Audio–Visual Deep Clustering for Speech Separation

Rui Lu, *Student Member, IEEE*, Zhiyao Duan, *Member, IEEE*, and Changshui Zhang, *Fellow, IEEE*

***Abstract*—Speech separation aims to separate individual voices from an audio mixture of multiple simultaneous talkers. Audio-only approaches show unsatisfactory performance when the speakers are of the same gender or share similar voice characteristics. This is due to challenges on learning appropriate feature representations for separating voices in single frames and streaming voices across time. Visual signals of speech (e.g., lip movements), if available, can be leveraged to learn better feature representations for separation. In this paper, we propose a novel audio–visual deep clustering model (AVDC) to integrate visual information into the process of learning better feature representations (embeddings) for Time–Frequency (T–F) bin clustering. It employs a two-stage audio–visual fusion strategy where speaker-wise audio–visual T–F embeddings are first computed after the first-stage fusion to model the audio–visual correspondence for each speaker. In the second-stage fusion, audio–visual embeddings of all speakers and audio embeddings calculated by deep clustering from the audio mixture are concatenated to form the final T–F embedding for clustering. Through a series of experiments, the proposed AVDC model is shown to outperform the audio-only deep clustering and utterance-level permutation invariant training baselines and three other state-of-the-art audio–visual approaches. Further analyses show that the AVDC model learns a better T–F embedding for alleviating the source permutation problem across frames. Other experiments show that the AVDC model is able to generalize across different numbers of speakers between training and testing and shows some robustness when visual information is partially missing.**

***Index Terms*—Speaker-independent speech separation, deep clustering, audio-visual fusion.**

I. Introduction

SPEECH separation, i.e., the *cocktail party problem*, aims at separating individual voices from audio mixtures of multiple simultaneous talkers. It is a fundamental problem in computer audition, and its success would greatly improve many speech applications such as Automatic Speech Recognition (ASR), dialogue systems, multimedia retrieval and hearing aids. When the number of channels is smaller than the number of sources, the problem is *under-determined*. In the extreme case, *single-channel* or *monaural* speech separation, is most difficult. Over the past decades, a variety of approaches have been proposed, however, state-of-the-art audio-based approaches still significantly under-perform humans.

In a cocktail party, human listeners use multiple cues from different modalities to pay attention to a target talker. Visual cues, especially lip movements, are among them. Thanks to the correlation between lip movements and acoustic signals, seeing lip movements greatly improves listeners' perception and understanding of a talker's speech [1]. Some people are even capable of lip reading, i.e., understanding what is being said by watching the lip movements without listening [2], [3].

It is thus a natural idea to design computational systems to leverage both the audio and visual cues toward better speech separation. In the past decade, this idea has been explored through a number of approaches. This includes non-deep methods [4]–[11] and deep approaches [12]–[14]. For non-deep methods, the main limitations are on the capabilities of learning from large datasets and generalization to different speakers. For deep approaches, many are limited to speaker-dependent settings [15] or cannot easily generalize over speech mixtures of different numbers of speakers [12], [14]. We will elaborate this in detail in Section II-B.

In this paper, we propose a system with a novel Audio-Visual Deep Clustering (AVDC) model for leveraging both audio and visual cues to solve the *cocktail party problem*. As shown in Figure 1, the proposed system takes both audio and visual data (optical flow and gray images) as inputs for spectrogram mask prediction. Separated audio waveforms can be obtained by multiplying the predicted masks with the complex spectrogram of the input mixture, followed by an inverse Short-Time Fourier Transform (STFT). As further explained in Figure 3, the core AVDC model performs a two-stage fusion of audio and visual information to compute the final audio-visual Time-Frequency (T-F) embeddings for clustering T-F bins for spectrogram mask prediction. The first stage extracts audio-visual features for each speaker from their lip movements and the audio mixture in each time frame. The second stage computes speaker-wise audio-visual embeddings from the audio-visual features, and then concatenates them with audio embeddings calculated by deep clustering [16], [17] from the audio mixture. The concatenated embeddings are the final audio-visual embeddings for clustering.

We carry out experiments on two publicly available audio-visual speech datasets. Results show that the proposed AVDC model outperforms the deep clustering [16], [17] and utterance-level Permutation Invariant Training (uPIT) [18] audio-only separation baselines and three state-of-the-art audio-visual

Manuscript received January 26, 2019; revised May 15, 2019 and June 27, 2019; accepted July 8, 2019. Date of publication July 15, 2019; date of current version August 1, 2019. This work was supported in part by the NSFC under Grants 61876095 and 61751308 and in part by the NSF under Grant 1741472. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alexey Ozerov. *(Corresponding author: Rui Lu.)*

R. Lu and C. Zhang are with the Institute for Artificial Intelligence, with, State Key Lab of Intelligent Technologies and Systems, with Beijing National Research Center for Information Science and Technology (BNRist), and also with Department of Automation,Tsinghua University, Beijing 100084, China (e-mail: lur13@mails.tsinghua.edu.cn; zcs@mail.tsinghua.edu.cn).

Z. Duan is with the Department of Electrical and Computer Engineering, and also with Department of Computer Science, University of Rochester, Rochester, NY 14627 USA (e-mail: zhiyao.duan@rochester.edu).

Digital Object Identifier 10.1109/TASLP.2019.2928140

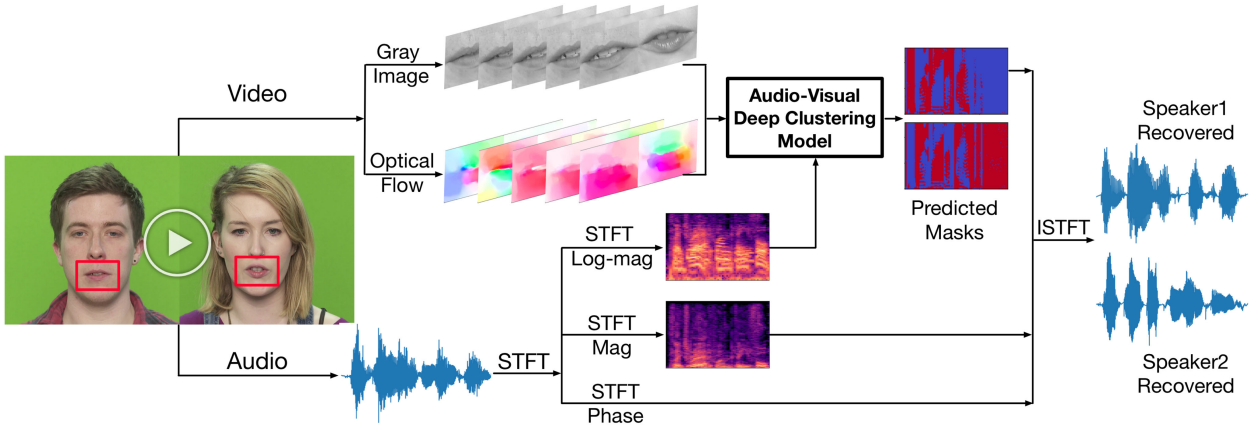


Fig. 1. Proposed audio-visual separation system: The Audio-Visual Deep Clustering (AVDC) model integrates both appearance (gray image) and motion (optical flow) information of the lip regions with the log magnitude spectrogram of the audio mixture to predict masks of source spectrograms. These masks are then applied to the complex spectrogram of the audio mixture followed by inverse-STFT for separation.

baselines [12], [14], [15]. Detailed analyses show that the integration of visual information greatly alleviates the source permutation problem. Further experiments also show the system's generalization ability across different numbers of speakers between training and testing and its robustness on partially observed videos. Ablation studies and visualization of the audio-visual embeddings provide further insights.

Our contributions in this paper come in threefold: First, the proposed AVDC model employs a novel two-stage audio-visual fusion strategy for speech separation, outperforming single-stage fusion in the experiments. Second, the AVDC model learns audio-visual "speaker-wise T-F embeddings". During separation, each speaker's Time-Frequency (T-F) embedding is shown to be able to separate the speaker from the other speakers. This explains how AVDC alleviates the source permutation problem that frequently happens for audio-only methods on same-gender speech mixtures. Third, the proposed model can generalize over speech mixtures of different numbers of speakers and shows a certain degree of robustness on partially observed videos.

The remainder of the paper is organized as follows: We first provide a comprehensive review of audio-based and audio-visual methods for speech separation in Section II. We then describe the proposed AVDC model in Section III. In Section IV, we compare our proposed method with state-of-the-art audio-only and audio-visual methods through experiments, and quantitatively and qualitatively verify the effectiveness, generalization ability and robustness of the proposed method in various experimental conditions. Finally, we conclude the paper in Section V.

II. Related Work on Single-Channel Speech Separation

A. Audio-Based Methods

Most speech separation methods only take the audio modality into consideration. Traditional methods mainly include Computational Auditory Scene Analysis (CASA) [19], [20], Hidden Markov Models (HMM) [21]–[23], and Non-negative Matrix Factorization (NMF) [4], [24]–[27]. CASA-based methods group T-F units of different speakers according to a variety of grouping cues including harmonicity, common modulation of pitch and amplitude, and timbre consistency to achieve separation. How to effectively fuse and balance different cues for different audio mixtures remains a challenging problem. HMM-based methods [21]–[23] typically use a factorial structure, i.e., one Markov chain for each speaker, to model the speaker-wise temporal dynamics and inter-speaker acoustic fusion. This factorial structure grows the state space exponentially with the number of simultaneous speakers, limiting the approach from well performing on mixtures with three or more speakers. Most NMF-based methods decompose the magnitude spectrogram of the audio mixture into speaker-specific spectral dictionaries and their temporal activations, where the dictionaries are usually pre-learned on clean speech of all (*supervised* separation) [25] or part (*semi-supervised* separation) of the speakers [28]. When sparsity penalties are introduced [27], NMF based models are also able to deal with noisy training corpus. However, NMF models are linear and lack the ability to learn more complex patterns from large amounts of training data. Moreover, most of them assume the availability of clean training speech for part or all of the speakers in the mixture, which cannot be satisfied in many speaker-independent scenarios in practice.

In recent years, deep learning techniques have been introduced to dramatically improve the separation performance [16]–[18], [29]–[32], thanks to their strong feature learning and data fitting capabilities. Many deep learning methods treat speech separation as a spectrogram mask estimation problem. Regression-based methods for mask estimation [29]–[31], [33] work well in speaker-dependent scenarios. However, they fail to operate properly in speaker-independent scenarios when clean speech training data is unavailable for target speakers. A typical error they make is the *source permutation problem*: The system can well separate the mixture into source signals at each time frame, but their assignments to sources over time are inconsistent. Estimating a consistent assignment is also called *sequential grouping* or *streaming* in Auditory Scene Analysis (ASA) and CASA [19].

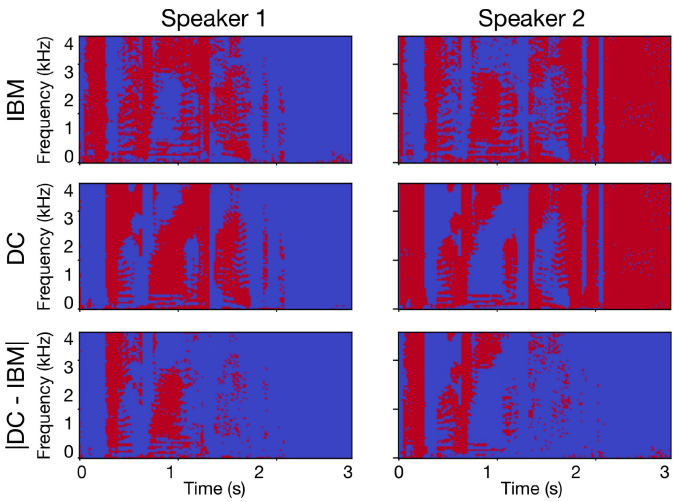


Fig. 2. Source permutation problem in separating a mixture of two male speakers. First row: The Ideal Binary Masks (IBM) of two speakers. Second row: Masks predicted by the audio-only Deep Clustering (DC) method [16]. Third row: Differences between the DC predicted masks and the IBM. The first second of each predicted mask is permuted to the wrong speaker.

