VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency

Ruohan Gao[1,2] Kristen Grauman[1,3]
[1]The University of Texas at Austin [2]Stanford University [3]Facebook AI Research
rhgao@cs.stanford.edu, grauman@fb.com

Abstract

We introduce a new approach for audio-visual speech separation. Given a video, the goal is to extract the speech associated with a face in spite of simultaneous background sounds and/or other human speakers. Whereas existing methods focus on learning the alignment between the speaker's lip movements and the sounds they generate, we propose to leverage the speaker's face appearance as an additional prior to isolate the corresponding vocal qualities they are likely to produce. Our approach jointly learns audio-visual speech separation and cross-modal speaker embeddings from unlabeled video. It yields state-of-the-art results on five benchmark datasets for audio-visual speech separation and enhancement, and generalizes well to challenging real-world videos of diverse scenarios. Our video results and code: http://vision.cs.utexas.edu/projects/VisualVoice/.

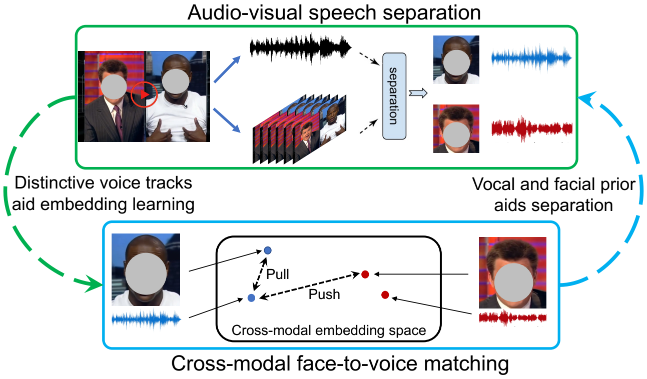

Figure 1: We propose a multi-task learning framework to jointly learn audio-visual speech separation and cross-modal face-voice embeddings. Our approach leverages the complementary cues between lip movements and cross-modal speaker embeddings for speech separation. The embeddings serve as a prior for the voice characteristics that enhances speech separation; the cleaner separated speech in turn produces more distinctive audio embeddings.

1. Introduction

Human speech is rarely observed in a vacuum. Amidst the noisy din of a restaurant, we concentrate to parse the words of our dining partner; watching a heated presidential debate, we disentangle the words of the candidates as they talk over one another; on a Zoom call we listen to a colleague while our children chatter and play a few yards away. Presented with such corrupted and entangled sounds, the human perceptual system draws heavily on visual information to reduce ambiguities in the audio [62] and modulate attention on an active speaker in a busy environment [29]. Automating this process of *speech separation* has many valuable applications, including assistive technology for the hearing impaired, superhuman hearing in a wearable augmented reality device, or better transcription of spoken content in noisy in-the-wild Internet videos.

While early work in automatic speech separation relied solely on the audio stream [53, 78, 18], recent work explores ways to leverage its close connections to the visual stream as well [21, 2, 19, 55, 15]. By analyzing the facial motion in concert with the emitted speech, these methods steer the audio separation module towards the relevant portions of the sound that ought to be separated out from the full audio track. For example, the mouth articulates in different shapes consistent with the phonemes produced in the audio, making it possible to mask a spectrogram for the target human speaker based on audio-visual (AV) consistency. However, solely relying on lip movements can fail when lip motion becomes unreliable, *e.g.*, the mouth region is occluded by the microphone or the speaker turns their head away.

While AV synchronization cues are powerful, we observe that the consistency between the speaker's facial appearance and their voice is also revealing for speech separation. Intuitively, attributes like gender, age, nationality, and body weight are often visible in the face *and* give a prior for what sound qualities (tone, pitch, timbre, basis of articulation) to listen for when trying to separate that person's speech from interfering sounds. For example, female speakers often register in higher frequencies, a heavier person may exhibit a wider range of sound intensities [9], and an American speaker may sound more nasal. The face-voice association, supported by cognitive science studies [11], is today often leveraged for speaker *identification* given the recording of a single speaker [50, 49, 39, 74]. In contrast, the speech *separation* problem demands discovering a cross-modal association in the presence of multiple overlapping sounds.

