FILTER-RECOVERY NETWORK FOR MULTI-SPEAKER AUDIO-VISUAL SPEECH SEPARATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We aim at audio-visual speech separation task. Given the face information for each speaker, the goal is to separate the corresponding speech in the speech mixture. Existing works are designed for a controlled setting with a fixed number of speakers, mostly 2 or 3 speakers, which is not easily scalable in practical application. To deal with this, we focus on separating voices for variable number of speakers with a single model, and build concrete mixture test sets for a fair comparison. There are two prominent issues in complex multi-speaker separation results: 1) There exists some noisy voice pieces belong to other speakers; 2) Part of the target speech is missing. To deal with these, we propose a valid method BFRNet, including a basic audio-visual speech separator and a Filter-Recovery Network (FRNet). The FRNet filters the noisy speech and recovery the missing parts for the output of the basic separator. Our method achieves the state-of-the-art results on audio-visual speech separation datasets. Besides, we apply the FRNet to other methods and achieve general performance improvements, which proves the effectiveness of the proposed FRNet.

1 INTRODUCTION

Audio-visual speech separation has wide applications in practice, such as speech recognition (Radford et al.; Chan et al., 2015), assistive hearing technologies, and online video meeting. As human voices are usually mixed together in public, it would be challenging if we want to directly extract the information of interest from the raw and mixed audio-visual signal with multiple speakers. As a result, separating audio signals for each speaker could serve as an effective pre-processing block for further audio-visual signal analysis.

Audio-visual speech separation has made prominent progress along with the development of convolutional neural networks (Gogate et al., 2018; Makishima et al., 2021; Gao & Grauman, 2021) and transformers (Ramesh et al., 2021; Montesinos et al., 2022; Rahimi et al., 2022). For example, previous works (Lee et al., 2021; Gao & Grauman, 2021; Montesinos et al., 2022) have studied the audio-visual speech separation but only focus on two-speaker separation. Although other works (Ephrat et al., 2018; Afouras et al., 2018b; 2019) manage to separate voices under multi-speaker scenarios, Ephrat et al. (2018) require customized models for different number of speakers, and (Afouras et al., 2018b) is more likely to enhance the target voice. Furthermore, Afouras et al. (2019) make use of pre-enrolled speaker embeddings to extract the corresponding speech, but still leave a performance gap compared with un-enrolled speakers. Therefore, it still remains as a challenge that how to efficiently perform audio-visual speech separation in a multi-speaker scenario. The difficulty mainly come from two aspects: the more essential challenge of recovering individual speech from a multi-speaker mixture, and the lack of a well-established benchmark for speech separation of multi-speaker scenario.

Through our explorations, most of the currently works show their superiority on two-speaker speech separation but yield disappointing results with more speakers (*e.g.*, 3, 4 or even 5 speakers). It exhibits a simple yet important fact that the complexity of speech separation strongly correlates with the number of speakers in the mixed audios. As shown in Fig. 1, the core problems for multi-speaker speech separation can be empirically summarized as two folds: 1) The **noisy parts** (red box in Figure 1) contain the components from other speakers within the separated audio and usually occur in complex scenarios. 2) Some **missing parts** (gray box in Figure 1) represent there are some

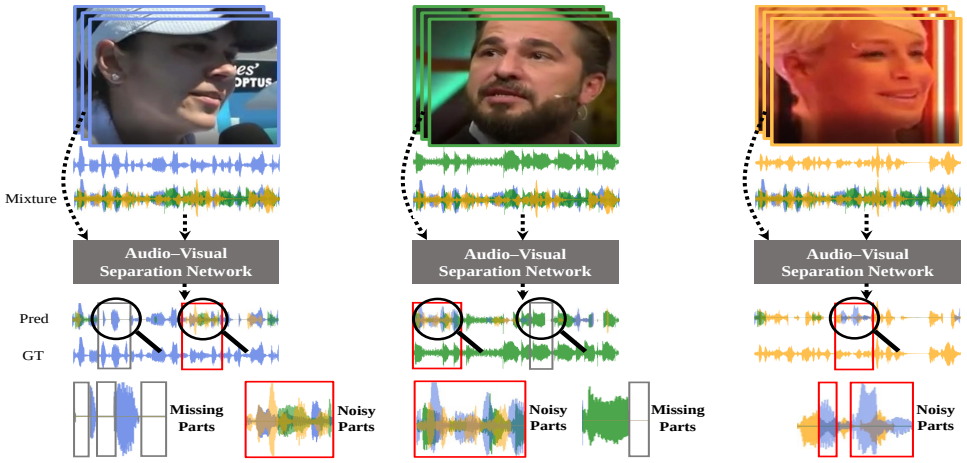

Figure 1: **Illustration of the two critical issues in multi-speaker audio-visual speech separation.**

target audio waves dropped by model. These two problems usually occur in complex scenarios. It reflects that the models are incapable of accurately separating target audio signal under complex scenarios. As a result, there is still a long way to go towards solving audio-visual speech separation among multi-speakers.

To this end, we come up with an effective method, termed as **BFRNet**, to conquer aforementioned challenges under audio-visual speech separation from multi-speakers. As illustrated in Figure 2, the BFRNet is composed of a basic audio-visual speech separator and a Filter-Recovery Network (FRNet). The FRNet aims at solving the two issues in the results of basic separator. It is composed of two serial modules, *i.e.*, **Filter** and **Recovery**. Firstly, Filter module utilizes visual features of target speaker to query corresponding audio signals from the coarse prediction and suppress component from other speakers. Then, the Recovery module use the clean audio yielded by Filter module to query the signals that belong to target speaker from coarse prediction. Essentially, our method aims to calibrate the coarsely separated audio predicted by off-the-shelf models (Gao & Grauman, 2021; Montesinos et al., 2022; Chuang et al., 2020). As a result, the noisy parts and missing parts are greatly reduced by our Filter-Recovery operation.