Deep Clustering (DC) [16], [17], [32], [34]–[36] and utterance-level Permutation Invariant Training (uPIT) [18] methods are the state-of-the-art for dealing with the source permutation problem. The DC method projects each T-F bin of the mixture's spectrogram into a high-dimensional space where embeddings of the same speaker can be clustered to form its separation mask. The uPIT algorithm recovers speaker masks by minimizing the separation loss at the utterance level; this greatly reduces permutation errors in the original PIT algorithm [37], whose loss is defined at the frame level.

Despite the advantages of DC and uPIT over other approaches, they still suffer from the permutation problem. An example is shown in Figure 2, where the second row are predicted masks from the DC method on a mixture of two male speakers. Comparing to the Ideal Binary Masks (IBM) in the first row, we can see that there is a severe permutation problem in the first second of the separated masks. This problem frequently occurs when the two speakers are of the same gender, since they have similar vocal characteristics which are difficult to distinguish by audio-only method.

B. Audio–Visual Methods

Humans employ various cues at a cocktail party to pay attention to a speaker's voice. These cues include not only the consistency and continuity of the timber of speakers, but also the activity and location information of the speakers. In particular, visual information such as lip movement provides fine activity information about speakers. The correspondence between lip movements and speech utterances tells us which speakers are talking as well as what are being said [38], [39].

This idea of audio-visual speech separation has been explored by many non-deep methods [6]–[11], [40]. Casanovas *et al.* [8] performed audio-visual source separation using sparse representations. This method first clusters video atoms and then assigns the audio atoms to different speakers based on the correlation between audio and visual modalities. The method in [9] fuses the audio-visual dictionary learning with T-F masking, which not only renders an effective global representation but also captures the local information within the signals. Methods in [6], [7], [10], [11] establish audio-visual coherence through statistical modelling and recover audio sources by maximizing the learned coherence given visual representations. These non-deep algorithms, however, have limited capability to learn from large amounts of data and the generalization ability to different speakers is also limited.

Deep approaches for audio-visual speech enhancement and separation have been proposed in recent years [12]–[15], [41]–[44]. The basic intuition of audio-visual enhancement is to use visible lip region movements to isolate the target speaker's voice from background noise, either by feeding the combined audio-visual features into an enhancement network [15], [41], [42] or deriving the recovered clean audio from the speaker's corresponding visual information [13]. Audio-visual separation methods mainly exploit the audio-visual fusion strategy [12] or learn an audio-visual correspondence to assist the separation process [14]. The "Noise-Invariant" model [15] trains a speaker-dependent speech enhancement network, and its generalization ability to different speakers is limited. Although the "Noise-Invariant" model was initially designed for speech enhancement, it can be implemented in separation tasks, we thus exploit this model as one of our audio-visual separation baselines. The speaker-independent separation model "Look-to-Listen" [12] exploits an audio-visual fusion strategy and shows a significant improvement over the "Noise-Invariant" model, not only in terms of the separation performance, but also the generalization ability. Nevertheless, training schemes of the "Look-to-Listen" model for different numbers of speakers cannot be naturally shared. The "AV-Match" method [14] exploits the underlying correspondence between lip movements and acoustic speech signals to correct separation results of an audio-only model. By assigning snippets of the separated signal to the correct sources over time, this method relieves the permutation problem by a large margin for 2-speaker mixtures. However, the "AV-Match" method cannot be easily generalized to multi-speaker mixtures due to the exponential increase of possible permutations. The "Sound-of-Pixel" [44] model uses an audio analysis network to separate the mixture audio spectrogram into components, and uses a visual analysis network to compute visual embeddings for pixels in the visual scene. When applied to source separation, an aggregated visual embedding is calculated through spatial pooling of pixel-level embeddings of each source, before weighing and combining the audio components. Implicit audio-visual fusion happens after audio mixture is separated into components. This model successfully learns the correspondence between the appearance and sounds of various musical instruments. However, when the multiple audio sources share similar timbre characteristics, this method would not work well, since it does not help address the source permutation problem.

Most of these methods assume the synchronization between audio and visual streams. To relax this requirement, multiple instance learning has been adopted [43]. Although dealing with

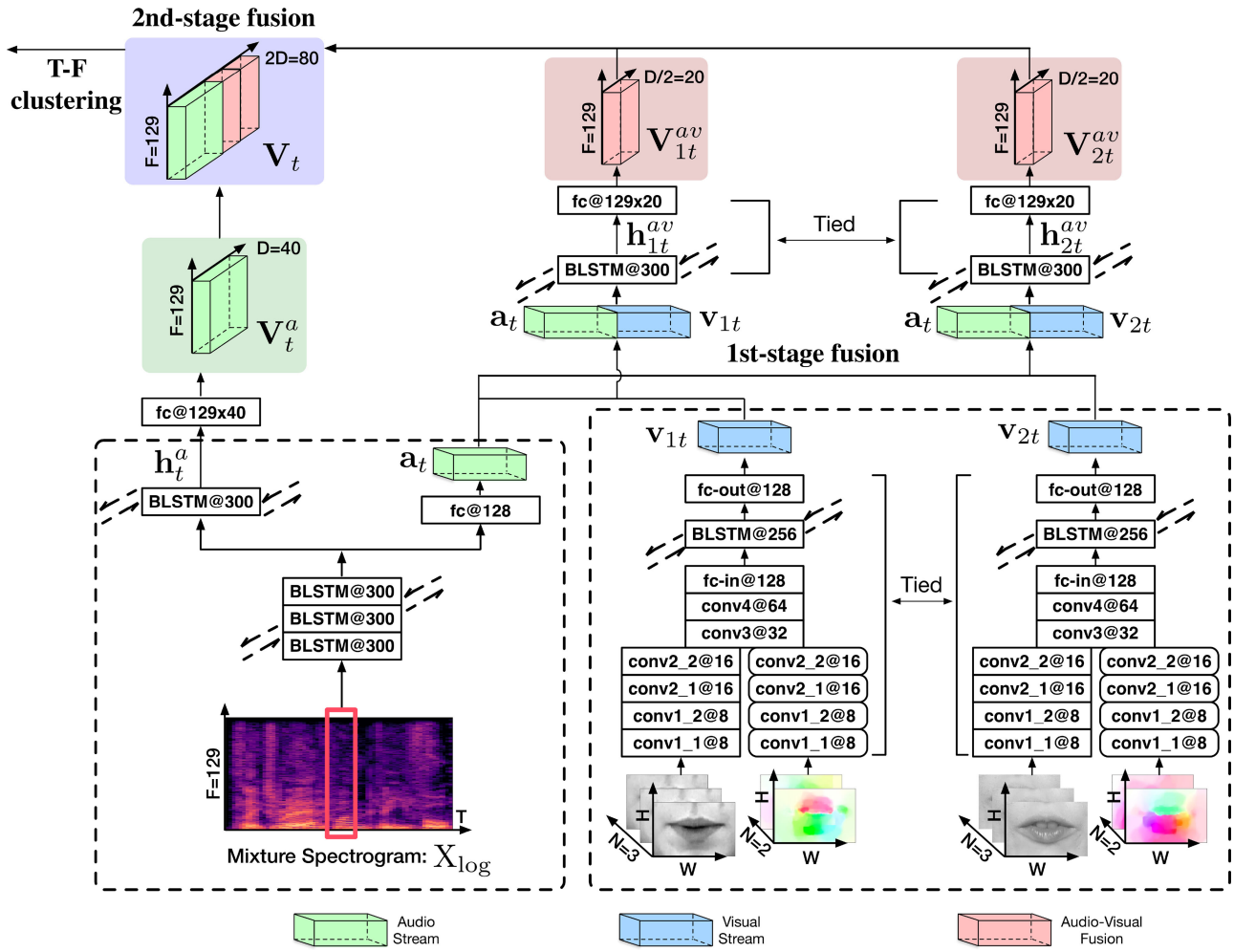


Fig. 3. Audio-Visual Deep Clustering (AVDC) model with a two-stage fusion strategy. First stage: The audio mixture and visual signals of each speaker are first encoded as frame-wise features through the audio-only networks (the left dashed box) and visual-only networks (the right dashed box), respectively. The audio features and visual features are then concatenated into audio-visual features for each speaker. Second stage: Audio-visual features of each speaker go through an embedding network to calculate speaker-wise audio-visual T-F embeddings ($\mathbf{V}_{1t}^{av}$ and $\mathbf{V}_{2t}^{av}$). The audio mixture also goes through an embedding network for audio-embeddings ($\mathbf{V}_t^a$). All such embeddings are concatenated into the final audio-visual embedding ($\mathbf{V}_t$) for clustering T-F bins during separation.

the unsynchronization between audio and visual information is meaningful for practical applications, our proposed model focuses on the audio-visual fusion strategy for speech separation. Therefore, we assume the synchronization between audio and visual streams as most other existing work does.

III. Method

As shown in Figure 1, the proposed system takes both the audio (STFT log-magnitude spectrogram) and visual (gray image and optical flow vectors of lip regions) modalities and integrates them through the "Audio-Visual Deep Clustering" (AVDC) model. This model predicts T-F masks for the speakers, and the masks are finally used to reconstruct source signals using the speech mixture's magnitude and phase spectrograms. The core part of this system is the "Audio-Visual Deep Clustering" model, which is illustrated in Figure 3. It adopts a two-stage fusion strategy to integrate the audio and visual modalities. The first-stage fusion computes speaker-wise audio-visual T-F embeddings for each speaker in the mixture, while the second-stage fusion concatenates these audio-visual embeddings with the audio-only embedding computed using an audio-only Deep Clustering (DC) method in Section III-A for the final clustering of T-F bins. This architecture will be detailed in Section III-B. As DC is a part of this architecture, we first describe DC in Section III-A.

A. Audio-Only Separation Model: Deep Clustering

Deep Clustering (DC) [16], [17] is one of the state-of-the-art audio-only speech separation methods. It aims to learn an embedding vector for each T-F bin of the speech mixture, based on which clustering can be performed to estimate a binary T-F mask for the speakers. Input to the DC model is the log-amplitude spectrogram of the speech mixture, $\mathbf{X}_{\log} \in \mathbb{R}^{F \times T}$, where F is the number of frequency bins and T is the number of time frames. DC then outputs a D-dimensional embedding vector for each T-F bin, constituting an embedding matrix $\mathbf{V} \in \mathbb{R}^{(F \times T) \times D}$.

Throughout this paper, our supervision information are the Ideal Binary Mask (IBM), $\mathbf{Y} \in \mathbb{R}^{(F\times T)\times C}$ [16], where C is the number of speakers in the mixture. Let $Y_{i,j} = 1$ when source j dominates T-F bin $i = (t, f)$, otherwise let $Y_{i,j} = 0$. Training the DC model is to learn the network weights that adjust the embedding matrix $\mathbf{V}$ to approximate the affinity matrix of the IBM:

$$L_{DC} = \left|\left|VV^T - YY^T\right|\right|_F^2. \tag{1}$$

During the testing phase, the speech mixture is first passed through the learned network to calculate the embedding vectors for all T-F bins. K-means clustering is then performed to cluster the embedding vectors to estimate a binary T-F mask for each speaker.

B. Audio–Visual Deep Clustering

The Audio-Visual Deep Clustering (AVDC) model is the core of the proposed audio-visual speech separation approach. It uses a two-stage fusion strategy to integrate audio and visual information extracted by the audio networks and visual networks described below.

1) Audio Networks: The left dashed box in Figure 3 illustrates the audio networks. Given the log-spectrogram of the mixture $\mathbf{X}_{\log} \in \mathbb{R}^{F\times T}$, we feed each frame to the first three Bidirectional Long Short-Term Memory (`BLSTM`) layers of the audio network for feature extraction. Each `BLSTM` layer has a hidden size of 300. Outputs of the stacked `BLSTM` layers act as shared features which are fed into the following two subnetworks. One subnetwork is an Fully-Connected (`fc`) layer that outputs features of the audio stream: $\mathbf{a}_t \in \mathbb{R}^{128}$ $(t = 1, ..., T)$. The other subnetwork is a `BLSTM` layer of which the output is denoted as $\mathbf{h}_t^a \in \mathbb{R}^{600}$, which will be used to calculate the audio-only embedding in Eq. 3.

