multi-channel target speech separation," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 530–541, Mar. 2020.
- [34] K. Tan, Y. Xu, S. Zhang, M. Yu, and D. Yu, "Audio-visual speech separation and dereverberation with a two-stage multimodal network," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 542–553, Mar. 2020.
- [35] S. Zhang, M. Lei, B. Ma, and L. Xie, "Robust audio-visual speech recognition using bimodal DFSMN with multi-condition training and dropout regularization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6570–6574.
- [36] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.
- [37] D. Povey *et al.*, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. INTERSPEECH*, 2016, pp. 2751–2755.
- [38] Y. Shao, Y. Wang, D. Povey, and S. Khudanpur, "Pychain: A fully parallelized pytorch implementation of LF-MMI for end-to-end ASR," in *Proc. INTERSPEECH*, 2020, pp. 561–565.
- [39] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2536–2544.
- [40] R. Gu *et al.*, "End-to-end multi-channel speech separation," 2019, *arXiv:1905.06286*.
- [41] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: [10.1109/TPAMI.2018.2889052](https://doi.org/10.1109/TPAMI.2018.2889052).
- [42] G.-L. Chao, W. Chan, and I. Lane, "Speaker-targeted audio-visual models for speech recognition in cocktail-party environments," in *Proc. INTERSPEECH*, 2016, pp. 2120–2124.
- [43] T. von Neumann *et al.*, "End-to-end training of time domain audio separation and recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7004–7008.
- [44] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends Amplification*, vol. 12, no. 4, pp. 332–353, 2008.
- [45] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
- [46] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.
- [47] Y. Hu *et al.*, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. INTERSPEECH* 2020, pp. 2472–2476.
- [48] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.
- [49] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.
- [50] T. Afouras, J. S. Chung, and A. Zisserman, "Deep lip reading: A comparison of models and an online application," in *Proc. INTERSPEECH*, 2018, pp. 3514–3518.

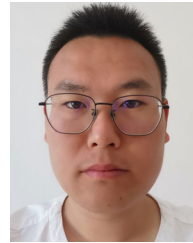
- [51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [52] J. Yu *et al.*, "Audio-visual multi-channel recognition of overlapped speech," in *Proc. INTERSPEECH*, 2020, pp. 3496–3500.
- [53] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, "Audio-visual speech recognition with a hybrid CTC/attention architecture," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 513–520.
- [54] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graphics (TOG)*, vol. 37, no. 4, pp. 1–11, 2018.
- [55] S. Dupont and J. Luetttin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, Sep. 2000.
- [56] G. Potamianos, C. Neti, J. Luetttin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *Issues Vis. Audio-Vis. Speech Process.*, Cambridge University Press, pp. 193–247, 2012.
- [57] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7596–7599.
- [58] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Appl. Intell.*, vol. 42, no. 4, pp. 722–737, 2015.
- [59] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4580–4584.
- [60] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "Flat-start single-stage discriminatively trained HMM-based models for ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 1949–1961, Nov. 2018.
- [61] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free MMI," in *Proc. INTERSPEECH*, 2018, pp. 12–16.
- [62] M. W. Lam, J. Wang, X. Liu, H. Meng, D. Su, and D. Yu, "Extract, adapt and recognize: An end-to-end neural network for corrupted monaural speech recognition," in *Proc. INTERSPEECH*, 2019, pp. 2778–2782.
- [63] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3444–3453.
- [64] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 269–277, 2008.



**Jianwei Yu** received the B.Sc. degree in physics from Nanjing University, Nanjing, China, in 2017. He is currently the Ph.D. student with the Chinese University of Hong Kong. His current research interests include noise robust speech recognition, audio-visual speech recognition, Bayesian modeling, and language model.



**Shi-Xiong Zhang** (Member, IEEE) received the M.Phil. degree in electronic and information engineering from The Hong Kong Polytechnic University and the Ph.D. degree with the Machine Intelligence Laboratory, Engineering Department, Cambridge University, in 2014. From 2014 to 2018, he was a Senior Speech Scientist with Microsoft, Speech Group. He is currently a Principal Researcher with Tencent America. His research interests include speech recognition, speaker verification, speech separation, multi-modal learning and machine learning (particularly structured prediction, graphical models, kernel methods and Bayesian non-parametric methods). He was granted the "IC Greatness award" in Microsoft in 2015. Shi-Xiong Zhang was nominated a 2011 Interspeech Best Student Paper Award for his paper "Structured Support Vector Machines for Noise Robust Continuous Speech Recognition". He was awarded Best Paper Award in 2008 IEEE Signal Processing Postgraduate Forum for his paper "Articulatory-Feature based Sequence Kernel For High-Level Speaker Verification". He was as a Session Chair of the ICASSP 2021.