Furthermore, we have noticed that there is still one obstacle resided in evaluation of audio-visual speech separation. As we found most of works (Gao & Grauman, 2021; Montesinos et al., 2022) use unfixed number of samples which are generated by randomly mixing the audio signals of different identities during inference, it sometimes leads to hard reproductions and unfair comparisons. Consequently, to unify the evaluation protocols, we curate a comprehensive benchmark to fairly verify the models' performance. Specifically, Specifically, for each type of mixture, we randomly sample test videos without replacement to make up the speech mixtures. The constructed fixed test sets serve for all experiments to ensure fairness and reproducibility of the results.

To sum up, our contribution can be summarized as follows:

- *First*, we systematically study the advance in audio-visual speech separation under multi-speaker scenarios.
- *Second*, to address the observed problems, we design the Filter-Recovery network to further improve the quality of multi-speaker speech separation.
- *Finally*, we define a well-established test benchmark for fair comparison. The exhaustive experiments on this benchmark validate the effectiveness of the proposed method.

2 Related Works

Audio-Only Speech Separation. Using only audio modality for speech separation faces the problem of speaker agnosticism. Some works (Liu et al., 2019; Wang et al., 2018) utilize the speaker's voice embedding as the hint to isolate the target speech. Current methods mostly treat the audio-only

speech separation as a label permutation problem. Chen et al. (2017) cluster the similar speech to perform speech separation. Luo et al. (2018) requires no priori information of speaker number. Luo & Mesgarani (2019) adopts a deep learning network consisting of a series of convolutional layers and train the network with permutation invariant loss (PIT) (Yu et al., 2017).

Visual Sound Separation. There are various kinds of sound separation explored in the literature. One type focuses on music separation (Zhao et al., 2018; Gao & Grauman, 2019; Xu et al., 2019; Gan et al., 2020), where the diverse shape of musical instruments and the distinguished patterns of music sounds are the key clues for music separation. Other works concentrate on the in-the-wild sounds (Gao et al., 2018; Tzinis et al., 2020; Chen et al., 2020), such as animal sound, vehicle sound, etc. Gao et al. (2018) incorporate the image recognition results as the advisor and learn prototypical spectral patterns for each sounding object. Tzinis et al. (2020) extend to unsupervised, open-domain audio-visual sound separation and develop a large new dataset. Chen et al. (2020) propose a algorithm to generate temporal synchronized sound given mismatched visual information.

Visual Speech Separation. Speech, the audio most closely associated with humans, is extensively studied and has a broad range of applications. Due to the natural correlation between face and speech, many works (Gabbay et al., 2018; Lu et al., 2018; Ephrat et al., 2018; Afouras et al., 2019; Chung et al., 2020; Hegde et al., 2021; Rahimi et al., 2022) employ face-related information to separate speech from a mixture in the literature. Chung et al. (2020) use only the still images containing facial appearance to isolate speech, with the assistance of the consistency of face identity and speech identity. Numerous methods (Afouras et al., 2018b; Gao & Grauman, 2021; Ephrat et al., 2018) explore the simultaneous lip motions and voice fluctuations clues. Lu et al. (2018) integrate optical flow and lip movements to predict the spectrogram masks. Hegde et al. (2021) propose synthesizing a virtual visual stream to deal with the situation where the visual stream is unreliable or completely absent. Another family of works (Owens & Efros, 2018; Afouras et al., 2020; Truong et al., 2021) combines multiple tasks for joint learning.

The most relevant works to ours are (Afouras et al., 2018b), Shi et al. (2020), and Yao et al. (2022). Afouras et al. (2018b) also pays attention to the uncontrolled environment with several speakers, but it only focus on enhancing the target speech and suppressing the noisy voices. Our method further separates every speech component for different mixtures in a single model. Shi et al. (2020) and (Yao et al., 2022) perform a coarse-to-fine procedure by adopting an additional refining separation phase. However, the refining phase applies the same model as the coarse phase. Thus if the coarse phase cannot achieve a clean separation, the refining results will likewise be sub-optimal.

3 METHOD

We are the first who pay attention to the audio-visual speech separation in a multi-speaker environment, where there are various numbers of speakers in speech mixtures during training. In this section, we first elaborate on the proposed multi-speaker training setting in Sec. 3.1, including the details about the multi-speaker training setting; Next we introduce the basic audio-visual speech separation network in Sec. 3.2, which takes the visual information and the speech mixture as input and outputs isolated speech; Then we present the Filter-Recovery network in Sec. 3.3; Finally we present the objective function for training model in Sec. 3.4.

3.1 OVERVIEW

Given a video containing S simultaneous speakers, our goal is to isolate the individual speech for each speaker. Formally, we denote time-domain speech mixture as $x = \sum_{i=1}^{S} x_i, x_i \in \mathbb{R}^{T_x}$, where T_x represents the time length, and x_i is the individual speech of the i-th speaker. Since acquiring the exact individual speech from the actual mixtures is impossible, we follow previous works (Afouras et al., 2018b; Gao & Grauman, 2021) to synthesize mixtures by adding individual speech together.

Current works are designed for definite speakers in the mixture, mostly 2 or 3 speakers. However, these models present poor effects for practical applications, where there are usually variable multiple number of speakers. Alternatively, they take an inefficient strategy of training separate models for mixtures with different numbers of speakers. To cope with the practical scenes, we create different mixtures with various numbers of speakers during training, i.e., S is ranged from 2 to 5. The network is required to isolate each speech component for all kinds of mixtures.