2) Visual Networks: The right dashed box in Figure 3 illustrates the visual networks. We exploit the lip regions' gray images and optical flow as inputs to our visual network. Every N consecutive gray images centered at the current frame are stacked to form an input patch of size $H \times W \times N$, where H and W are height and width of the lip region image in pixels. Similarly, optical flow data are of size $H \times W \times 2$, where the two channels represent horizontal and vertical motions.

Inspired by other audio-visual models [39], [45]–[47], we adopt the VGG-style [48] Convolutional Neural Network (CNN) as the feed-forward part of our visual network. In early layers, we have two separate CNNs (`conv1_1` $-$ `conv2_2`) to process the gray image and the optical flow respectively. `conv1_1` and `conv1_2` have 8 filters of size 3×3 and stride 1, followed by a `max` $-$ `pooling` layer with kernel size 2 and stride 2. `conv2_1` and `conv2_2` have the same hyperparameters as the previous set except that they have 16 filters. Output features of the early layers from both streams are concatenated to go through further convolutional layers (`conv3` and `conv4`). `conv3` layer has 32 filters of size 3×3 and stride 1, followed by a `max` $-$ `pooling` layer with kernel size 2 and stride 2, while `conv4` layer has 64 filters. Every convolutional layer is followed by batch-normalization [49] and ReLU non-linearity. It is noted that such a "separate-merge strategy" is effective in human action recognition [50], [51] and also works best in our task.

Following the convolutional layers, the `fc` $-$ `in` layer is an `fc` layer with output dimension of 128 and ReLU non-linearity. Finally, we add a `BLSTM` layer on top of the feed-forward networks to model the long-term temporal evolution of visual information. The `BLSTM` layer has a hidden size of 256. Its output is fed to another fully-connected layer `fc` $-$ `out` to obtain the 128-d frame-wise visual feature vectors $\mathbf{v}_{it}$ $(t = 1, ..., \lceil T/r \rceil, i = 1, ..., C)$, where r is the ratio between audio sampling rate and and video sampling rate.

It is worth to mention that the weights of all the feed-forward parts of our visual network (all the `conv` layers, the `fc` $-$ `in` layer and the `fc` $-$ `out` layer) are tied not only across each frame but also for different speakers, as shown in Figure 3. This reduces the complexity of the architecture and makes the feed-forward network act as a general visual feature extractor to capture features of the mouth regions [38].

3) First-Stage Audio–Visual Fusion: The proposed audio-visual fusion differs from previous works [12], [15] in that we carry out a two-stage fusion strategy. The first stage is that we concatenate $\mathbf{a}_t$ with the aligned $\mathbf{v}_{it}$ to form the audio-visual features as inputs to another `BLSTM` layer:

$$\mathbf{h}_{it}^{av} = \text{BLSTM}(\text{concate}[\mathbf{a}_t, \mathbf{v}_{it}]), \tag{2}$$

where $\mathbf{h}_{it}^{av} \in \mathbb{R}^{600}$ for $i = 1, ..., C$. This `BLSTM` network is expected to capture the audio-visual correspondence between components of the audio mixture and the lip movements of each speaker.

4) Second-Stage Audio–Visual Fusion: The second stage of audio-visual fusion is the concatenation of the audio embedding and audio-visual embeddings of all speakers. The audio embedding $\mathbf{V}_t^a \in \mathbb{R}^{F\times D}$ is calculated by passing $\mathbf{h}_t^a$ through an `fc` layer with an output dimension of $F \times D$, same as DC:

$$\mathbf{V}_t^a = \text{FC}_a(\mathbf{h}_t^a). \tag{3}$$

The audio-visual embedding $\mathbf{V}_{it}^{av} \in \mathbb{R}^{F\times(D/2)}$ is obtained by passing $\mathbf{h}_{it}^{av}$ through an `fc` layer with an output dimension of $F \times (D/2)$ for each speaker i:

$$\mathbf{V}_{it}^{av} = \text{FC}_{av}(\mathbf{h}_{it}^{av}). \tag{4}$$

Finally, the audio embedding and audio-visual embeddings of all speakers are concatenated to form the final embedding $\mathbf{V}_t \in \mathbb{R}^{F\times(D+C\times D/2)}$ at frame t for T-F bin clustering:

$$\mathbf{V}_t = \text{concate}[\mathbf{V}_t^a, \mathbf{V}_{1t}^{av}, ..., \mathbf{V}_{Ct}^{av}]. \tag{5}$$

Concatenating all embeddings across time, the final embedding matrix can be re-arranged to $\mathbf{V} \in \mathbb{R}^{(T\times F)\times(D+C\times D/2)}$. In other words, each T-F bin is assigned an embedding vector of the dimension of $D + C \times D/2$, where the audio embedding takes D dimensions and the audio-visual embedding for each speaker takes $D/2$ dimensions. K-means clustering is used to group the T-F bins for separation. During clustering, it is expected that the audio embedding dimensions behave similarly to DC, while the audio-visual embedding dimensions for each speaker help to differentiate T-F bins dominated by that speaker versus all the other speakers.

Algorithm 1: k-POD.

Input: Partially Observed Embedding Matrix $\mathbf{V}$
Output: T-F bin assignments $\hat{\mathbf{Y}}$
1: Initialize $\hat{\mathbf{Y}}^{(0)}$ and cluster centroids $\mathbf{B}^{(0)}$
2: **repeat**
3: $\quad \mathbf{V}^{(m)} \leftarrow \mathcal{P}_{\Omega}(\mathbf{V}) + \mathcal{P}_{\Omega^c}(\hat{\mathbf{Y}}^{(m)}\mathbf{B}^{(m)})$
4: $\quad (\hat{\mathbf{Y}}^{(m+1)}, \mathbf{B}^{(m+1)}) \leftarrow$ k-means$(\mathbf{V}^{(m)})$
5: **until** convergence

It can be seen that this audio-visual fusion model can be trained and tested on mixtures of different number of speakers, showing superiority over the "Look-to-Listen" model [12], of which the network structure and training scheme is limited to train and test on the same number of speakers' mixtures.

C. Partially Observed Video

In practice, visual information of certain speakers may be missing, due to occlusion and the camera scope. In this case, audio-visual features in Eq. (2) and embeddings in Eq. (4) for those speakers will be missing. To perform k-means algorithm on $\mathbf{V}$ with missing dimensions, we apply the k-POD [52] algorithm detailed in Algorithm III-C. Let $\Omega \subset \{1, ..., T \times F\} \times \{1, ..., d\}$ $(d = D + C \times (D/2))$ be a subset of indexes that correspond to the observed entries of embedding matrix $\mathbf{V}$. The projection operator of matrix $\mathbf{V}$ onto an index subset Ω is given by:

$$[\mathcal{P}_{\Omega}(\mathbf{V})]_{ij} = \begin{cases} \mathbf{V}_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{if } (i,j) \in \Omega^c \end{cases}, \tag{6}$$

where Ω^c is the complement of Ω. In Algorithm III-C, $\hat{\mathbf{Y}} \in \mathbb{R}^{(T\times F)\times C}$ is the recovered mask for T-F bins and rows of $\mathbf{B} \in \mathbb{R}^{C\times d}$ are the centroids of the clusters. $\hat{\mathbf{Y}}^{(m)}\mathbf{B}^{(m)}$ in Algorithm III-C indicates the matrix multiplication, with the resulted matrix of size $\mathbb{R}^{(T\times F)\times d}$ and each row of which is filled with the corresponding k-means centroids. To sum up, this algorithm iteratively fills the missing values of rows of $\mathbf{V}$ and then perform k-means clustering until convergence. We exploit this algorithm to verify the robustness of our proposed model when videos of certain speakers are missing.

IV. Experiments

In this section, we first carry out experiments on 2-speaker and 3-speaker mixtures to illustrate the performance improvement brought by our proposed AVDC model compared with previous state-of-the-art methods. We also train and test our model in unmatched conditions (i.e., the training and testing mixtures have different number of speakers) to verify its flexibility in generalization. We also quantitatively and qualitatively show the benefits of the proposed two-stage audio-visual fusion strategy in obtaining high performance for both same-gender mixtures and different-gender mixtures. The final experiment shows robustness of AVDC model against partially missing visual information in the videos.

A. Datasets

We make use of the 2-speaker mixtures and 3-speaker mixtures of the GRID [53] and TCD-TIMIT [54] datasets to illustrate the effectiveness of our proposed method.

1) GRID Dataset: The GRID [53] dataset provides both audio and visual data useful for audio-visual speech separation research [13], [15]. This dataset contains 34 speakers and each of them has 1000 frontal face video recordings. Each video has a duration of 3 seconds with a frame rate of 25 frames per second (FPS). We use its high-quality version with an image resolution of 720×576 pixels (height $\times$ width). The corpus has a relatively small vocabulary since it contains only 51 different words. In our experiments, we exclude Speakers 's8' and 's21', as many files of 's8' are not well recorded (these files only contain around 0.3 seconds videos) and there is no video data for 's21'. This results in a total of 17 male speakers and 15 female speakers. We randomly select 3 males and 3 females to construct a validation set of 2.5 hours and another 3 males and females for a test set of 2.5 hours. The rest of the speakers form the training set of 30 hours. To construct a 2-speaker mixture, we randomly choose two different speakers, mix their audio at a random Signal-to-Noise Ratio (SNR) between 0 dB and 5 dB as in [16], [17] and put their videos side by side. It is worth to mention that we balance the amount of same-gender mixtures and different-gender mixtures during the data generation process.

2) TCD-TIMIT Dataset: TCD-TIMIT is another audio-visual speech dataset [54] commonly used in audio-visual speech separation research [12], [15]. We use the frontal-face videos from the 59 volunteers in the dataset, which are comprised of 32 males and 27 females. Each speaker reads 98 sentences from the TIMIT [55] corpus, resulting in durations of around 5 seconds for each video. All of the videos have a resolution of 1920×1080 pixels with a frame rate of 29.97 FPS. We randomly select 6 males and 5 females to construct a validation set of 2.5 hours and another 6 males and 5 females for a test set of 2.5 hours. The rest of the speakers form the training set of 30 hours. The generation process is similar to that of the GRID dataset.

3) Data Preprocessing: All audio recordings are downsampled to 8 kHz to compute the Short Time Fourier Transform (STFT) with a window size of 32 ms and a hop size of 8 ms without zero padding. This gives us $F = 129$ frequency bins in the spectrogram in Section III. The log-amplitude spectrogram $\mathbf{X}_{\log}$ is used as the input to the AVDC model as shown in Figure 1. We build the IBM ($\mathbf{Y}$) by setting the mask value of the dominant source to 1 in each T-F bin. This data preprocessing process is similar to those in [16], [34].

Mouth areas in the videos are detected with the Dlib library [56]: For the GRID dataset, mouth areas have a size of 80×120 while for TCD-TIMIT dataset, it is 150×225 $(H \times W)$. We employ the Python implementation[1] of the Coarse2Fine [57], [58] algorithm for optical flow extraction. For gray images, we find that the stack of 3 consecutive images $(N = 3)$ achieves the best performance. We set $D = 40$ for the T-F embedding

[1] https://github.com/pathak22/pyflow

matrices $\mathbf{V}$, $\mathbf{V}^a$ and $\mathbf{V}_i^{av}$. We normalize each log-amplitude spectrogram to zero mean and unit standard deviation, where the mean and standard deviation are computed across all the T-F bins of the log-amplitude spectrograms of the training data.

For gray images, the pixel intensities are first rescaled to the range of 0 to 1 [41]. We then compute the mean and std for each pixel across all the training data independently, resulting in a mean and an std image of the same size with the gray images [47]. Gray images are then normalized with the computed mean and std. Since the optical flow data have two channels, we compute the mean and std for each channel across all the training data, resulting in the global mean and std with size $1 \times 1 \times 2$, representing the statistics of vertical and horizontal axis, such mean and std for optical flow data perform the best in our model. Both channels of the optical flow data are then normalized to zero mean and unit std.