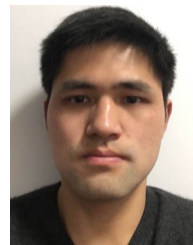


speech enhancement and recognition.

**Bo Wu** received the B.Eng. degree in electronic information engineering from Southwest University, Chongqing, China, in 2012. He received the Ph.D. degree from the National Laboratory of Radar Signal Processing, Xidian University, Xi'an, China, in 2018. From September 2014 to October 2016, he was a Visiting Student with the Center for Signal and Information Processing, Georgia Institute of Technology, Atlanta, GA, USA. He is currently a Senior Research Scientist with Tencent AI lab. His current research interests include signal processing, machine learning,



**Shansong Liu** received the B.E. degree in automation from Sichuan University, Chengdu, China, in 2014, and the M.S. degree in control science and engineering from Tsinghua University, Beijing, China, in 2017. He is currently the Ph.D. student with the Chinese University of Hong Kong. His current research interests include disordered speech recognition and multi-modal speech recognition.



**Shoukang Hu** (Graduate Student, Member, IEEE) received the B.E. degree in mechanical and electrical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2017. He is currently the Ph.D. student with the Chinese University of Hong Kong. His current research interests include speech recognition, Bayesian modeling and neural architecture search.



**Mengzhe Geng** received the B.Sc. degree in mathematics and information engineering from the Chinese University of Hong Kong in 2019. He is currently the Ph.D. student with the Chinese University of Hong Kong. His current research interests include data augmentation and speaker adaptation.



**Xunying Liu** (Member, IEEE) received the Ph.D. degree in speech recognition and the M.Phil. degree in computer speech and language processing both from the University of Cambridge, prior to his undergraduate study Shanghai Jiao Tong University. He had been a Senior Research Associate with the Machine Intelligence Laboratory of the Cambridge University Engineering Department, and from 2016, has been an Associate Professor with the Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong. He and his students recipient of the number of Best Paper Awards and Nominations, including a Best Paper Award at ISCA Interspeech2010 for the paper titled "Language Model Cross Adaptation for LVCSR System Combination," and a Best Paper Award at IEEE ICASSP2019 for their paper titled "BLHUC: Bayesian Learning of Hidden Unit Contributions for Deep Neural Network Speaker Adaptation". His current research interests include machine learning, large vocabulary continuous speech recognition, statistical language modeling, noise robust speech recognition, audio-visual speech recognition, computer aided language learning, speech synthesis, and assistive technology. Dr. Xunying Liu is a Member of IEEE and ISCA.



**Helen Meng** (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA. In 1998, she joined the Chinese University of Hong Kong, Hong Kong, where she is currently the Chair Professor with the Department of Systems Engineering and Engineering Management. She was the former Department Chairman and the Associate Dean of Research with the faculty of Engineering. Her research interests include human-computer interaction via multimodal and multilingual spoken language

systems, spoken dialog systems, computer-aided pronunciation training, speech processing in assistive technologies, health-related applications, and big data decision analytics. She was the Editor-in-Chief of the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING between 2009 and 2011. She was the recipient of the IEEE Signal Processing Society Leo L. Beranek Meritorious Service Award in 2019. She was also on the Elected Board Member of the International Speech Communication Association (ISCA) and an International Advisory Board Member. She is a ISCA, HKCS, and HKIE.



**Dong Yu** (Fellow, IEEE) is a Distinguished Scientist and Vice General Manager with Tencent AI Lab. Prior to joining Tencent in 2017, he was a Principal Researcher with Microsoft Research (Redmond), Microsoft, where he joined in 1998. He has been focusing his research on speech recognition and processing and has published two monographs and more than 250 papers. His works have been cited for more than 40000 times per Google Scholar and have been recognized by the prestigious IEEE Signal Processing Society 2013, 2016, 2020 Best Paper Award. Dr. Dong

Yu is currently the Chair of the IEEE Speech and Language Processing Technical Committee (SLPTC) and Technical Co-Chair of ICASSP 2021 and DSLW 2021. He was a Member of the IEEE SLPTC (2013–2018), a Distinguished Lecturer of APSIPA (2017–2018), an Associate Editor of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (TransASLP) (2011–2015) and an Associate Editor of the IEEE SIGNAL PROCESSING MAGAZINE (2008–2011).