Our key insight is that these two tasks—cross-modal matching and speech separation—are mutually beneficial. The cleaner the sound separation, the more accurately an embedding can link the voice to a face; the better that embedding, the more distinctive is the prior for the voice characteristics which will in turn aid separation. We thus aim to "visualize" the voice of a person based on how they look to better separate that voice's sound. See Figure 1.

To this end, we propose VISUALVOICE, a multi-task learning framework to jointly learn audio-visual speech separation together with cross-modal speaker embeddings. We introduce a speech separation network that takes video of a human speaker talking in the presence of other sounds (speech or otherwise) and returns the isolated sound track for just their speech. Our network relies on facial appearance, lip motion, and vocal audio to solve the separation task, augmenting the conventional "mix-and-separate" paradigm for audio-visual separation to account for a cross-modal contrastive loss requiring the separated voice to agree with the face. Notably, our approach requires no identity labels and no enrollment of speakers, meaning we can train and test with fully unlabeled video.

Our main contributions are as follows. Firstly, we introduce an audio-visual speech separation framework that leverages complementary cues from facial motion and cross-modal face-voice attributes. Secondly, we devise a novel multi-task framework that successfully learns both separation and cross-modal embeddings in concert. Finally, through experiments on 5 benchmark datasets, we demonstrate state-of-the-art results for audio-visual speech separation and enhancement in challenging scenarios. The embedding learned by our model additionally improves the state of the art for unsupervised cross-modal speaker verification, emphasizing the yet-unexplored synergy of the two tasks.

2. Related Work

Audio-Only Speech Separation. Sound source separation is studied extensively in auditory perception [10]. Speech separation is a special case of sound source separation where the goal is to separate the speech signal of a target speaker from background interference, including non-speech noise [75, 42] and/or interfering speech from other speakers [79, 48]. While early work assumes access to multiple microphones to facilitate separation [53, 78, 18], some methods tackle the "blind" separation problem with monaural audio [65, 68, 69], including recent deep learning approaches [37, 73, 33, 40, 48]. Our work also targets single-channel speech separation, but unlike traditional audio-only methods, we use visual information to guide separation.

Audio-Visual Source Separation. Early methods that leverage audio-visual (AV) cues for source separation use techniques such as mutual information [34, 20], audio-visual independent component analysis [67, 60], and non-negative matrix factorization [58, 24]. Recent deep learning approaches focus on separating musical instruments [26, 81, 77, 80, 22], speech [21, 2, 19, 55, 3, 15], or other sound sources from in-the-wild videos [24, 72]. Current deep AV speech separation methods typically leverage face detection and tracking to guide the separation process [21, 2, 19], while some forgo explicit object detection and process video frames directly [55, 4]. These methods give impressive results by learning the association between lip movements and speech. For robustness to occluded lips, recent work incorporates a (non-visual) identity-specific voice embedding for the audio channel [3]. When only a profile image of the speaker is available, rather than video, learned identity embeddings extracted from a fixed pre-trained network for face images can benefit separation [15].

In contrast to prior AV methods, our model solves for source separation by incorporating both lip motion and cross-modal face-voice attributes. In particular, we propose a multi-task learning framework to jointly learn audio-visual speech separation and cross-modal speaker embeddings. The latter helps learn separation from unlabeled video (i.e., no identity labels, no enrollment of users) by surfacing the sound properties consistent with different facial appearances, as we show in the results.

Cross-Modal Learning with Faces and Voices. There are strong links between how a person's face looks and how their voice sounds. Leveraging this link, cross-modal learning methods explore a range of interesting tasks: face reconstruction from audio [54], talking face generation [82], emotion recognition [5], speaker diarization [13], speech recognition [16], and speaker identification [50, 49, 39, 74, 17]. Unlike any of the above, our work tackles audio-visual speech separation. We jointly learn cross-modal embeddings with the goal of enhancing separation results, with the new insight that hearing voice elements consistent with a face's appearance can help disentangle speech from other overlapping sounds.

Audio-Visual Learning. Apart from source separation, recent inspiring work integrates both audio and visual cues on an array of other tasks including self-supervised representation learning [57, 6, 8, 55, 41, 23], localizing sounds in video frames [7, 66, 71, 36], generating sounds from video [56, 84, 25, 47, 83], and action recognition [38, 27]. Different from all of them, our work leverages the visual cues in faces for the task of audio-visual speech separation.