B. Baselines and Evaluation Measures

For the "audio-only" baselines, we use the DC model [16] and the uPIT model [18], which are the state of the arts. We use our own implementations and follow [17] and [18] to exploit their optimal structures. The finally adopted structure of the DC model is comprised of four `BLSTM` layers, each having a hidden size of 300 while that of the uPIT model is comprised of three `BLSTM` layers, each having a hidden size of 896 units. All models are trained by the Adam algorithm [59] with a learning rate of $\lambda = 0.001$, we stop training if the loss on the validation set does not decrease for 5 epochs in succession.

For audio-visual baselines, we compare with our previous approach AV-Match [14], which was designed to fix source permutation errors in audio-only separation using audio-visual match as post-processing. Since this method can not be easily generalized across different numbers of speakers, we only report its performance on the 2-speaker case. We also compare with two other state-of-the-art audio-visual methods, Look-to-Listen [12] and Noise-Invariant [15], however, since we have no access to their code, we simply use their reported results in the 2-speaker case on the TCD-TIMIT dataset.

For evaluation measures, we use delta Signal-to-Distortion Ratio (ΔSDR), Signal-to-Distortion Ratio (SDR), Signal-to-Artifacts Ratio (SAR) and Signal-to-Interference Ratio (SIR) [60]. We conduct 5-fold cross validation for each experiment to measure the separation performance.

C. Hyper-Parameter Tuning

We carry out experiments on the GRID dataset to discuss how the hyper-parameters affect the performance of our proposed AVDC model. As shown in Tables I and II, we report the mean ΔSDR on different kinds of mixtures in both 2-speaker and 3-speaker settings. "Def" indicates the default assignment of the separated audio frames while "Opt" indicates that we apply the optimal permutation on the outputs of our model.

In Table I, we inspect how the dimensions of our audio and visual vectors in the first fusion stage ($\mathbf{a}_t$ and $\mathbf{v}_{it}$) affect the separation performance. We vary this dimension Dim from 64, 128, 256, to 512, and find that Dim $= 128$ outperforms other settings by a large margin on different-gender mixtures. We thus set Dim $= 128$ in all our following experiments.

TABLE I
ΔSDR (dB) OF THE AVDC MODEL ON THE GRID DATASET WITH DIFFERENT AUDIO AND VISUAL VECTOR DIMENSIONS BEFORE THE FIRST-STAGE FUSION

Mixtures	Dim	Same-Gender		Diff-Gender		Overall	
		Def	Opt	Def	Opt	Def	Opt
2-SPK Mixtures	64	7.37	7.50	8.21	8.24	7.73	7.82
	128	7.35	7.57	10.16	10.11	**8.88**	**8.95**
	256	7.65	7.74	8.56	8.72	8.21	8.33
	512	7.61	7.69	8.25	8.30	7.88	7.95
3-SPK Mixtures	64	2.91	3.71	4.56	5.21	4.43	4.96
	128	3.32	4.30	5.50	5.83	**4.56**	**4.99**
	256	3.88	4.26	4.53	4.83	4.46	4.77
	512	3.57	4.44	3.53	4.07	3.53	4.20

TABLE II
ΔSDR (dB) OF THE AVDC MODEL ON THE GRID DATASET WITH DIFFERENT T-F EMBEDDING DIMENSIONS

Mixtures	Setting	Same-Gender		Diff-Gender		Overall	
		Def	Opt	Def	Opt	Def	Opt
2-SPK Mixtures	D/2 = 10	7.15	7.26	7.59	7.65	7.34	7.43
	D/2 = 20	7.35	7.57	10.16	10.11	**8.88**	**8.95**
	D/2 = 30	7.97	8.00	8.35	8.31	8.13	8.12
3-SPK Mixtures	D/2 = 10	3.71	4.13	4.14	4.59	4.10	4.55
	D/2 = 20	3.32	4.30	5.50	5.83	**4.56**	**4.99**
	D/2 = 30	3.40	4.03	4.15	4.73	4.08	4.57

In Table II, we compare the separation performance of the AVDC model with different audio-visual T-F embedding dimensions. This hyper-parameter mainly affects the second-fusion stage. D / 2 $= 20$ performs the best among all the settings, we thus use it in all our following experiments.

It is interesting to notice that the optimal hyper-parameters are changing depending on the same-gender or different-gender condition, as observed in Tables I and II. One possible reason is that patterns of different-gender mixtures are more complicated than those of same-gender mixtures, thus given the fixed amount of training data, it is easier to encounter the overfitting problem when we increase the model complexity.

D. Separation Results on 2-Speaker Mixtures

We compare the separation performance of the proposed Audio-Visual Deep Clustering (AVDC) method with baselines in Table III. In addition to the overall ΔSDR, we report results on Female-Female (F-F) mixtures, Male-Male (M-M) mixtures and Female-Male (F-M) mixtures separately. Methods with a subscript "spk3$''$ report results of the model trained on 3-speaker mixtures and evaluated on 2-speaker mixtures. This allows us to assess the generalization ability to different number of speakers of these methods. Methods with a superscript $\star$ show results when the optimal source permutation is used: For each time frame t, we select the permutation that has the highest correlation between the ground-truth mask $\mathbf{Y}_t$ (IBM) and the predicted

TABLE III
COMPARISON OF SEPARATION RESULTS (MEAN±STD) ON 2-SPEAKER MIXTURES IN BOTH THE GRID AND TCD-TIMIT DATASETS. METHODS WITH A SUPERSCRIPT ⋆ SHOW RESULTS WHEN THE OPTIMAL SOURCE PERMUTATION IS USED. METHODS WITH A SUBSCRIPT "SPK3" SHOW RESULTS OF THE MODEL TRAINED ON 3-SPEAKER MIXTURES BUT EVALUATED ON 2-SPEAKER MIXTURES

Dataset	Method	ΔSDR (dB)				SDR (dB)	SIR (dB)	SAR (dB)
		F-F	M-M	F-M	overall	overall	overall	overall
GRID	DC [17]	5.14 ± 0.91	4.05 ± 1.27	9.92 ± 0.48	7.72 ± 0.44	8.13 ± 0.45	14.05 ± 0.53	10.24 ± 0.37
	uPIT [18]	4.92 ± 0.60	4.68 ± 0.41	9.56 ± 0.57	7.60 ± 0.24	8.01 ± 0.24	14.15 ± 0.26	10.09 ± 0.36
	AV-Match [14]	6.96 ± 0.49	4.93 ± 0.95	9.80 ± 0.47	8.11 ± 0.45	8.52 ± 0.46	14.66 ± 0.48	10.36 ± 0.50
	AVDC	$\mathbf{8.32 \pm 0.54}$	$\mathbf{6.18 \pm 0.67}$	$\mathbf{10.16 \pm 0.49}$	$\mathbf{8.88 \pm 0.42}$	$\mathbf{9.30 \pm 0.43}$	$\mathbf{15.60 \pm 0.44}$	$\mathbf{11.06 \pm 0.38}$
	DC_{spk3}	1.84 ± 0.70	1.64 ± 0.53	8.46 ± 0.68	5.69 ± 0.25	6.11 ± 0.26	11.63 ± 0.30	8.81 ± 0.21
	$AVDC_{spk3}$	4.39 ± 0.98	4.08 ± 0.51	7.95 ± 0.80	6.41 ± 0.61	6.83 ± 0.61	12.76 ± 0.68	9.11 ± 0.47
	DC⋆	7.24 ± 0.51	5.30 ± 0.98	9.83 ± 0.43	8.30 ± 0.35	8.72 ± 0.36	14.87 ± 0.36	10.49 ± 0.33
	uPIT⋆	5.87 ± 0.50	4.79 ± 0.58	9.77 ± 0.46	7.82 ± 0.31	8.21 ± 0.32	14.30 ± 0.34	10.31 ± 0.32
	AVDC⋆	$\mathbf{8.52 \pm 0.46}$	$\mathbf{6.43 \pm 0.59}$	$\mathbf{10.11 \pm 0.47}$	$\mathbf{8.95 \pm 0.35}$	$\mathbf{9.37 \pm 0.35}$	$\mathbf{15.70 \pm 0.34}$	$\mathbf{11.04 \pm 0.34}$
	$DC^{\star}_{spk3}$	4.87 ± 0.77	3.69 ± 0.56	8.39 ± 0.55	6.64 ± 0.26	7.06 ± 0.26	12.85 ± 0.33	9.11 ± 0.24
	$AVDC^{\star}_{spk3}$	4.88 ± 0.85	4.30 ± 0.33	7.84 ± 0.74	6.47 ± 0.52	6.89 ± 0.52	12.75 ± 0.62	8.97 ± 0.43
	IBM	12.61 ± 0.83	11.28 ± 0.35	12.76 ± 0.43	12.39 ± 0.43	12.80 ± 0.43	20.08 ± 0.48	13.90 ± 0.41
TCD-TIMIT	DC [17]	5.09 ± 0.67	3.51 ± 0.78	$\mathbf{9.98 \pm 0.53}$	6.32 ± 0.38	6.71 ± 0.39	12.78 ± 0.52	9.18 ± 0.25
	uPIT [18]	4.98 ± 0.27	4.03 ± 0.38	9.78 ± 0.32	6.37 ± 0.32	6.76 ± 0.32	13.07 ± 0.24	9.19 ± 0.36
	AVDC	$\mathbf{6.73 \pm 0.55}$	$\mathbf{5.49 \pm 0.97}$	9.91 ± 0.66	$\mathbf{7.47 \pm 0.35}$	$\mathbf{7.86 \pm 0.37}$	$\mathbf{14.45 \pm 0.40}$	$\mathbf{9.92 \pm 0.28}$
	DC_{spk3}	4.14 ± 0.51	3.25 ± 0.58	9.62 ± 0.51	5.81 ± 0.28	6.20 ± 0.30	12.24 ± 0.37	8.93 ± 0.14
	$AVDC_{spk3}$	5.12 ± 0.49	4.96 ± 0.64	8.30 ± 0.30	6.20 ± 0.26	6.60 ± 0.26	13.20 ± 0.28	9.04 ± 0.12
	DC⋆	6.92 ± 0.44	5.85 ± 0.61	$\mathbf{9.85 \pm 0.60}$	7.62 ± 0.45	8.00 ± 0.46	14.32 ± 0.59	9.85 ± 0.36
	uPIT⋆	6.03 ± 0.39	5.65 ± 0.58	9.66 ± 0.42	7.38 ± 0.46	7.79 ± 0.45	14.19 ± 0.61	9.69 ± 0.79
	AVDC⋆	$\mathbf{7.40 \pm 0.50}$	$\mathbf{6.38 \pm 0.65}$	9.78 ± 0.69	$\mathbf{7.92 \pm 0.46}$	$\mathbf{8.31 \pm 0.47}$	$\mathbf{14.83 \pm 0.55}$	$\mathbf{10.10 \pm 0.36}$
	$DC^{\star}_{spk3}$	6.20 ± 0.34	5.65 ± 0.60	9.63 ± 0.52	7.25 ± 0.38	7.65 ± 0.40	13.85 ± 0.49	9.60 ± 0.28
	$AVDC^{\star}_{spk3}$	5.79 ± 0.46	5.66 ± 0.73	8.19 ± 0.39	6.62 ± 0.31	7.02 ± 0.32	13.30 ± 0.42	9.10 ± 0.24
	IBM	15.97 ± 0.61	13.45 ± 0.67	15.01 ± 0.53	14.67 ± 0.39	15.06 ± 0.39	22.61 ± 0.43	16.11 ± 0.38

mask $\hat{\mathbf{Y}}_t$. This allows us to assess the relative significance of errors due to frame-wise separation or cross-frame source assignment.