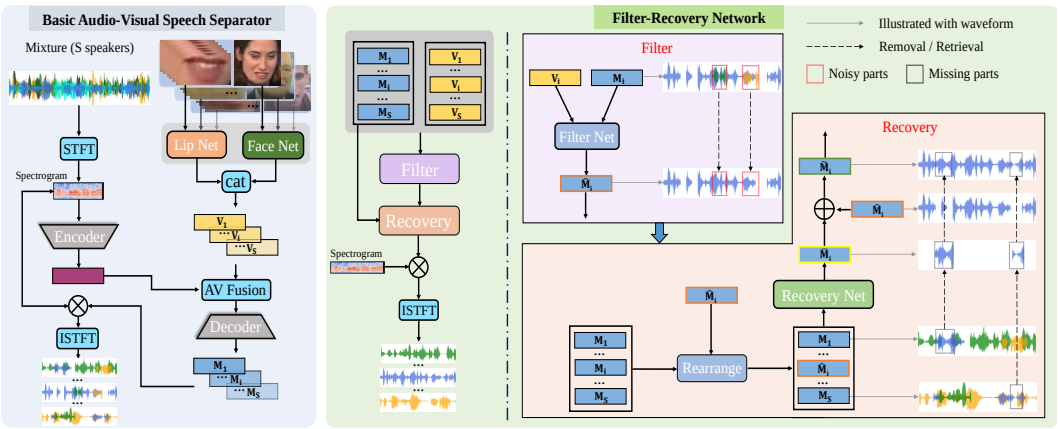

Figure 2: **Overview of the proposed framework.** It consists of a basic audio-visual speech separator and a Filter-Recovery network. To better illustrate the effect of the FRNet, we visualize masks into the form of waveforms by $\rightarrow$ to present speech components. The red/gray boxes and the corresponding $\dashrightarrow$ represents the noisy/missing speech pieces are removed or retrieved.

Fig. 2 shows the architecture of the proposed model, consisting of a basic audio-visual speech separator and an additional Filter-Recovery Network (FRNet). During training, we feed all kinds of mixtures into the audio-visual speech separator and output all separated speech. The model takes the time-frequency mixture spectrogram $X \in \mathbb{R}^{2\times F\times T_X}$ as input, which is converted from time-domain speech via Short-Time Fourier Transform (STFT). F indicates frequency dimension and T_X denotes the time dimension of the spectrogram. Each time-frequency bin contains the real and imaginary part of the corresponding complex spectrogram. The model outputs mask $M_i \in \mathbb{R}^{2\times F\times T_X}$ for input mixture spectrogram X and the extracted visual feature V_i for i-th speaker.

Afterwards, for each mixture, the separated S masks are sent to the FRNet, in order to filter the mixed speech and recover the missing voices. The FRNet outputs improved mask for each speaker in the mixture, which further adopted to restore the exact time-domain speech.

3.2 Basic Audio-Visual Speech Separator

In the proposed framework, the basic audio-visual speech separator can be replaced to any network which outputs the mask of mixture spectrogram. Following (Gao & Grauman, 2021), we adopt both face and lip clues to guide target speech separation. Specifically, a face net is used to explore face attributes, and a lip net is taken to dig out the consistency of lip motion and speech pronunciation.

Lip Net. Following the previous structure (Ma et al., 2021; Gao & Grauman, 2021), we feed T_v consecutive frames of lip regions into a 3D convolutional layer followed by a ShuffleNet v2 (Ma et al., 2018) network to extract mouth features. A temporal convolutional network is further utilized to output the lip motion features $Lip_i \in \mathbb{R}^{C_l\times T_v}$ for i-th speaker.

Face Net. The Face Net is aimed at leveraging the consistency between face attributes and speech. A ResNet-18 network takes a single face image as input, and outputs a face embedding for i-th speaker. We repeat the face embedding along the time dimension to obtain $Face_i \in \mathbb{R}^{C_f\times T_v}$.

Encoder-Decoder Separator. As for the speech analysis and separation end, we adopt a U-Net network consisting of an encoder, a fusion module, and a decoder. The encoder consists of multiple convolution and pooling layers. It takes the mixture spectrogram X as input and outputs an audio feature Au_i of dimension $C_a \times T_v$. Following (Xiong et al., 2022), we take an AV-Fusion module to learn joint audio-visual representation. We first concatenate the lip feature Lip_i and face feature $Face_i$ together along the time dimension to get visual feature $V_i \in \mathbb{R}^{C_v\times T_v}$, then feed V_i and Au_i to AV-Fusion module to obtain a enhanced feature. The enhanced feature is delivered into the decoder to predict a mask M_i of input spectrogram, which denotes the proportion of the projection of the prediction on the mixture. Note that the proportion might be positive and negative, as the spectrograms have both positive and negative values.

3.3 FILTER AND RECOVERY NETWORK

Due to the high complexity of speech mixture with multi-speakers (especially above 3 speakers), the separated speech remains two apparent issues: 1) There still exists voices from others; 2) The separated speech is partially missing compared to the ground truth speech. To solve the two problems, we introduce the Filter-Recovery network (FRNet), which takes as input the visual information and separated speech mask for each speaker, and output a more precise target speech mask.

As the Fig. 2 shows, the FRNet consists of a Filter net and a Recovery net. The former utilizes visual features to remove the noise voices from the separated speech, and the latter learns the correlation between all separated speech, in order to extract the missing voices from others.