3. Approach

Our goal is to perform audio-visual speech separation. We first formally define our problem (Sec. 3.1); then we present our audio-visual speech separation network (Sec. 3.2); next we introduce how we learn audio-visual speech separation and cross-modal face-voice embeddings in a multi-task learning framework (Sec. 3.3); finally we present our training criteria and inference procedures (Sec. 3.4).

3.1. Problem Formulation

Given a video V with multiple speakers, we denote $x(t) = \sum_{k=1}^{K} s_k(t)$ as the observed single-channel linear mixture of the voices for these K speakers, where $s_k(t)$ are time-discrete signals responsible for each speaker. Our goal in audio-visual speech separation is to separate the sound $s_k(t)$ for each speaker from $x(t)$ by leveraging the visual cues in the video. For simplicity we describe the sources as speakers throughout, but note that the mixed sound can be something other than speech, as we will demonstrate in results with speech enhancement evaluation.

To generate training examples, we follow the commonly adopted "mix-and-separate" paradigm [19, 55, 2, 81, 26] and generate synthetic audio mixtures by mixing human speech segments. These speech segments are accompanied by the face tracks of the corresponding speakers, which are extracted automatically from "in the wild" videos with background chatter, laughter, pose variation, *etc*.

Suppose we have two speech segments $s_{\mathcal{A}_1}(t)$, $s_{\mathcal{A}_2}(t)$ from video $V_{\mathcal{A}}$ for speaker $\mathcal{A}$, and $s_{\mathcal{B}}(t)$ from video $V_{\mathcal{B}}$ for speaker $\mathcal{B}$.[1] Let $F_{\mathcal{A}_1}, F_{\mathcal{A}_2}, F_{\mathcal{B}}$ denote the face tracks associated with the speech segments $s_{\mathcal{A}_1}(t)$, $s_{\mathcal{A}_2}(t)$, $s_{\mathcal{B}}(t)$, respectively. We create two mixture signals $x_1(t)$ and $x_2(t)$:

$$x_1(t) = s_{\mathcal{A}_1}(t) + s_{\mathcal{B}}(t), \quad x_2(t) = s_{\mathcal{A}_2}(t) + s_{\mathcal{B}}(t). \quad (1)$$

The mixture speech signals are then transformed into complex audio spectrograms X_1 and X_2.

Our training objective is to jointly separate $s_{\mathcal{A}_1}(t)$, $s_{\mathcal{A}_2}(t)$ and $s_{\mathcal{B}}(t)$ for face tracks $F_{\mathcal{A}_1}$, $F_{\mathcal{A}_2}$ and $F_{\mathcal{B}}$ from the two mixed signals $x_1(t)$ and $x_2(t)$. In Sec. 3.3 we present a speaker consistency loss that regularizes the separation process with the two mixtures. To perform separation, we predict complex ideal ratio masks (cIRM) [76] $M_{\mathcal{A}_1}$, $M_{\mathcal{A}_2}$, $M_{\mathcal{B}_1}$ and $M_{\mathcal{B}_2}$ to separate clean speech for the corresponding speakers from X_1 and X_2. Note that we separately predict a mask for speaker $\mathcal{B}$ from each mixture. The predicted spectrograms for the separated speech signals are obtained by complex masking the mixture spectrograms:

$$S_{\mathcal{A}_i} = X_i * M_{\mathcal{A}_i}, \;\; S_{\mathcal{B}_i} = X_i * M_{\mathcal{B}_i}, \;\; i \in \{1, 2\}, \quad (2)$$

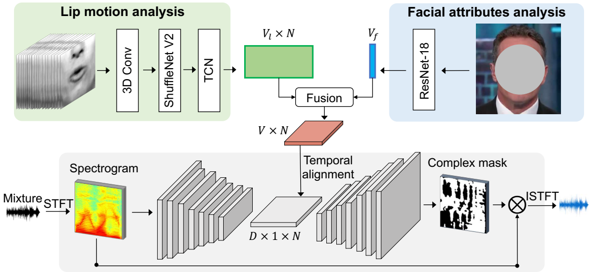

Figure 2: Our audio-visual speech separator network takes a mixed speech signal as input and analyses the lip motion and facial attributes in the face track to separate the portion of sound responsible for the corresponding speaker.

where $*$ indicates complex multiplication. Finally, using the inverse short-time Fourier transform (ISTFT) [30], we reconstruct the separated speech signals.