Several interesting observations can be made from Table III. First, the proposed AVDC model outperforms the audio-based DC method and the uPIT method on both the GRID and TCD-TIMIT datasets. On the GRID dataset, AVDC achieves an improvement of 1.16 dB over DC on mean ΔSDR and 1.28 dB over uPIT; While on the TCD-TIMIT dataset, these improvements are 1.15 dB and 1.10 dB, respectively. By inspecting results on different types of mixtures, we can see that these improvements are more pronounced on same-gender mixtures: On GRID, the improvements over DC are 3.18 dB for F-F and 2.13 dB for M-M while those over uPIT are 3.4 dB for F-F and 1.5 dB for M-M; On TCD-TIMIT, the improvements over DC are 1.64 dB for F-F and 1.98 dB for M-M while those over uPIT are 1.75 dB for F-F and 1.46 dB for M-M. We suggest that this is because audio-only separation methods mainly leverage the differences of timbre information of speakers [16], [17], [37], which can be quite similar when they are of the same gender. In this case, not only sources are more difficult to separate in each frame, but also the separated sources in different frames are harder to assign to the correct source consistently (i.e., permutation error).

When visual cues are introduced, the permutation errors can be greatly reduced, as the correlation between lip movements and acoustic features provides a natural cue for the assignment of separated sources. Indeed, when the ground-truth permutation is used in DC, leading to DC⋆, a big improvement of ΔSDR can be observed across types of mixtures and datasets. The result suggests that the uPIT method can better handle the "source permutation" problem: the improvements of uPIT⋆ over uPIT on the GRID dataset are relatively small. However, when we test on the TCD-TIMIT dataset which has a more abundant vocabulary, the improvements of uPIT⋆ over uPIT are still significant. The claim that AVDC helps address the source permutation problem can be further verified by comparing AVDC and AVDC⋆, which uses the ground-truth source assignment in each time frame. We can see that AVDC⋆ only improves over AVDC slightly, suggesting that the source permutation problem is not a major issue anymore in AVDC.

When visual cues are introduced, separation errors not associated with permutation errors are also reduced. This can be seen by comparing AVDC⋆ with DC⋆ and uPIT⋆, where the ground-truth source assignment is used. We still see a significant improvement of mean ΔSDR when comparing AVDC⋆ with DC⋆ in same-gender mixtures (1.28 dB for F-F and 1.13 dB for M-M on GRID; 0.48 dB for F-F and 0.53 dB for M-M

on TCD-TIMIT). This suggests that the introduction of visual information and audio-visual matching also helps learning better feature embeddings for clustering. When we compare AVDC* with uPIT* on both datasets, we observe even more significant mean ΔSDR improvements (2.65 dB for F-F and 1.64 dB for M-M on GRID; 1.37 dB for F-F and 0.73 dB for M-M on TCD-TIMIT). This suggests that though the uPIT method has advantages over the DC method in solving the "source permutation" problem, it is not as good as DC in improving the frame-wise separation performance.

Second, the proposed AVDC approach significantly outperforms three other state-of-the-art audio-visual approaches. The AV-Match [14] method is designed to fix source permutation problems in DC, hence its performance is bounded by DC*. As described earlier, AVDC not only fixes the permutation problem, but also learns better embeddings for clustering. The Look-to-Listen model [12] and Noise-Invariant model [15] report an average SDR of 4.1 dB and 0.4 dB on the TCD-TIMIT dataset respectively, while the proposed AVDC achieves 7.86 dB. In fact, both baselines seem very weak on the TCD-TIMIT dataset, as they even underperform the audio-only baselines, DC and uPIT. Since we have no access to the implementation nor result details of these two methods, we simply copied their reported SDR results on TCD-TIMIT from [12].

To make sure our comparison with Look-to-Listen [12] and Noise-Invariant [15] methods is fair, we checked the experimental setup for their reported results. However, we could not find details on how their test data was generated, nor did we know whether cross validation was employed as ours. Therefore, there could be a difference on the distributions of speakers, genders and utterances between their test data and ours. This makes a direct comparison on the overall results less rigorous; one may need to compare their reported results with the different types of mixtures of our results to partially mitigate this issue. Nevertheless, it is worth to mention that the Noise-Invariant method is speaker-dependent and it requires training a dedicated model for each speaker [15]. This gives it a disadvantage over AVDC and Look-to-Listen. On the other hand, it is also noted that the Look-to-Listen [12] model was trained on 2000 hours of video clips, about two orders of magnitude higher than our training set.

There are three possible reasons that the Look-to-Listen [12] model performs worse than our AVDC model: 1) the training data for Look-to-Listen model (which does not include TCD-TIMIT) might be too different from the TCD-TIMIT test data. 2) the two-stage fusion strategy of the AVDC model has an advantage over the single-stage fusion of Look-to-Listen, in terms of exploiting a balance between the audio-only branch and the audio-visual branch for separating different kinds of mixtures. It is also worth to mention that the entire network architecture of our proposed AVDC model does not depend on the number of speakers, hence it is easier to generalize across different numbers of speakers compared to the Look-to-Listen model. 3) the proposed AVDC method is able to generalize from 3-speaker-mixture training to 2-speaker-mixture testing. Although the performance degrades much from AVDC to AVDC$_{\text{spk3}}$, by comparing DC$_{\text{spk3}}$ with AVDC$_{\text{spk3}}$, we see that the introduction of visual information still improves the performance in this unmatched condition. The uPIT [18] and AV-Match [14] baselines, however, cannot be easily generalized across different numbers of speakers. The Noise-Invariant [15] method is speaker-dependent, and the network design of Look-to-Listen [12] method is limited to train and test on the same number of speakers' mixtures.

Last but not least, there is still a big gap between the best results (mostly achieved by AVDC*) and IBM. This suggests that there is a large room for learning a better feature embedding for source separation.

E. Separation Results on 3-Speaker Mixtures

Besides the experiments on 2-speaker mixtures, we also compare our proposed AVDC model with the DC model and the uPIT model on 3-speaker mixtures in Table IV. In addition to the overall ΔSDR, we report results on Female-Female-Female (F3) mixtures, Female-Female-Male (F2-M1) mixtures, Female-Male-Male (F1-M2) mixtures and Male-Male-Male (M3) mixtures separately. Similar to Table III, methods with a superscript $\star$ show results of models using the ground-truth frame-wise source assignment. Methods with a subscript spk2 show results of methods trained on 2-speaker mixtures but evaluated on 3-speaker mixtures.

Comparing the two tables, we can see that 3-speaker separation is much more difficult than 2-speaker separation, as the performance of all methods in all conditions drop significantly. Nevertheless, the proposed AVDC approach outperforms DC and uPIT on both datasets and all of the evaluation metrics. On GRID, the improvement of AVDC over DC in terms of mean ΔSDR is 1.45 dB while that over uPIT is 0.82 dB; For TCD-TIMIT, the improvement over DC is 0.86 dB while that over uPIT is 0.47 dB. Similar to the case of 2-speaker mixtures, the performance improvement mainly comes from the same-gender mixtures (e.g., improvements over DC being 1.98 dB for F3 and 1.97 dB for M3 on GRID and 1.16 dB for F3 and 2.03 dB for M3 on TCD-TIMIT).

It is also worth to mention that the permutation problem in 3-speaker separation is more pronounced as the number of possible permutations is larger; making the advantages of AVDC better shown, since one of its advantages is alleviating permutation errors through audio-visual matching. This can be observed by comparing the performance gap between DC-DC*, uPIT-uPIT* and AVDC-AVDC*: On the GRID dataset, the gap between DC and DC* is 2.80 dB for F3 mixtures and 2.16 dB for M3 mixtures, the gap between uPIT and uPIT* is 1.61 dB for F3 mixtures and 1.03 dB for M3 mixtures, which verifies uPIT's advantage over DC in alleviating the permutation problem, while the gap between AVDC and AVDC* is further reduced to 0.71 dB and 0.54 dB, respectively.

On the other hand, when we compare AVDC* with DC*, we see a similar performance on the GRID dataset and a slight decrease on the TCD-TIMIT dataset. This suggests that the learned audio and audio-visual embeddings by AVDC are not better than the audio embeddings learned by DC for clustering. Reasons for this result are not clear yet. One possibility is that

TABLE IV
COMPARISON OF SEPARATION RESULTS (MEAN±STD) ON 3-SPEAKER MIXTURES IN BOTH THE GRID AND TCD-TIMIT DATASETS. METHODS WITH A SUPERSCRIPT ⋆ SHOW RESULTS WHEN THE OPTIMAL SOURCE PERMUTATION IS USED. METHODS WITH A SUBSCRIPT "SPK2" SHOW RESULTS OF THE MODEL TRAINED ON 2-SPEAKER MIXTURES BUT EVALUATED ON 3-SPEAKER MIXTURES

Dataset	Method	ΔSDR (dB)					SDR (dB)	SIR (dB)	SAR (dB)
		F3	F2-M1	F1-M2	M3	overall	overall	overall	overall
GRID	DC [17]	$1.93 \pm .52$	$4.52 \pm .92$	$4.77 \pm .47$	$1.41 \pm .57$	$3.11 \pm .22$	$0.53 \pm .21$	$4.75 \pm .32$	$5.05 \pm .13$
	uPIT [18]	$2.59 \pm .36$	$5.45 \pm .49$	$4.99 \pm .57$	$2.14 \pm .29$	$3.74 \pm .33$	$1.16 \pm .34$	$5.29 \pm .42$	$5.26 \pm .46$
	AVDC	$\mathbf{3.91 \pm .92}$	$\mathbf{5.54 \pm .65}$	$\mathbf{5.45 \pm .55}$	$\mathbf{3.44 \pm .41}$	$\mathbf{4.56 \pm .43}$	$\mathbf{1.98 \pm .45}$	$\mathbf{6.83 \pm .47}$	$\mathbf{5.57 \pm .32}$
	DC_{spk2}	$1.57 \pm .39$	$2.73 \pm .23$	$2.67 \pm .43$	$0.97 \pm .60$	$1.96 \pm .12$	$-0.63 \pm .12$	$2.87 \pm .15$	$4.91 \pm .13$
	$AVDC_{spk2}$	$4.12 \pm .84$	$4.68 \pm .75$	$4.32 \pm .73$	$2.84 \pm .69$	$3.87 \pm .59$	$1.29 \pm .61$	$5.51 \pm .76$	$5.49 \pm .36$
	DC⋆	$\mathbf{4.73 \pm .59}$	$\mathbf{6.36 \pm .73}$	$\mathbf{5.79 \pm .27}$	$3.57 \pm .49$	$4.97 \pm .11$	$2.39 \pm .10$	$7.17 \pm .16$	$5.35 \pm .11$
	uPIT⋆	$4.20 \pm .44$	$5.89 \pm .70$	$5.10 \pm .55$	$3.17 \pm .48$	$4.49 \pm .54$	$1.91 \pm .55$	$6.45 \pm .68$	$5.17 \pm .24$
	AVDC⋆	$4.62 \pm .73$	$5.96 \pm .59$	$5.70 \pm .30$	$\mathbf{3.98 \pm .19}$	$\mathbf{4.99 \pm .23}$	$\mathbf{2.41 \pm .24}$	$\mathbf{7.25 \pm .27}$	$\mathbf{5.38 \pm .21}$
	$DC^{\star}_{spk2}$	$4.77 \pm .49$	$5.78 \pm .36$	$4.95 \pm .42$	$3.11 \pm .55$	$4.52 \pm .32$	$1.94 \pm .33$	$6.43 \pm .37$	$5.13 \pm .23$
	$AVDC^{\star}_{spk2}$	$5.56 \pm .69$	$6.20 \pm .46$	$5.45 \pm .40$	$3.90 \pm .38$	$5.13 \pm .43$	$2.55 \pm .44$	$7.26 \pm .50$	$5.52 \pm .34$
	IBM	$12.83 \pm .72$	$12.97 \pm .48$	$12.38 \pm .35$	$11.46 \pm .35$	$12.32 \pm .44$	$9.74 \pm .44$	$16.00 \pm .47$	$11.13 \pm .41$
TCD-TIMIT	DC [17]	$2.26 \pm .49$	$4.83 \pm .24$	$5.26 \pm .26$	$1.67 \pm .48$	$3.47 \pm .26$	$0.81 \pm .26$	$5.45 \pm .33$	$5.28 \pm .11$
	uPIT [18]	$2.93 \pm .55$	$4.94 \pm .75$	$5.32 \pm .47$	$2.04 \pm .42$	$3.86 \pm .12$	$1.20 \pm .31$	$5.29 \pm .42$	$5.60 \pm .30$
	AVDC	$\mathbf{3.42 \pm .87}$	$\mathbf{4.95 \pm .52}$	$\mathbf{5.36 \pm .46}$	$\mathbf{3.70 \pm .76}$	$\mathbf{4.33 \pm .49}$	$\mathbf{1.67 \pm .50}$	$\mathbf{6.89 \pm .58}$	$\mathbf{5.64 \pm .26}$
	DC_{spk2}	$1.98 \pm .25$	$3.15 \pm .23$	$3.06 \pm .25$	$1.15 \pm .36$	$2.26 \pm .15$	$-0.40 \pm .16$	$3.60 \pm .15$	$4.78 \pm .19$
	$AVDC_{spk2}$	$3.41 \pm .22$	$4.15 \pm .28$	$4.63 \pm .36$	$3.01 \pm .34$	$3.77 \pm .27$	$1.11 \pm .28$	$5.81 \pm .32$	$5.46 \pm .22$
	DC⋆	$\mathbf{5.59 \pm .24}$	$\mathbf{7.41 \pm .31}$	$\mathbf{7.15 \pm .38}$	$5.20 \pm .50$	$\mathbf{6.31 \pm .35}$	$\mathbf{3.65 \pm .36}$	$\mathbf{8.81 \pm .42}$	$\mathbf{6.28 \pm .27}$
	uPIT⋆	$5.27 \pm .47$	$6.56 \pm .40$	$6.48 \pm .59$	$5.13 \pm .62$	$6.07 \pm .52$	$3.41 \pm .42$	$8.35 \pm .52$	$5.93 \pm .44$
	AVDC⋆	$5.38 \pm .34$	$6.71 \pm .25$	$6.36 \pm .39$	$\mathbf{5.30 \pm .63}$	$5.91 \pm .41$	$3.25 \pm .42$	$8.47 \pm .50$	$5.97 \pm .31$
	$DC^{\star}_{spk2}$	$4.91 \pm .37$	$6.44 \pm .54$	$5.72 \pm .54$	$4.49 \pm .54$	$5.35 \pm .48$	$2.69 \pm .49$	$7.45 \pm .54$	$5.68 \pm .36$
	$AVDC^{\star}_{spk2}$	$5.51 \pm .58$	$6.53 \pm .51$	$6.09 \pm .49$	$5.19 \pm .48$	$5.82 \pm .50$	$3.16 \pm .51$	$8.19 \pm .60$	$5.95 \pm .35$
	IBM	$15.80 \pm .49$	$15.20 \pm .46$	$14.35 \pm .50$	$13.45 \pm .64$	$14.51 \pm .34$	$11.85 \pm .34$	$18.54 \pm .35$	$13.09 \pm .34$