Since the attention mechanism (Vaswani et al., 2017) enhances some parts of the input data while declining other parts, it is natural for us to adopt attention as the component module. For the sake of fluent description, we define some operations here. Given three tensors $\mathbf{q} \in \mathbb{R}^{D_h \times N_q}$, $\mathbf{k}$, $\mathbf{v} \in \mathbb{R}^{D_h \times N_{kv}}$, we compute the weighted sum of $\mathbf{v}$:

$$[\mathbf{q}',\ \mathbf{k}',\ \mathbf{v}'] = [\mathbf{U}_q\mathbf{q},\ \mathbf{U}_k\mathbf{k},\ \mathbf{U}_v\mathbf{v}], \qquad \mathbf{U}_q, \mathbf{U}_k, \mathbf{U}_v \in \mathbb{R}^{D_h \times D_h}, \tag{1}$$

$$\mathrm{W} = \mathrm{softmax}\left((\mathbf{q}')^\top \mathbf{k}'/\sqrt{\mathrm{D_h}}\right), \qquad \mathrm{W} \in \mathbb{R}^{\mathrm{N_q} \times \mathrm{N_{kv}}}, \tag{2}$$

$$\mathrm{Attn}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \mathbf{v}'\mathrm{W}^\top, \qquad \mathrm{Attn} \in \mathbb{R}^{\mathrm{D_h} \times \mathrm{N_q}}. \tag{3}$$

Filter Net. Due to the entanglement of speech in the mixture, the more speakers in the mixture, the poorer the separation results. Since the visual information is highly correlated with speech, we utilize each speaker's visual knowledge to reduce irrelevant voices in the separation results. Filter Net consists of L basic layers, each consisting of Attn and MLP modules, where the MLP block contains two fully-convolutional layers and a ReLU activation function. In the whole FRNet, we concatenate the real and imaginary parts of M_i obtained by the basic separator for speaker i to form dimension $2F \times T_X$. For the sake of convenience, we still adopt M_i to denote. Beside, to align the dimension, we transform V_i to V_i' of shape $2F \times T_X$ with a convolutional layer.

For layer l in the Filter Net, we feed V_i' and M_i^l into the layer and get the output M_i^{l+1} by the following steps, where $M_i^0 = M_i$:

$$z^{l+1} = \mathrm{LN}\left(\mathrm{Attn}(V_i', M_i^l, M_i^l) + M_i^l\right), \qquad z^{l+1} \in \mathbb{R}^{2F \times T_X}, \tag{4}$$

$$M_i^{l+1} = \mathrm{LN}\left(\mathrm{MLP}(z^{l+1}) + z^{l+1}\right), \qquad M_i^{l+1} \in \mathbb{R}^{2F \times T_X}, \tag{5}$$

For i-th speaker, the Filter Net output mask $\hat{M}_i$ that the noise speech pieces are removed.

Recovery Net. We design the Recovery Net to pull out the missing voice pieces based on the similarity of $\hat{M}_i$ and others' coarse masks $M_{1,\ldots,i-1,i+1,\ldots,S}$. For convenience, we describe the rearrange process in advance. For speaker i, we stack the $S-1$ coarse masks together:

$$\mathrm{temp} = [M_1; \cdots; M_{i-1}; M_{i+1}; \cdots; M_S], \quad \mathbf{M}_i \xleftarrow{\mathrm{reshape}} \mathrm{temp}, \quad \mathbf{M}_i \in \mathbb{R}^{T_X \times 2F \times (S-1)}, \tag{6}$$

where the $\xleftarrow{\mathrm{reshape}}$ operation rearranges the input tensor to the target dimension.

Because of the uncertainty and complexity of speech patterns (voice loudness changes, speed changes, etc.), there may be some missing parts in the separated speech. In the Recovery Net, for i-th speaker, we adopt the clean speech mask $\hat{M}_i$ as the cue to extract missing speech mask $\bar{M}_i$ by learning its association with the coarse separation masks of other speakers. The Recovery Net consists of L basic layers, which is similar to the decoder of Transformer (Vaswani et al., 2017). Specifically, the output of layer $l+1$ can be conveyed by the following equations:

$$\hat{q}_i^l = \mathrm{LN}(\mathrm{Attn}(q_i^l, q_i^l, q_i^l) + q_i^l), \qquad \bar{q}_i^l \xleftarrow{\mathrm{reshape}} \hat{q}_i^l, \qquad \hat{q}_i^l \in \mathbb{R}^{2F \times T_X}, \bar{q}_i^l \in \mathbb{R}^{T_X \times 2F \times 1}, \tag{7}$$

$$z^l[t] = \mathrm{LN}\left(\mathrm{Attn}\left(\bar{q}_i^l[t], \mathbf{M}_i[t], \mathbf{M}_i[t]\right) + \bar{q}_i^l[t]\right), \qquad z^l[t] \in \mathbb{R}^{2F \times 1}, \tag{8}$$

$$\hat{z}^l[t] = \mathrm{LN}\left(\mathrm{MLP}(z^l[t]) + z^l[t]\right), \qquad \hat{z}^l[t] \in \mathbb{R}^{2F \times 1} \tag{9}$$

$$\hat{z}^l = [\hat{z}^l[0]; ...; \hat{z}^l[t]; ...; \hat{z}^l[T_X]], \quad q_i^{l+1} \xleftarrow{\mathrm{reshape}} \hat{z}^l, \quad \hat{z}^l \in \mathbb{R}^{T_X \times 2F \times 1}, q_i^{l+1} \in \mathbb{R}^{2F \times T_X}, \tag{10}$$

where $q_i^0 = \hat{M}_i, \bar{M}_i = \mathrm{FC}(q_i^L)$. We hold the view that the speech separation issues analysed above do not always manifest to the same degree at different time slots. So in Equ. 8, the attention operations are performed for each tensor individually in each time slice t to recover missing parts.