3.2. Audio-Visual Speech Separator Network

Next we present the architecture of our audio-visual speech separator network, which leverages the complementary visual cues of both lip motion and cross-modal facial attributes. Later in Sec. 3.3 we will introduce our multi-task learning framework to learn both audio-visual speech separation and cross-modal face-voice embeddings, and describe how we jointly separate speech from $x_1(t)$ and $x_2(t)$.

We use the visual cues in the face track to guide the speech separation for each speaker. The visual stream of our network consists of two parts: a lip motion analysis network and a facial attributes analysis network (Figure 2).

Following the state-of-the-art in lip reading [46, 44], the lip motion analysis network takes N mouth regions of interest (ROIs)[2] as input and it consists of a 3D convolutional layer followed by a ShuffleNet v2 [43] network to extract a time-indexed sequence of feature vectors. They are then processed by a temporal convolutional network (TCN) to extract the final lip motion feature map of dimension $V_l \times N$.

For the facial attributes analysis network, we use a ResNet-18 [32] network that takes a single face image randomly sampled from the face track as input to extract a face embedding $\mathbf{i}$ of dimension V_f that encodes the facial attributes of the speaker. We replicate the facial attributes feature along the time dimension to concatenate with the lip motion feature map and obtain a final visual feature of dimension $V \times N$, where $V = V_l + V_f$.

The facial attributes feature represents an identity code whose role is to identify the space of expected frequencies or other audio properties for the speaker's voice, while the role of the lip motion is to isolate the articulated speech specific to that segment. Together they provide complementary

[1] No identity labels are used during training. $s_{\mathcal{A}_1}(t)$ and $s_{\mathcal{A}_2}(t)$ come from the same training video, so we assume they share the same identity.

[2] The ROIs are derived from the face track through facial landmark detection and alignment to a mean reference face. See Supp. for details.

visual cues to guide the speech separation process.

On the audio side, we use a U-Net [64] style network tailored to audio-visual speech separation. It consists of an encoder and a decoder network. The input to the encoder is the complex spectrogram of the mixture signal of dimension $2 \times F \times T$, where F, T are the frequency and time dimensions of the spectrogram. Each time-frequency bin contains the real and imaginary part of the corresponding complex spectrogram value. The input is passed through a series of convolutional layers with frequency pooling layers in between, which reduces the frequency dimension while preserving the time dimension. In the end we obtain an audio feature map of dimension $D \times 1 \times N$, where D is the channel dimension.

We then concatenate the visual and audio features along the channel dimension to generate an audio-visual feature map of dimension $(V + D) \times 1 \times N$. The decoder takes the concatenated audio-visual feature as input. It has symmetric structure with respect to the encoder, where the convolutional layer is replaced by an upconvolutional layer and the frequency pooling layer is replaced by a frequency upsampling layer. Finally, we use a `Tanh` layer followed by a `Scaling` operation on the output feature map to predict a bounded complex mask of the same dimension as the input spectrogram for the speaker.

We build an audio-visual feature map for each speaker in the mixture to separate their respective voices. Alternatively, to build a model tailored to two-speaker speech separation, we concatenate the visual features of both speakers in the mixture with the audio feature to generate an audio-visual feature map of dimension $(2V + D) \times 1 \times N$ and simultaneously separate their voices. This leads to slightly better performance due to the additional context of the other speaker being provided (see Supp. for a comparison), while a model trained with the visual feature of a single speaker can be used in the general case where the number of speakers is unknown at inference time. We use the applicable case in experiments. See Supp. for the network details.

3.3. Cross-Modal Matching for Separation

Next we introduce our multi-task learning framework that simultaneously learns AV speech separation and cross-modal face-voice embeddings. The framework includes several novel loss functions to regularize learning.