the amount of 3-speaker mixtures in training is not sufficient to cover the various kinds of mixing situations. This hypothesis can be partially verified by comparing $AVDC^{\star}_{spk2}$ with $DC^{\star}_{spk2}$, where only 2-speaker mixtures are used for training. With the same amount of training mixtures, a much better coverage of mixing conditions can be obtained on 2-speaker mixtures than 3-speaker mixtures. At the same time, a 0.61 dB and 0.47 dB improvement on the overall mean ΔSDR is observed on GRID and TCD-TIMIT, respectively.

Finally, we inspect the generalization ability of the AVDC approach to different numbers of speakers. Comparing $AVDC_{spk2}$ with AVDC, we only see a slight decrease, but comparing $AVDC_{spk2}$ with DC_{spk2}, we see a significant improvement on both datasets (1.91 dB mean ΔSDR on GRID and 1.51 dB on TCD-TIMIT). This suggests that even in unmatched conditions, visual information shows great advantages for speech separation. Interestingly, $AVDC_{spk2}$ even outperforms DC on both datasets on all overall metrics.

F. Ablation Study

The most significant difference between our method and previous approaches [12], [14], [15] is that our proposed model exploits a two-stage audio-visual fusion strategy which can better integrate information from both modalities, thus taking advantages of both the audio-only approach and the audio-visual approach. We verify our claim through ablation experiments on both datasets with 2-speaker and 3-speaker mixtures. For comparison, we delete the "Audio-Only Embedding″ part of the AVDC model in Figure 3, and denote it as the "AVDC-WOA″ model in Figures 5 and 6.

Figure 5 shows 5-fold cross validation results of ΔSDR on 2-speaker mixtures as boxplots. There are two interesting observations. First, across all kinds of mixtures and both datasets, the proposed AVDC model significantly outperforms the AVDC-WOA model. This suggests that the integration of audio embeddings in the second-stage audio-visual fusion of AVDC significantly improves the separation performance. Second, on same-gender mixtures, the AVDC-WOA model significantly outperforms DC, while on different-gender mixtures, it significantly underperforms DC. In fact, on different-gender mixtures, DC already achieves a similar performance as the proposed AVDC. This suggests that the audio-visual embedding learned in the first-stage fusion is only helpful for same-gender mixtures, where the source permutation problem is the major source of error. On different-gender mixtures, audio-visual matching is not that helpful for improving speech separation performance.

Figure 6 shows 5-fold cross validation results of ΔSDR on 3-speaker mixtures. It shows a similar trend. The differences, however, are twofold: First, the performance gap between AVDC-WOA and AVDC is smaller on same-gender mixtures. This suggests that the source permutation problem is even

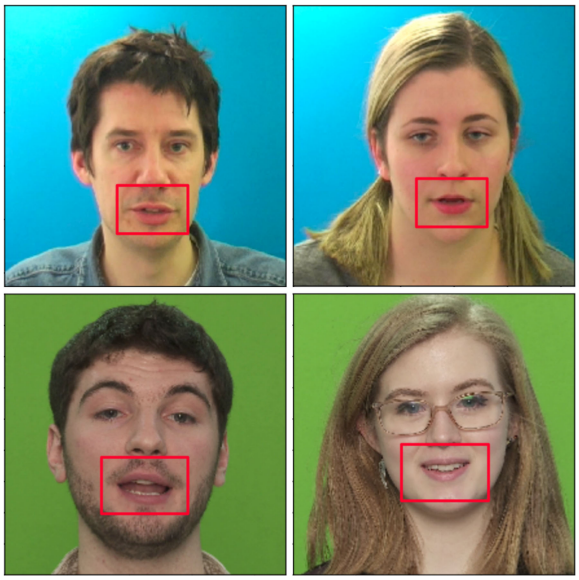


Fig. 4. Screen shots of speech videos in our experiments. The first row are a male and a female speaker from the GRID dataset. The second row are a male and a female speaker from the TCD-TIMIT dataset. Bounding boxes mark the lip regions where visual information is extracted and fed to our model.

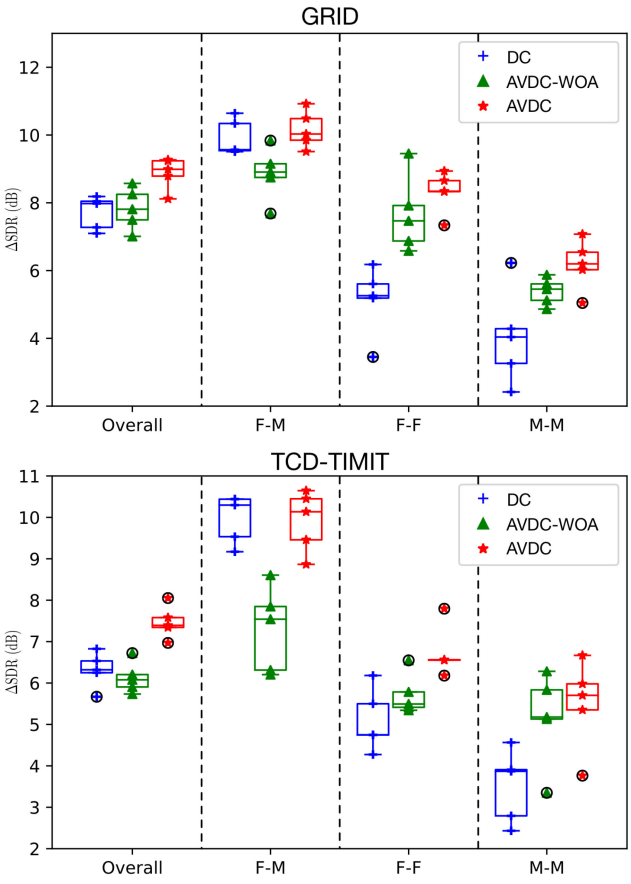


Fig. 5. Ablation study of the proposed AVDC model on 2-speaker mixtures in both the GRID and TCD-TIMIT datasets. Boxplots of ΔSDR with 5-fold cross validation are shown on different types of speech mixtures for DC (AVDC without the AV-branch in second-stage fusion), AVDC-WOA (AVDC without the audio-branch in second-stage fusion), and AVDC.

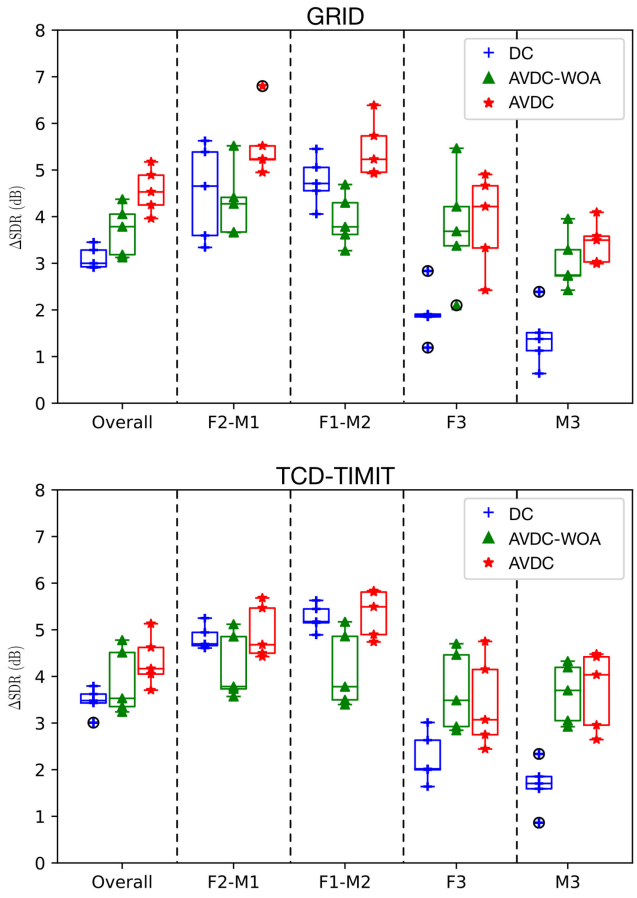


Fig. 6. Ablation study of the proposed AVDC model for 3-speaker mixtures in both the GRID and TCD-TIMIT datasets. Boxplots of ΔSDR with 5-fold cross validation are shown on different types of speech mixtures for DC (AVDC without the AV-branch in second-stage fusion), AVDC-WOA (AVDC without the audio-branch in second-stage fusion), and AVDC.

more dominant in 3-speaker same-gender mixtures. Second, on different-gender mixtures (F2-M1 and F1-M2), both AVDC and AVDC-WOA are slightly lifted relative to DC, resulting in AVDC slightly outperforming DC while AVDC-WOA slightly underperforming DC. This, again, suggests that the source permutation problem weighs higher in 3-speaker different-gender mixtures than the 2-speaker different-gender mixtures; after all, there are always two speakers with the same gender.