After obtaining the noise-filtered result $\hat{M}_i$ and the recovered missing part $\bar{M}_i$, we add them up to obtain the clean and complete separation mask:

$$\tilde{M}_i = \hat{M}_i + \bar{M}_i \tag{11}$$

3.4 Loss Function

The output M_i of the basic audio-visual separator and $\tilde{M}_i$ of the FRNet both need to be optimized. We take the scale-invariant signal-to-noise ratio (SI-SNR) loss (Le Roux et al., 2019) as the loss function following (Pan et al., 2022). The predicted masks M_i and $\tilde{M}_i$ are separately multiplied by the mixture spectrogram X to get the separated spectrograms, which are finally transformed by inverse STFT to restore the time-domain speech y_i and $\tilde{y}_i$. Given the predicted speech $\hat{s}$ and the target speech s, we can compute the SI-SNR loss with the following formula:

$$\mathcal{L}_{SI\text{-}SNR}(\hat{s}, s) = -10\log_{10}\left(\frac{\|\frac{\langle \hat{s}, s\rangle s}{\|s\|^2}\|^2}{\|\hat{s} - \frac{\langle \hat{s}, s\rangle s}{\|s\|^2}\|^2}\right). \tag{12}$$

The overall optimizing target for $k\text{-}th$ speaker is as follows:

$$\mathcal{L}_i = \lambda \mathcal{L}_{SI\text{-}SNR}(y_i, x_i) + (1-\lambda)\mathcal{L}_{SI\text{-}SNR}(\tilde{y}_i, x_i), \tag{13}$$

where λ is the factor to control the ratio of the two losses. For a batch of a total N speakers, the training loss is the average of all speakers.

4 Experiments

4.1 Datasets

VoxCeleb2 (Chung et al., 2018). This dataset is organized with the identity labels, where there are 5994 speakers in the training set and another 118 speakers in the test set. It contains more than 1 million samples, and each consisting of an utterance and the synchronized face tracks. Following (Gao & Grauman, 2021), we hold out two videos for each speaker in the original training set and utilize the rest videos as our training set. For the remained videos in the original training set, we randomly sample 7200 videos to build the seen-unseen test set. Our unseen-unheard test set consists of 7200 videos randomly chosen from the original test set. All of the rest of the videos in the Voxceleb2 dataset are used to form our validation set.

Lip Reading Sentences 2&3 (Afouras et al., 2018a;c). LRS2 and LRS3 datasets contain 144k and 151k videos clips from TV programs. To evaluate the transferability of all models, we only train the models on Voxceleb2 and test them on LRS2 and LRS3 datasets without any fine-tuning. For each dataset, our test set consists of 1200 randomly selected videos from the original test set.

Test Benchmark. During test, we thoroughly evaluate the performance on mixtures with various numbers of speakers, i.e., from mixtures with 2 speakers to 5 speakers. Meanwhile, to eliminate the uncertainty of results caused by testing on randomly created mixtures, we construct fixed test sets of mixtures.

Table 1: **Numbers of each kind of mixtures for each test set.**

Test Dataset	# 2-mix	# 3-mix	# 4-mix	# 5-mix
Vox2 nseen-unheard	3600	2400	1800	1440
Vox2 seen-heard	3600	2400	1800	1440
LRS2	600	400	300	240
LRS3	600	400	300	240

For test videos of each dataset, we randomly mix them to form 2-mix, 3-mix, 4-mix and 5-mix test sets without any subjective considerations. Each video appears only once in each mixture set. To be more explicit, we list the amount of each kind of mixture for all test sets in Tab. 1.

4.2 Implementation Details

Data Process. We randomly cut 2.55 seconds long clips from videos and the corresponding audios sampled at 16kHz as the training pairs. For all utterances that make up the mixture, we normalize

the energy of each one to the same, which roughly corresponds to the same loudness of each speech. Then we add up all normalized speech to obtain the mixture. STFT is conducted on the mixture waveform using a Hann window length of 400, hop size of 160, and FFT window size of 512 to output the complex spectrogram of dimension $2 \times 257 \times 256$, which is taken as the input to the U-Net encoder. A randomly selected frame from the video is rescaled to 224×224 and input to the face analysis network. The input to the lip reading network is cropped gray mouth regions of interest (ROIs) of dimension 88×88. We adopt the official implement of 2D face landmark detection (Bulat & Tzimiropoulos, 2017) to detect the mouth landmarks and crop the mouth region.

Training Setting. To achieve great separation results for mixtures containing different numbers of speakers, we separate all types of mixtures into individual speeches simultaneously in each batch. In each batch, the ratio of the number of 2-mix, 3-mix, 4-mix, and 5-mix is 2:1:1:1 and the total number of speakers is 256. As for how to select utterances to create mixtures, some methods (Gao & Grauman, 2021; Rahimi et al., 2022) perform put-back random sampling. Such an approach results in some utterances being sampled multiple times while others being ignored, in which case the model might overfit some samples and underfit others. Instead, at the beginning of each training epoch, we shuffle all samples and pick them one by one to ensure no duplication or omission.

Optimization. We adopt the Adam optimizer to train the network with weight decay of 1e-4 and learning rate 1e-4. We train model for 19 epochs, and drops the learning rate by a factor of 0.1 after epochs 12 and 15. λ in 13 is set to 0.5.

Evaluation. Following previous methods (Afouras et al., 2018b; Ephrat et al., 2018; Gao & Grauman, 2021; Rahimi et al., 2022), we adopt the standard blind source separation metric Signal-to-Distortion-Ratio (SDR) (Vincent et al., 2006), which measures the ratio between the energy of the target signal and of the errors. To further assess the speech quality and intelligibility, we also employ the Perceptual Evaluation of Speech Quality (PESQ) (Rix et al., 2001) metric.

4.3 Comparison with State-of-the-art Methods

Tab. 2 and 3 compare our method to other open-source state-of-the-art models on the test sets. All experiments follow the same test and evaluation protocol. We choose two types of methods for comparison: one is audio-only methods and the other is audio-visual methods.

Audio-Only Methods:

U-Net-AO. By removing the face net and the lip net from the basic audio-visual speech separator and modifying the U-Net, we obtain an audio-only speech separator and train it with permutation invariant loss (PIT).

VoiceFilter (Wang et al., 2018). This method takes the mixture and the reference audio as input to extract corresponding speech from the mixtue. Note that the reference audio is a piece of randomly sampled utterance from the target speaker.