Mask prediction loss: As shown in Fig. 3, we predict complex masks $M_{\mathcal{A}_1}$, $M_{\mathcal{A}_2}$, $M_{\mathcal{B}_1}$, $M_{\mathcal{B}_2}$ to separate speech for the corresponding speakers from X_1 and X_2, respectively. We compute the following loss on the predicted complex masks:

$$L_{\textit{mask-prediction}} = \sum_{i \in \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{B}_1, \mathcal{B}_2\}} \|M_i - \mathcal{M}_i\|_2, \quad (3)$$

where $\mathcal{M}_i$ denotes the ground-truth complex masks, which are obtained by taking the complex ratio of the spectrogram of the clean speech to the corresponding mixture speech spectrogram. This loss provides the main supervision to enforce the separation of clean speech.

Cross-modal matching loss: To capture the desired cross-modal facial attributes to guide the separation process, we jointly learn cross-modal face-voice embeddings. The idea aligns with prior work on cross-modal matching [50, 49, 39, 16, 74, 17, 51], but here our goal is audio separation—not person identification—and rather than a single-source input, in our case the audio explicitly contains *multiple* sources.

Similar to the facial attributes analysis network, we use a ResNet-18 network as the vocal attributes analysis network $\Phi(\cdot)$. We extract audio embeddings $\mathbf{a}^{\mathcal{A}_1}$, $\mathbf{a}^{\mathcal{A}_2}$, $\mathbf{a}^{\mathcal{B}_1}$, $\mathbf{a}^{\mathcal{B}_2}$ for each separated speech spectrogram:

$$\mathbf{a}^{\mathcal{A}_i} = \Phi(X_i * M_{\mathcal{A}_i}),\ \mathbf{a}^{\mathcal{B}_i} = \Phi(X_i * M_{\mathcal{B}_i}),\ i \in \{1, 2\}. \quad (4)$$

Let $\mathbf{i}^{\mathcal{A}}$ and $\mathbf{i}^{\mathcal{B}}$ denote the face image embeddings extracted from the facial attributes analysis network for speakers $\mathcal{A}$ and $\mathcal{B}$, respectively. We use the following triplet loss:

$$L_t(\mathbf{a}, \mathbf{i}^+, \mathbf{i}^-) = \max\{0, D(\mathbf{a}, \mathbf{i}^+) - D(\mathbf{a}, \mathbf{i}^-) + \mathtt{m}\}, \quad (5)$$

where $D(\mathbf{a}, \mathbf{i})$ is the cosine distance of the speech embedding and the face image embedding, and $\mathtt{m}$ represents the margin between the two distances. The cross-modal matching loss is defined as follows:

$$\begin{aligned} L_{\textit{cross-modal}} = {} & L_t(\mathbf{a}^{\mathcal{A}_1}, \mathbf{i}^{\mathcal{A}}, \mathbf{i}^{\mathcal{B}}) + L_t(\mathbf{a}^{\mathcal{A}_2}, \mathbf{i}^{\mathcal{A}}, \mathbf{i}^{\mathcal{B}}) \\ & + L_t(\mathbf{a}^{\mathcal{B}_1}, \mathbf{i}^{\mathcal{B}}, \mathbf{i}^{\mathcal{A}}) + L_t(\mathbf{a}^{\mathcal{B}_2}, \mathbf{i}^{\mathcal{B}}, \mathbf{i}^{\mathcal{A}}). \end{aligned} \quad (6)$$

This loss forces the network to learn cross-modal face-voice embeddings such that the distance between the embedding of the separated speech and the face embedding for the corresponding speaker should be smaller than that between the separated speech embedding and the face embedding for the other speaker, by a margin $\mathtt{m}$. It encourages the speech separation network to produce cleaner sounds so that a more accurate speech embedding can be obtained to link the voice to the face. Meanwhile, the better the face embedding, the more distinctive the facial attributes feature can be to guide the speech separation process.

Speaker consistency loss: The audio segments $s_{\mathcal{A}_1}(t)$ and $s_{\mathcal{A}_2}(t)$ come from the same speaker from video $V_{\mathcal{A}}$, so the voice characteristics of $s_{\mathcal{A}_1}(t)$ and $s_{\mathcal{A}_2}(t)$ should be more similar compared to $s_{\mathcal{B}}(t)$. Therefore, the audio embeddings for the separated speech segments for speaker $\mathcal{A}$ should also be more similar compared to that of speaker $\mathcal{B}$. To capture this, we introduce a speaker consistency loss on the audio embeddings of the separated speech:

$$L_{\textit{consistency}} = L_t(\mathbf{a}^{\mathcal{A}_1}, \mathbf{a}^{\mathcal{A}_2}, \mathbf{a}^{\mathcal{B}_1}) + L_t(\mathbf{a}^{\mathcal{A}_1}, \mathbf{a}^{\mathcal{A}_2}, \mathbf{a}^{\mathcal{B}_2}). \quad (7)$$

This loss further regularizes the learning process by jointly separating sounds using the two mixtures.