An interesting observation in Figures 5 and 6 is that for different-gender mixtures, the AVDC-WOA model underperforms the DC model. This is less intuitive as one would expect that although integrating audio and visual information may not improve the separation performance over audio-only methods, it would not degrade the performance either. We argue that a possible explanation for this is that the audio-visual embeddings that the AVDC-WOA model learns are helpful for solving the "source permutation" problem in audio-only methods. However, such embeddings may contain noisy and misleading correlations between audio and visual features and may degrade single-frame separation performance. An extreme situation is the ventriloquism effect, where the correlation is entirely misleading for separation. When the "source permutation" problem is not

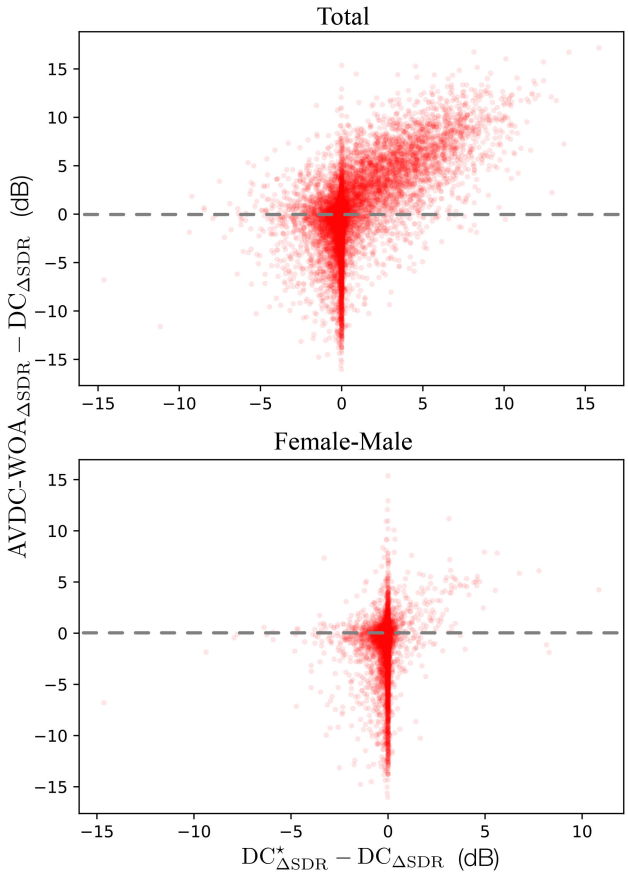


Fig. 7. Visualization of separation performance of different models on 2-speaker mixtures of the GRID dataset. The x-axis indicates the difference of ΔSDR between DC* and DC. The y-axis indicates the difference of ΔSDR between AVDC-WOA and DC.

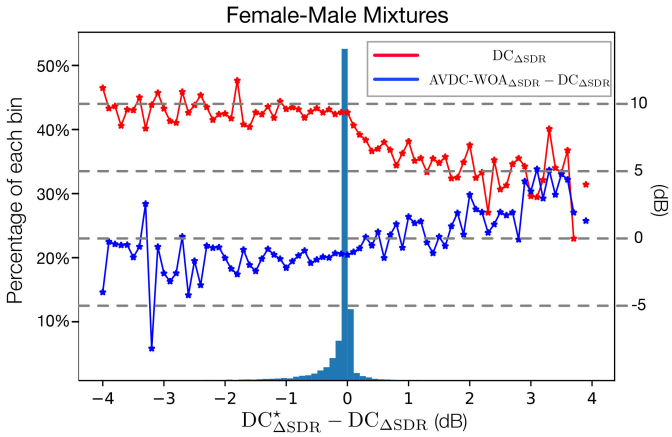


Fig. 8. Histogram shows the percentage of F-M mixtures in terms of the DC$^{\star}_{\Delta SDR}$ - DC$_{\Delta SDR}$ values. The red line shows average DC$_{\Delta SDR}$ value of mixtures in each histogram bin, while the blue line shows the average AVDC-WOA$_{\Delta SDR}$ - DC$_{\Delta SDR}$ value.

significant, as for different-gender mixtures, the advantages of AVDC-WOA are overtaken by its disadvantages and the performance drops from audio-only methods. It is noted that the proposed AVDC model, however, addresses the disadvantages of AVDC-WOA, through the second-stage fusion of audio and visual information.

We further explain this phenomenon in Figure 7. Each dot in Figure 7 represents a mixture in the test set of the GRID dataset. The x-axis is the difference of ΔSDR between DC* and DC. It indicates the degree of the permutation errors of the DC model: bigger DC$^{\star}_{\Delta SDR}$ − DC$_{\Delta SDR}$ values indicate more significant permutation errors. The y-axis is the difference of ΔSDR between AVDC-WOA and DC. It indicates the improvement due to the integration of visual information in the first-stage fusion. We draw all the mixtures together in the top subfigure and only the F-M mixtures in the bottom subfigure. The top subfigure shows a strong distribution along the diagonal, suggesting that there are many permutation errors in DC, and AVDC-WOA is able to improve separation performance in these cases. The bottom subfigure, however, does not show many points on the positive range of the x-axis; The DC model does not encounter many permutation errors for F-M mixtures and the AVDC-WOA model does not improve over DC. In fact, we find that AVDC-WOA performs worse than DC on about 55% of the F-M mixtures.

Similar observations can be made from Figure 8, where the x-axis has the same meaning as that of Figure 7. In this figure, we first draw the histogram of all the F-M mixtures with a bin size of 0.1 dB and show the percentage of each bin on the left y-axis. Then we compute the average DC$_{\Delta SDR}$ value of the mixtures in each bin and show them as the red line (right y-axis). Similarly, we also show the average AVDC-WOA$_{\Delta SDR}$ - DC$_{\Delta SDR}$ of the mixtures in each histogram bin as the blue line. We have three interesting observations from this figure: First, as can be seen from the histogram, the gap between DC$^{\star}_{\Delta SDR}$ and DC$_{\Delta SDR}$ for more than 72% of the F-M mixtures is below 0.1 dB, indicating that permutation errors are not significant for different-gender mixtures. Second, the negative part of the histogram and the red line indicate that when the audio-only DC model has excellent separation performance (DC$_{\Delta SDR}$ values around 10 dB), optimal assignment across separated frames often deteriorates the final separation results (as DC$^{\star}_{\Delta SDR}$ < DC$_{\Delta SDR}$). Under such circumstances, the integration of visual information also often degrades the final separation performance. Third, the positive part of the histogram and points show that when there are significant permutation errors (DC$^{\star}_{\Delta SDR}$ > DC$_{\Delta SDR}$), the DC$_{\Delta SDR}$ value drops significantly and the improvement of ΔSDR from DC to AVDC-WOA increases consistently, indicating that visual information helps alleviate source permutation errors. However, the DC model only show significant permutation problems on 33% of the F-M mixtures, resulting in the phenomenon that the AVDC-WOA model has worse average performance than the DC model in general on F-M mixtures.

G. Speaker-Wise Audio–Visual T–F Embeddings

After the second-stage fusion, the final T-F embedding matrix of our model is a concatenation of two matrices: The audio-only

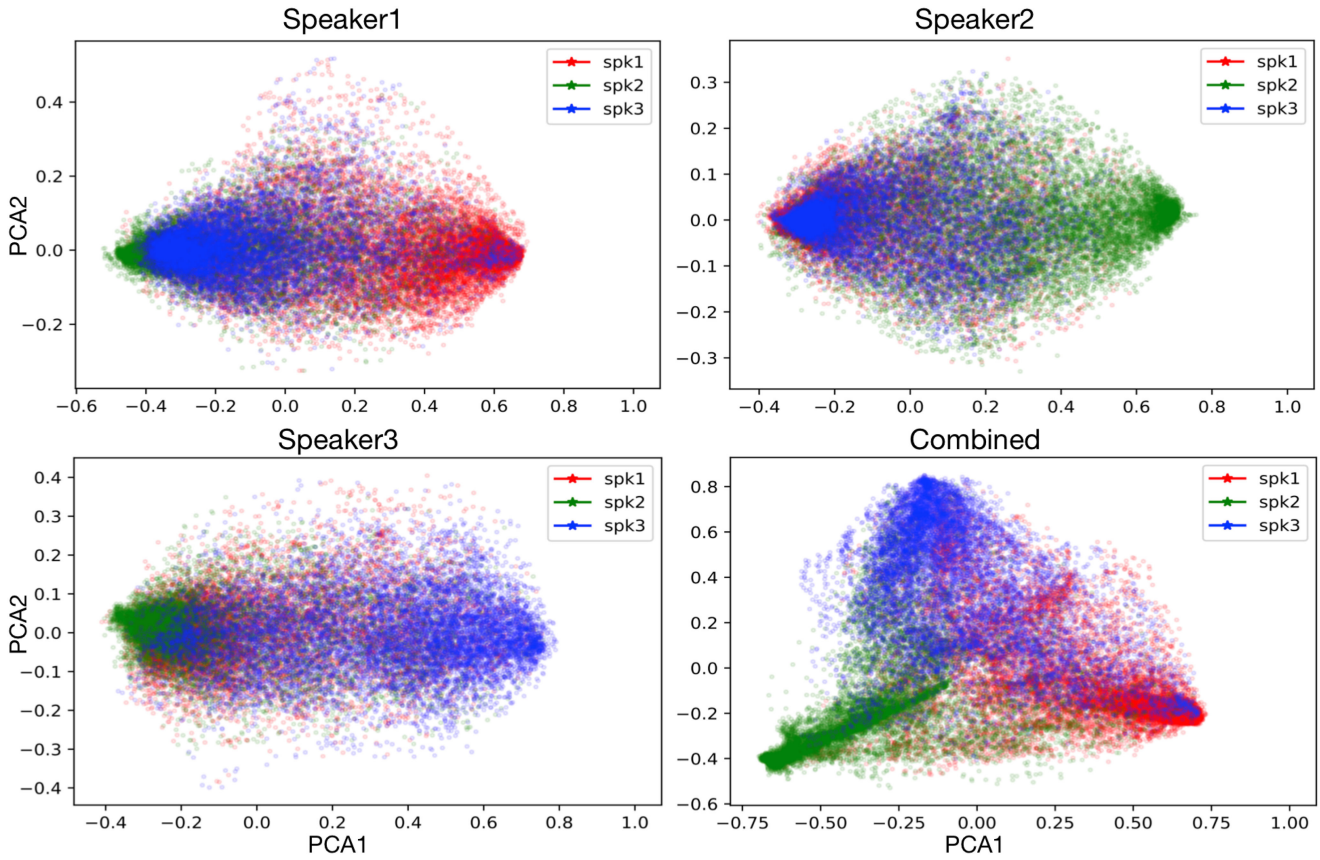


Fig. 9. Visualization of audio-visual embeddings of a 3-speaker (3-female) mixture in the test set of the GRID dataset. Each of the first three subfigures shows the PCA in two dimensions of a speaker's audio-visual embedding vectors of all T-F bins. The target speaker is separated from the other two speakers that are still mixed. The last subfigure shows the PCA in two dimensions of the combined (concatenated) 3-speaker embedding vectors of all T-F bins. All of the speakers are separated.

embedding matrix (same as that of DC), and the audio-visual embedding matrix obtained after the first-stage audio-visual fusion. To further analyze the proposed model, in this section we visualize these audio-visual embeddings. It is noted that these embeddings are speaker-dependent; each embedding captures the audio-visual correspondence for a specific speaker in the mixture. Therefore, each embedding should show a 2-cluster clustering effect differentiating the target speaker from all the other speakers. The Principal Component Analysis (PCA) visualization on a 3-speaker mixture in Figure 9 indeed shows this phenomenon. This is an F3 (3-female) mixture from the test set of the GRID dataset. The AV embeddings are calculated by passing the AV data of the mixture through the trained 1-stage of the network. Although each speaker goes through the same network, their resulted embeddings are different. In Figure 9, each of the first three subfigures shows the PCA on 2 dimensions of a speaker's embedding vectors of all T-F bins. Each subfigure clearly shows a 2-cluster clustering effect, where the target speaker is separated from the other two speakers that are essentially mixed up. The last subfigure, on the other hand, performs PCA on the combined (concatenated) embedding vectors of all of the 3 speakers. It can be seen that all of the 3 speakers are well separated. This analysis provides further evidence showing that the AV embeddings help separate mixture audio into source components and consistently assign source components to correct sources.