Conv-TasNet (Luo & Mesgarani, 2019). This method adopts a fully-convolutional time-domain audio separation network, which takes the time-domain mixture as input and outputs all speech components simultaneously. It trains the model with the permutation invariant loss.

Audio-Visual Methods:

LAVSE (Chuang et al., 2020). To reduce processing costs, it employs a lightweight framework, where a lip encoder extracts the lip motion feature as the synchronization signal for target speech extraction.

VisualVoice (Gao & Grauman, 2021). The method contains the same visual networks and U-Net as our basic audio-visual separator, except the U-Net simply concatenates the audio and visual features together instead of AV-Fusion. Besides, it contains an additional vocal analysis network. Since two loss functions built on the vocal analysis network in this method are designed for only 2-mix separation, we remove the vocal analysis network for adaptation to our multi-speaker speech separation.

VoViT (Montesinos et al., 2022). The newly proposed method uses a landmark-based graph convolutional network (Yan et al., 2018) to capture the facial motion cues. It adopts an AV spectro-temporal transformer for target speech separation.

Table 2: **Results on Voxceleb2 dataset.** All methods are training in our proposed multi-speaker setting. Our method achieves the best performance for all types of mixtures. In addition, each audio-visual method combined with the FRNet results in a large improvement.

Method	Unseen								Seen							
	SDR (dB)				PESQ				SDR (dB)				PESQ			
	2-mix	3-mix	4-mix	5-mix	2-mix	3-mix	4-mix	5-mix	2-mix	3-mix	4-mix	5-mix	2-mix	3-mix	4-mix	5-mix
Audio-Only																
U-Net-AO	0.06	-2.93	-4.66	-5.87	1.55	1.15	0.98	0.89	0.06	-2.92	-4.65	-5.87	1.59	1.17	1.01	0.94
VoiceFilter Wang et al. (2018)	7.07	4.21	1.98	-0.57	2.39	2.08	1.87	1.22	7.56	4.53	2.12	0.13	2.54	2.27	1.99	1.53
Conv-TasNet Luo & Mesgarani (2019)	9.46	4.57	1.50	-0.62	2.67	2.11	1.78	1.66	9.88	4.76	1.54	-0.65	2.70	2.13	1.79	1.66
Audio-Visual																
LAVSE Chuang et al. (2020)	5.38	1.53	-0.6	-2.13	2.17	1.67	1.43	1.28	5.54	1.45	-0.72	-2.31	2.20	1.67	1.43	1.28
LAVSE + FRNet	8.35	4.52	2.06	0.25	2.57	2.06	1.76	1.56	8.61	4.56	1.97	0.05	2.59	2.06	1.75	1.54
VisualVoice Gao & Grauman (2021)	8.97	4.59	2.05	0.25	2.64	2.12	1.82	1.62	8.61	3.96	1.39	-0.4	2.61	2.05	1.74	1.54
VisualVoice + FRNet	10.64	7.01	4.61	2.76	2.84	2.43	2.15	1.92	10.8	6.91	4.34	2.37	2.86	2.42	2.1	1.86
VoViT Montesinos et al. (2022)	9.62	5.08	2.06	-0.21	2.66	2.12	1.77	1.53	9.91	5.29	2.17	-0.11	2.69	2.15	1.80	1.54
VoViT + FRNet	10.63	6.69	3.78	1.37	2.75	2.28	1.93	1.65	10.95	6.93	3.93	1.48	2.78	2.31	1.95	1.67
BFRNet (ours)	**11.06**	**7.48**	**5.13**	**3.26**	**2.87**	**2.48**	**2.20**	**1.97**	**11.27**	**7.48**	**4.89**	**2.86**	**2.89**	**2.47**	**2.16**	**1.91**

Table 3: **Results on LRS 2&3 datasets.** The models are trained on the Voxceleb2 and validated on LRS2/LRS3 without fine-tuning. BFRNet outperforms all base methods that do not combine FRNet. In some metrics, other methods that combined with FRNet achieve the best performance, which also proves the effectiveness of the proposed FR network.

Method	LRS2								LRS3							
	SDR(dB)				PESQ				SDR(dB)				PESQ			
	2-mix	3-mix	4-mix	5-mix	2-mix	3-mix	4-mix	5-mix	2-mix	3-mix	4-mix	5-mix	2-mix	3-mix	4-mix	5-mix
Audio-Only																
U-Net-AO	0.25	-2.65	-4.26	-5.37	1.51	1.14	0.97	0.90	0.19	-2.70	-4.34	-5.45	1.40	1.03	0.85	0.76
VoiceFilter Wang et al. (2018)	6.98	4.10	1.78	-0.77	2.35	2.01	1.69	1.17	8.02	4.58	2.95	-0.31	2.48	2.08	1.86	1.26
Conv-TasNet Luo & Mesgarani (2019)	9.26	4.16	1.13	-1.10	2.48	1.88	1.51	1.30	10.35	4.68	3.45	-0.58	2.64	1.98	1.79	1.38
Audio-Visual																
LAVSE Chuang et al. (2020)	5.74	1.68	-0.54	-2.16	2.04	1.54	1.30	1.17	6.10	1.80	-0.39	-1.96	2.05	1.50	1.26	1.10
LAVSE + FR	8.72	4.46	1.74	-0.38	2.45	1.88	1.56	1.34	9.42	5.03	2.15	0.18	2.52	1.92	1.57	1.35
VisualVoice Gao & Grauman (2021)	10.31	5.55	2.62	0.56	2.61	2.04	1.68	1.47	11.21	6.09	3.20	1.04	2.72	2.09	1.73	1.47
VisualVoice + FR	11.52	**7.78**	**4.95**	**2.80**	2.78	**2.34**	**1.99**	**1.72**	12.51	**8.47**	**5.73**	**3.50**	2.92	**2.45**	**2.10**	**1.79**
VoViT Montesinos et al. (2022)	10.03	4.66	1.15	-1.48	2.56	1.94	1.55	1.29	10.64	5.37	1.41	-0.88	2.64	1.97	1.51	1.26
VoViT + FR	11.19	6.43	2.75	-0.18	2.67	2.11	1.67	1.39	11.90	7.08	3.34	0.67	2.75	2.16	1.70	1.39
BFRNet (ours)	**11.68**	7.37	4.15	1.86	**2.78**	2.28	1.86	1.58	**12.74**	8.37	5.38	2.80	**2.92**	2.42	2.03	1.69

Besides, we combine all audio-visual methods with our FRNet to verify its generality. We call these methods *+FR.