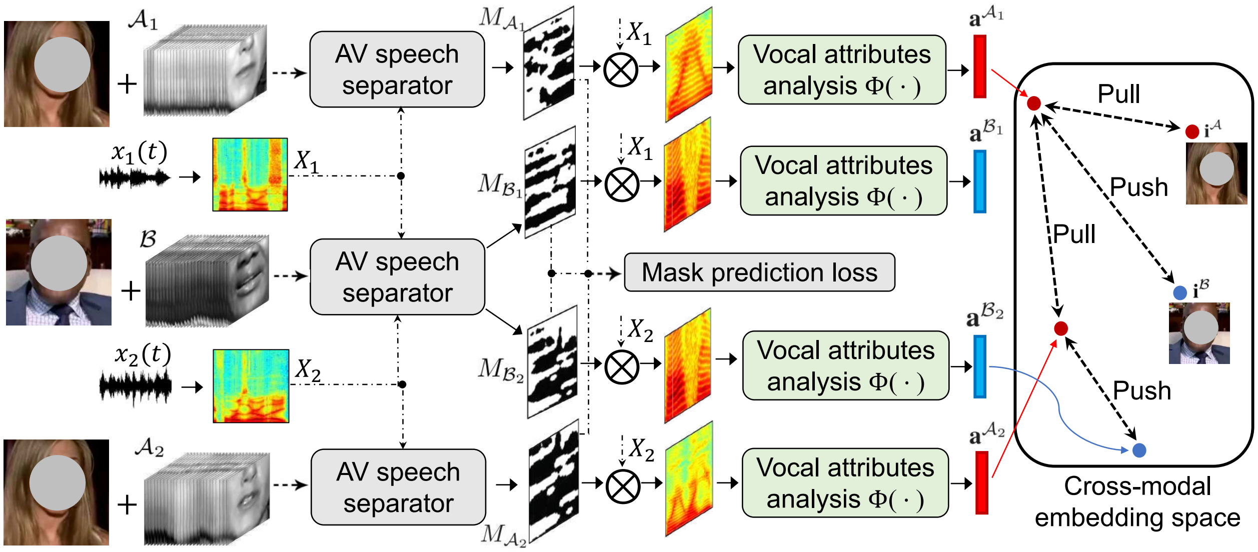
$\mathcal{A}_1$
AV speech separator
$M_{\mathcal{A}_1}$
X_1
Vocal attributes analysis $\Phi(\cdot)$
$\mathbf{a}^{\mathcal{A}_1}$
Pull
$\mathbf{i}^{\mathcal{A}}$
$x_1(t)$
X_1
$M_{\mathcal{B}_1}$
X_1
Vocal attributes analysis $\Phi(\cdot)$
$\mathbf{a}^{\mathcal{B}_1}$
Push
Pull
$\mathcal{B}$
AV speech separator
Mask prediction loss
$\mathbf{i}^{\mathcal{B}}$
$\mathbf{a}^{\mathcal{B}_2}$
$x_2(t)$
X_2
$M_{\mathcal{B}_2}$
X_2
Vocal attributes analysis $\Phi(\cdot)$
Push
$\mathcal{A}_2$
AV speech separator
X_2
$\mathbf{a}^{\mathcal{A}_2}$
Vocal attributes analysis $\Phi(\cdot)$
$M_{\mathcal{A}_2}$
Cross-modal embedding space

Figure 3: Our multi-task learning framework that jointly learns audio-visual speech separation and cross-modal face-voice embeddings. The network is trained by minimizing the combination of the mask prediction loss, the cross-modal matching loss, and the speaker consistency loss defined in Sec. 3.3.