H. Partially Observed Videos

In this section, we evaluate the robustness of our proposed model when only partial visual information is observed. To perform clustering on embedding vectors with missing dimensions, we exploit the k-POD algorithm as described in Algorithm III-C. For 2-speaker mixtures, we randomly select one speaker and occlude the speaker during the middle third segment of each video. The results on these mixtures are denoted as AVDC$_{occ1}$ in Table V. For 3-speaker mixtures, we create two kinds of partially observed videos. The first kind, denoted as AVDC$_{occ1}$ in Table VI, is to randomly select one speaker and to occlude the speaker in the middle third segment of the video; While the second kind, denoted as AVDC$_{occ2}$, is to randomly select two speakers and to occlude both speakers in the middle third segment of the video. The audio is not altered in these videos.

As shown in Table V, when one speaker's visual information is partially missing in the 2-speaker mixtures, the separation performance drops for the same-gender mixtures. On GRID, the mean ΔSDR decreases by 0.96 dB, while on TCD-TIMIT it decreases by 1.00 dB. For different-gender mixtures, however, the

TABLE V
EXPERIMENTS ON THE PARTIALLY OBSERVED VIDEOS OF 2-SPEAKER MIXTURES

Dataset	Method	ΔSDR (dB)		
		same-gender	different-gender	overall
GRID	DC [17]	4.58 ± 0.87	9.92 ± 0.48	7.72 ± 0.44
	AVDC	7.07 ± 0.67	10.16 ± 0.49	8.88 ± 0.42
	AVDC$_{occ1}$	6.11 ± 0.55	10.00 ± 0.53	8.39 ± 0.44
	IBM	11.87 ± 0.56	12.76 ± 0.43	12.39 ± 0.43
TCD-TIMIT	DC [17]	4.18 ± 0.38	9.98 ± 0.53	6.32 ± 0.38
	AVDC	6.04 ± 0.32	9.91 ± 0.66	7.47 ± 0.35
	AVDC$_{occ1}$	5.04 ± 0.31	9.63 ± 0.63	6.73 ± 0.38
	IBM	14.47 ± 0.32	15.01 ± 0.53	14.67 ± 0.39

TABLE VI
EXPERIMENTS ON THE PARTIALLY OBSERVED VIDEOS OF 3-SPEAKER MIXTURES

Dataset	Method	ΔSDR (dB)		
		same-gender	different-gender	overall
GRID	DC [17]	1.59 ± 0.35	4.63 ± 0.64	3.11 ± 0.22
	AVDC	3.64 ± 0.45	5.48 ± 0.58	4.56 ± 0.43
	AVDC$_{occ1}$	3.15 ± 0.43	5.16 ± 0.54	4.15 ± 0.40
	AVDC$_{occ2}$	2.57 ± 0.37	4.80 ± 0.56	3.68 ± 0.36
	IBM	11.99 ± 0.54	12.65 ± 0.40	12.32 ± 0.44
TCD-TIMIT	DC [17]	1.88 ± 0.35	5.06 ± 0.18	3.47 ± 0.26
	AVDC	3.51 ± 0.62	5.16 ± 0.47	4.33 ± 0.49
	AVDC$_{occ1}$	2.62 ± 0.56	4.43 ± 0.35	3.52 ± 0.45
	AVDC$_{occ2}$	1.72 ± 0.27	3.88 ± 0.14	2.80 ± 0.17
	IBM	14.29 ± 0.28	14.73 ± 0.44	14.51 ± 0.34

performance only drops slightly (0.16 dB on GRID and 0.28 dB on TCD-TIMIT). This again verifies that visual information is specifically effective for same-gender mixtures. Moreover, it is worth noting that for these partially observed videos, the proposed AVDC model still outperforms the audio-only DC model.

Table VI shows results on 3-speaker mixtures with partial visual information. On the GRID dataset, we observe a slight decrease when one speaker is partially missing (AVDC$_{occ1}$). The decrease is more significant when two speakers are missing (AVDC$_{occ2}$), but the performance is still significantly higher than DC, especially for same-gender mixtures. On the TCD-TIMIT dataset, when we occlude one speaker (AVDC$_{occ1}$), we observe a similar ΔSDR with the DC model; When two speakers' videos are occluded (AVDC$_{occ2}$), ΔSDR drops dramatically. We suggest that the different behaviors on the two datasets are due to the phonetic richness of the datasets. When the dataset is phonetically rich such as the TCD-TIMIT dataset, visual information plays a more important role in speech separation, and if missing, a more significant drop will be observed on the separation performance.

V. CONCLUSION

In this paper, we proposed an Audio-Visual Deep Clustering model for speaker-independent speech separation. The proposed two-stage audio-visual fusion strategy learns speaker-wise audio-visual T-F embeddings in the first stage. It then concatenates the audio-visual embeddings and audio embeddings in the second stage for T-F clustering. The proposed method outperforms the audio-only deep clustering and uPIT approaches and three other state-of-the-art audio-visual approaches in our experiments. The proposed model shows flexibility in generalization across different numbers of speakers between training and testing. It also shows robustness against partially missing visual information in the videos. Future work includes the expansion of the dataset for training and testing to include speech recordings with more natural and rich poses and speech contents.

REFERENCES

[1] E. Z. Golumbic, G. B. Cogan, C. E. Schroeder, and D. Poeppel, "Visual input enhances selective speech envelope tracking in auditory cortex at a cocktail party," *J. Neuroscience*, vol. 33, no. 4, pp. 1417–1426, 2013.

[2] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[3] W. J. Ma, X. Zhou, L. A. Ross, J. J. Foxe, and L. C. Parra, "Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space," *PLoS One*, vol. 4, no. 3, 2009, Art. no. e4638.

[4] S. Parekh, S. Essid, A. Ozerov, N. Q. Duong, P. Pérez, and G. Richard, "Motion informed audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 6–10.

[5] B. Rivet, L. Girin, and C. Jutten, "Visual voice activity detection as a help for speech source separation from convolutive mixtures," *Speech Commun.*, vol. 49, no. 7–8, pp. 667–677, 2007.

[6] B. Rivet, L. Girin, and C. Jutten, "Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 96–108, Jan. 2007.

[7] W. Wang, D. Cosker, Y. Hicks, S. Saneit, and J. Chambers, "Video assisted speech source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. 425–428.

[8] A. L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation based on sparse redundant representations," *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 358–371, Aug. 2010.

[9] Q. Liu, W. Wang, P. J. Jackson, M. Barnard, J. Kittler, and J. Chambers, "Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5520–5535, Nov. 2013.

[10] D. Sodoyer, J.-L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten, "Separation of audio-visual speech sources: A new approach exploiting the audio-visual coherence of speech stimuli," *EURASIP J. Appl. Sign. Process.*, vol. 2002, no. 1, pp. 1165–1173, 2002.

[11] Q. Liu, W. Wang, and P. Jackson, "Use of bimodal coherence to resolve the permutation problem in convolutive BSS," *Sign. Process.*, vol. 92, no. 8, pp. 1916–1927, 2012.

[12] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," in *Proc. Assoc. Comput. Machinery (ACM) Special Interest Group Comput. Graphics Interactive Techn. (SIGGRAPH)*, vol. 37, no. 4, pp. 112:1–112:11, 2018.

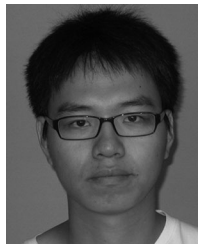
[13] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg, "Seeing through noise: Speaker separation and enhancement using visually-derived speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 3051–3055.

[14] R. Lu, Z. Duan, and C. Zhang, "Listen and look: Audio–visual matching assisted speech source separation," *IEEE Signal Process. Lett.*, vol. 25, no. 9, pp. 1315–1319, Sep. 2018.

[15] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *Proc. Interspeech*, 2018, pp. 1170–1174.

[16] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 31–35.

[17] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, 2016, pp. 545–549.
[18] 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.
[19] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley-IEEE press, 2006.
[20] K. Hu and D. Wang, "An unsupervised approach to cochannel speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 122–131, Jan. 2013.
[21] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 1996, pp. 472–478.
[22] T. Virtanen, "Speech recognition using factorial hidden Markov models for separation in the feature space," in *Proc. Interspeech*, 2006, pp. 89–92.
[23] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 45–66, 2010.
[24] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2001, pp. 556–562.
[25] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 1–12, Jan. 2007.
[26] N. Seichepine, S. Essid, C. Févotte, and O. Cappé, "Soft nonnegative matrix co-factorization," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5940–5949, Nov. 2014.
[27] D. El Badawy, N. Q. Duong, and A. Ozerov, "On-the-fly audio source separation—novel user-friendly framework," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 2, pp. 261–272, Feb. 2017.
[28] P. Smaragdis, B. Raj, and M. V. S. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. Int. Conf. Independent Component Anal. Signal Separation (ICA)*, 2007, pp. 414–421.
[29] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.
[30] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.
[31] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.
[32] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 4, pp. 787–796, Apr. 2018.
[33] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 1562–1566.
[34] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 246–250.
[35] Z.-Q. Wang, J. L. Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 686–690.
[36] Z.-Q. Wang, J. Le Roux, D. Wang, and J. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," in *Proc. Interspeech*, 2018, pp. 2708–2712.
[37] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 241–245.
[38] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. Asian Conf. Comput. Vision (ACCV)*, 2016, pp. 87–103.
[39] J. Son Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2017, pp. 6447–6456.
[40] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 125–134, May 2014.
[41] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 117–128, Apr. 2018.
[42] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018, pp. 3244–3248.
[43] S. Parekh, A. Ozerov, S. Essid, N. Duong, P. Pérez, and G. Richard, "Identify, locate and separate: Audio-visual object extraction in large video collections using weak supervision," Preprint, 2018, *arXiv:1811.04000*.
[44] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, 2018, pp. 570–586.
[45] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2016, pp. 2405–2413.
[46] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, 2017, pp. 609–617.
[47] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, "3D convolutional neural networks for cross matching recognition," *IEEE Access*, vol. 5, pp. 22081–22091, 2017.
[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.
[49] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.
[50] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 568–576.
[51] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2016, pp. 1933–1941.
[52] J. T. Chi, E. C. Chi, and R. G. Baraniuk, "*k*-POD: A method for *k*-means clustering of missing data," *The Amer. Statistician*, vol. 70, no. 1, pp. 91–99, 2016.
[53] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The J. Acoustical Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424, 2006.
[54] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 603–615, May 2015.
[55] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Speech Input/Output Assessment Speech Databases*, pp. 121–124, 1989.
[56] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jul. 2009.
[57] C. Liu *et al.*, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 2009.
[58] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan, "Learning features by watching objects move," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2017, pp. 2701–2710.
[59] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.
[60] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.



Rui Lu received the B.S. degree from Tsinghua University, Beijing, China, in 2013. He is currently working toward the Ph.D. degree with the Department of Automation, Tsinghua University, Beijing, China.

His research interests primarily include the area of signal processing and machine learning toward music information retrieval, environmental sound classification and detection, and audio–visual source separation.

Zhiyao Duan (S'09–M'13) received the B.S. degree in automation and the M.S. degree in control science and engineering from Tsinghua University, Beijing, China, in 2004 and 2008, respectively, and the Ph.D. degree in computer science from Northwestern University, Evanston, IL, USA, in 2013.

He is an Assistant Professor in the Department of Electrical and Computer Engineering and the Department of Computer Science, University of Rochester, Rochester, NY, USA. His research interest is in the broad area of computer audition, i.e., designing computational systems that are capable of understanding sounds, including music, speech, and environmental sounds.

Dr. Duan co-presented a tutorial on automatic music transcription at the International Conference on Music Information Retrieval 2015. He received a best paper award at the 2017 Sound and Music Computing Conference, a best paper nomination at the 2017 International Conference on Music Information Retrieval, and a CAREER award from the National Science Foundation.

Changshui Zhang received the B.S. degree from the Peking University, Beijing, China, in 1986, and the Ph.D. degree from Tsinghua University, Beijing, China, in 1992.

He is currently a Professor with the Department of Automation, Tsinghua University.

Dr. Zhang is an Associate Editor of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE . His interests include machine learning, artificial intelligence, pattern recognition, and computer vision.