As seen in Tab. 2, Tab. 3 and Fig. 3, our method achieves the best results in the three datasets compared to the other baseline methods without the FRNet. It is worth mentioning that our method outperforms the best baseline methods by more than 1 dB SDR for every dataset. Besides, the FRNet combined with any audio-visual method can significantly improve performance. For example, the FRNet combined with (Gao & Grauman, 2021) can bring a 1.67 dB SDR improvement on the unseen-unheard 2-mix test set of voxceleb2. The difficulty of separation grows with the number of speakers in the mixture, resulting in the lower metrics. For the voxceleb2 dataset, although the speakers in the seen-heard test set appear in the training set, the results of the seen-heard set are not necessarily higher than that of the unseen-unheard set among some methods. This might be attributed to the speakers' voiceprint features not being critical for speech separation.

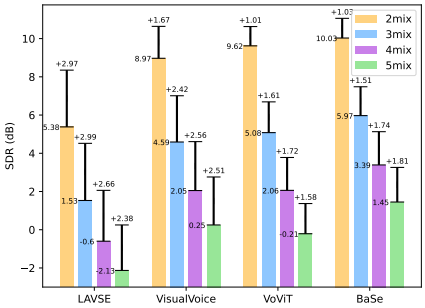

Figure 3: **Visualization of SDR metric of each method on the VoxCeleb2 unseen-unheard test set.** The increment on each bar represents the improvement after combining the FRNet.

4.4 Ablation Studies

We conduct ablation studies on the unseen-unheard and seen-heard test sets of Voxceleb2 and report the average of the results on the two sets.

The effect of modules in FRNet In order to validate the necessity of each module, we conduct experiments to observe the results of removing corresponding modules. When the Filter net is removed, the Recovery net adopts M_i to replace $\hat{M}_i$. When the Recovery net is removed, the output

of Filter net $\hat{M}_i$ is considered as the final output. As seen in Tab. 4, removing any module greatly drops the performance.

Module design of FRNet To validate the rationality of the modules of FRNet, we explore the structural design of the Filter and the Recovery networks. For Filter network, we try to filter voice noise from M_i with itself as the clue instead of simultaneous visual clue V_i, i.e., there is a self-attention in 4. We name this ablation as "filter-sa". As to the recovery network, we design an ablation which replaces the clean mask $\hat{M}_i$ with M_i as the clue to extract the missing voices from $\mathbf{M}_i$ and we name it "recovery-noise". The results are displayed in Tab. 5.

Layers of the FRNet Both the Filter and the Recovery nets are composed of L basic layers. We perform ablation experiments on the number of L. As can be seen in Tab. 6, there is a great improvement for network with $L = 2$ compared to $L = 1$. When the number of layers L increases to 3, the performance improvement is insignificant. To balance performance and efficiency, we adopt the 2-layer FRNet.

Table 4: **Ablation Study on the impact of FR modules.**

Modules		SDR (dB)				PESQ			
Filter	Recovery	2-mix	3-mix	4-mix	5-mix	2-mix	3-mix	4-mix	5-mix
		10.05	5.79	3.12	1.13	2.74	2.24	1.93	1.70
	✓	10.8	6.65	4.51	2.48	2.83	2.38	2.10	1.86
✓		10.3	6.30	3.75	1.81	2.85	2.37	2.05	1.82
✓	✓	11.17	7.48	5.01	3.06	2.88	2.48	2.18	1.94

Table 5: **Ablation Study on module design of FRNet.**

Exps	SDR (dB)				PESQ			
	2-mix	3-mix	4-mix	5-mix	2-mix	3-mix	4-mix	5-mix
Filter-sa	10.91	7.19	4.81	2.86	2.85	2.37	2.86	1.82
Recovery-noise	10.81	7.21	4.82	2.88	2.86	2.43	2.12	1.88
FR	11.17	7.48	5.01	3.06	2.88	2.48	2.18	1.94

Table 6: **Ablation Study on the layers L.**

Layers	SDR (dB)				PESQ			
	2-mix	3-mix	4-mix	5-mix	2-mix	3-mix	4-mix	5-mix
1	10.88	7.17	4.69	2.74	2.83	2.41	2.10	1.86
2	11.17	7.48	5.01	3.06	2.88	2.48	2.18	1.94
3	11.20	7.52	5.17	3.31	2.88	2.50	2.18	1.95

4.5 Qualitative Results

To further verify the effectiveness of the proposed approach, we visualize the spectrograms of results in Fig 4. The first row presents the ground truth, the separation result of our basic separator, and the result of BFRNet, respectively. The second row shows the separation results of the other approaches. The results demonstrate that the separation results of baseline methods are subject to the two issues we claim, including the results of the proposed method without the FRNet. We use black boxes to highlight the missing parts compared to GT, and red boxes to indicate the noisy parts from other speakers. As seen in the Fig., the two problems are greatly suppressed after utilizing the FRNet.