3.4. Training and Inference

The overall objective function for training is as follows:

$$L = L_{\textit{mask-prediction}} + \lambda_1 L_{\textit{cross-modal}} + \lambda_2 L_{\textit{consistency}}, \quad (8)$$

where λ_1 and λ_2 are the weight for the cross-modal matching and speaker consistency losses, respectively. During testing, we first detect faces in the video frames and extract the mouth ROIs for each speaker. For each speaker, we use the mouth ROIs and one face image (a randomly selected frame) as the visual input and predict a complex mask to separate the speech from the mixture signal. We use a sliding window approach to perform separation segment by segment for videos of arbitrary length.

Our audio-visual speech separation network is trained from scratch without using any identity labels, whereas prior methods often assume access to a pre-trained lip reading model [2, 3] or a pre-trained face recognition model [19] that sees millions of labeled faces. Furthermore, we do not need to pre-enroll the voice of the speakers as in [3]. Our framework can train and test with fully unlabeled video.

4. Experiments

Using a total of 6 benchmark datasets, we validate our approach for 1) audio-visual speech separation, 2) speech enhancement (Sec. 4.3), and 3) cross-modal speaker verification (Sec. 4.4).

4.1. Datasets

VoxCeleb2 [14]: This dataset contains over 1 million utterances with the associated face tracks extracted from YouTube videos, with 5,994 identities in the training set and 118 identities in the test set. We hold out two videos for each identity in the training set as our seen-heard test set, and we use 59 identities in the original test set as our validation set and the other 59 identities as our unseen-unheard test set. Note that we make use of the identity labels only for the purpose of making these evaluation splits. During testing, we randomly mix two test clips from different speakers to create the synthetic mixture. This ensures the ground-truth of the separated speech is known for quantitative evaluation, following standard practice [2, 19]. We randomly sample 2,000 test parings each from the seen-heard and unseen-unheard test sets. For speech enhancement experiments, we additionally mix the speech mixture with non-speech audios from AudioSet [28] as background noise during both training and testing. The types of noise include music, laughter, crying, engine, wind, *etc*. See Supp. for details and video examples.

Mandarin [35], TCD-TIMIT [31], CUAVE [59], LRS2[3] [1]: We evaluate on these four standard benchmark datasets to compare our model with a series of state-of-the-art audio-visual speech separation and enhancement methods in Sec. 4.3.2. See Supp. for details.

VoxCeleb1 [52]: This dataset contains over 100,000 utterances for 1,251 celebrities extracted from YouTube videos. We evaluate on this dataset for cross-modal speaker verification in Sec. 4.4. We use the same train/val/test split as in [49] to compare with their reported results.

We are conscious of the risk of biases in data-driven methods for human understanding and have taken measures

[3] All experiments on LRS2 were conducted at UT Austin.

	Reliable lip motion					Unreliable lip motion				
	SDR	SIR	SAR	PESQ	STOI	SDR	SIR	SAR	PESQ	STOI
Audio-Only [79]	7.85	13.7	9.97	2.61	0.82	7.85	13.7	9.97	2.61	0.82
AV-Conv [2]	8.91	14.8	11.2	2.73	0.84	7.23	11.4	9.98	2.51	0.80
Ours (static face)	7.21	12.0	10.6	2.52	0.80	7.21	12.0	10.6	2.52	0.80
Ours (lip motion)	9.95	16.9	11.1	2.80	0.86	7.57	12.7	10.0	2.54	0.81
Ours	**10.2**	**17.2**	**11.3**	**2.83**	**0.87**	**8.53**	**14.3**	**10.4**	**2.64**	**0.84**

Table 1: Audio-visual speech separation results on the VoxCeleb2 dataset. We show the performance separately for testing examples where the lip motion is reliable (left) or unreliable (right). See text for details. Higher is better for all metrics.

to mitigate them; see Supp. for broader impact discussion.