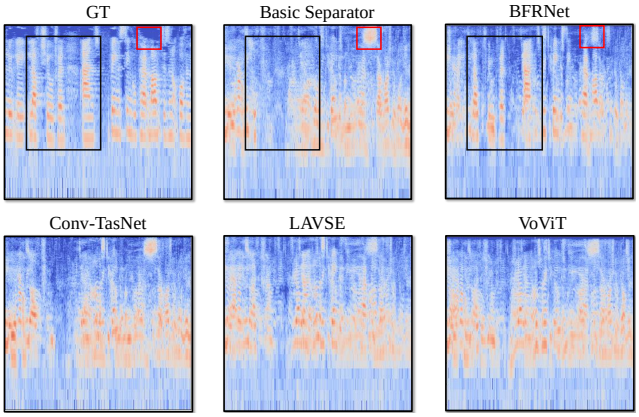

Figure 4: Visualization of the spectrograms of ground truth and predictions.

5 Conclusion

In this paper, we concentrate on the multi-speaker audio-visual speech separation task. We are the first to propose separating mixtures with a variable number of speakers simultaneously during training, and we also provide a fixed test benchmark for fair comparison. There are two significant problems for speech separation especially in multi-speaker setting: part of the voice is missing in the separated speech; the separated speech may still be mixed with others' voice. To deal with this, we propose a Filter and Recovery network to solve these two problems. The filter module filters out other people's voices, and the recovery module compensates for their missing voices. We have conducted various experiments to demonstrate the effectiveness of this module, and its addition to other audio-visual speech separation methods has led to considerable improvements.

REFERENCES

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

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

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018c.

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

Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision*, pp. 208–224. Springer, 2020.

Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*, 2015.

P Chen, Y Zhang, M Tan, H Xiao, D Huang, and C Gan. Generating visually aligned sound from videos. *IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society*, 2020.

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

Shang-Yi Chuang, Yu Tsao, Chen-Chou Lo, and Hsin-Min Wang. Lite audio-visual speech enhancement. *arXiv preprint arXiv:2005.11769*, 2020.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

Soo-Whan Chung, Soyeon Choe, Joon Son Chung, and Hong-Goo Kang. Facefilter: Audio-visual speech separation using still images. *Proc. Interspeech 2020*, pp. 3481–3485, 2020.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (TOG)*, 37(4):1–11, 2018.

Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. *Proc. Interspeech 2018*, pp. 1170–1174, 2018.

Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10478–10487, 2020.

Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3879–3888, 2019.

Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15490–15500. IEEE, 2021.

Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 35–53, 2018.

Mandar Gogate, Ahsan Adeel, Ricard Marxer, Jon Barker, and Amir Hussain. Dnn driven speaker independent audio-visual mask estimation for speech separation. *arXiv preprint arXiv:1808.00060*, 2018.

Sindhu B Hegde, KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Visual speech enhancement without a real visual stream. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1926–1935, 2021.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630. IEEE, 2019.

Jiyoung Lee, Soo-Whan Chung, Sunok Kim, Hong-Goo Kang, and Kwanghoon Sohn. Looking into your speech: Learning cross-modal affinity for audio-visual speech separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1336–1345, 2021.

Shuo Liu, Gil Keren, and Björn Schuller. Single-channel speech separation with auxiliary speaker embeddings. *arXiv preprint arXiv:1906.09997*, 2019.

Rui Lu, Zhiyao Duan, and Changshui Zhang. Listen and look: Audio–visual matching assisted speech source separation. *IEEE Signal Processing Letters*, 25(9):1315–1319, 2018.

Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8): 1256–1266, 2019.

Yi Luo, Zhuo Chen, and Nima Mesgarani. Speaker-independent speech separation with deep attractor network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(4): 787–796, 2018.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, 2018.

Pingchuan Ma, Brais Martinez, Stavros Petridis, and Maja Pantic. Towards practical lipreading with distilled and efficient models. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7608–7612. IEEE, 2021.

Naoki Makishima, Mana Ihori, Akihiko Takashima, Tomohiro Tanaka, Shota Orihashi, and Ryo Masumura. Audio-visual speech separation using cross-modal correspondence loss. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6673–6677. IEEE, 2021.

Juan F Montesinos, Venkatesh S Kadandale, and Gloria Haro. Vovit: Low latency graph-based audio-visual voice separation transformer. *arXiv preprint arXiv:2203.04099*, 2022.

Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 631–648, 2018.

Zexu Pan, Ruijie Tao, Chenglin Xu, and Haizhou Li. Selective listening by synchronizing speech with lips. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1650–1664, 2022.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision.

Akam Rahimi, Triantafyllos Afouras, and Andrew Zisserman. Reading to listen at the cocktail party: Multi-modal speech separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10493–10502, 2022.

Karthik Ramesh, Chao Xing, Wupeng Wang, Dong Wang, and Xiao Chen. Vset: A multimodal transformer for visual speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6658–6662. IEEE, 2021.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pp. 749–752. IEEE, 2001.

Ziqiang Shi, Rujie Liu, and Jiqing Han. Lafurca: Iterative refined speech separation based on context-aware dual-path parallel bi-lstm. *arXiv preprint arXiv:2001.08998*, 2020.

Thanh-Dat Truong, Chi Nhan Duong, Hoang Anh Pham, Bhiksha Raj, Ngan Le, Khoa Luu, et al. The right to talk: An audio-visual transformer approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1105–1114, 2021.

Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. *arXiv preprint arXiv:2011.01143*, 2020.

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

Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006.

Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John Hershey, Rif A Saurous, Ron J Weiss, Ye Jia, and Ignacio Lopez Moreno. Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking. *arXiv preprint arXiv:1810.04826*, 2018.

Junwen Xiong, Peng Zhang, Lei Xie, Wei Huang, Yufei Zha, and Yanning Zhang. Audio-visual speech separation based on joint feature representation with cross-modal attention. *arXiv preprint arXiv:2203.02655*, 2022.

Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 882–891, 2019.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Zengwei Yao, Wenjie Pei, Fanglin Chen, Guangming Lu, and David Zhang. Stepwise-refining speech separation network via fine-grained encoding in high-order latent domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:378–393, 2022.

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

Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 570–586, 2018.