4.2. Implementation Details

Our AV speech separation framework is implemented in PyTorch. For all experiments, we sub-sample the audio at 16kHz, and the input speech segment is 2.55s long. STFT is computed using a Hann window length of 400 with a hop size of 160 and FFT window size of 512. The complex spectrogram X is of dimension $2 \times 257 \times 256$. The input to the lip motion analysis network is $N = 64$ mouth regions of interest (ROIs) of size of 88×88, and the input to the face attributes analysis network is a face image of size 224×224. The lip motion feature is of dimension $V_l \times N$ with $V_l = 512$, $N = 64$. The dimension for both the face and voice embeddings is 128. The entire network is trained using an Adam optimizer with weight decay of 0.0001, batchsize of 128, and starting learning rate set to 1×10^{-4}. λ_1 and λ_2 are both set to 0.01 in Eq. 8. The loss terms are not normalized for scale, so the absolute values of the loss weights do not directly indicate their impact on learning. The margin $\mathtt{m}$ is set to 0.5 for the triplet loss. See Supp. for details of the network architecture and other optimization hyperparameters.

4.3. Results on Audio-Visual Speech Separation

We first evaluate on audio-visual speech separation and compare to a series of state-of-the-art methods [19, 2, 4, 15, 21, 60, 12, 35] and multiple baselines:

- **Audio-Only**: This baseline uses the same architecture as our method except that no visual feature is used to guide the separation process. We use the permutation invariant loss (PIT) [79] to train the network.
- **Ours (lip motion)**: An ablation of our method where only the lip motion analysis network is used to guide the separation process.
- **Ours (static face)**: An ablation of our method where only the facial attributes analysis network is used to guide the separation process.
- **AV-Conv [2]**: A state-of-the-art audio-visual speech separation method that predicts the magnitude and phase of the spectrogram separately through two subnetworks. Because the authors' code is available, we can use it for extensive experiments trained and evaluated on the same data as our method.
- **Ephrat *et al.* [19], Afouras *et al.* [4], Chung *et al.* [15], Gabbay *et al.* [21], Hou *et al.* [35], Casanovas *et al.* [12], Pu *et al.* [60]**: We directly quote results from [19, 4, 15] to compare to a series of prior state-of-the-art methods on standard benchmarks in Sec. 4.3.2.

We evaluate the speech separation results using a series of standard metrics including Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifacts Ratio (SAR) from the mir eval library [61]. We also evaluate using two speech-specific metrics: Perceptual Evaluation of Speech Quality (PESQ) [63], which measures the overall perceptual quality of the separated speech and Short-Time Objective Intelligibility (STOI) [70], which is correlated with the intelligibility of the signal.

4.3.1 Quantitative Results

Table 1 shows the speech separation results on the VoxCeleb2 dataset. We use the visual features of both speakers as input to guide the separation and simultaneously separate their voices. We present results separately for scenarios where the lip motion is reliable and unreliable. For the reliable case, we use the original mouth ROIs extracted automatically from the face tracks; for the unreliable case, we randomly shift the mouth ROI sequences in time by up to 1s and occlude the lip region for up to 1s per segment during both training and testing. These corruptions represent typical video artifacts (e.g., buffering lag) and mouth occlusions. Table 2 shows the speech enhancement results. The setting is the same as Table 1 except that the mixture contains additional background sounds (e.g., laughter, car engine, wind, *etc.*) sampled from AudioSet. The visual feature of only the target speaker is used to guide the separation for speech enhancement experiments.

Tables 1 and 2 show that in both scenarios, our method achieves the best separation results. It outperforms AV-Conv [2] by a good margin. The audio-only baseline benefits from our architecture design, and it has decent performance, though note that unlike AV methods, it cannot as-

	Reliable lip motion					Unreliable lip motion				
	SDR	SIR	SAR	PESQ	STOI	SDR	SIR	SAR	PESQ	STOI
Audio-Only [79]	3.56	10.9	5.71	2.00	0.66	3.56	10.9	5.71	2.00	0.66
AV-Conv [2]	5.32	11.9	7.52	2.20	0.71	3.99	9.43	6.92	2.02	0.67
Ours (static face)	3.48	8.43	6.91	1.96	0.68	3.48	8.43	6.91	1.96	0.68
Ours (lip motion)	6.31	13.3	7.72	2.32	0.76	4.21	9.78	6.85	2.03	0.69
Ours	**6.55**	**13.7**	**7.84**	**2.34**	**0.77**	**4.95**	**11.0**	**7.02**	**2.12**	**0.72**

Table 2: Audio-visual speech enhancement results on the VoxCeleb2 dataset with audios from AudioSet used as non-speech background noise. Higher is better for all metrics.

sign the separated speech to the corresponding speaker. We evaluate both possible matchings and report its best results (to the baseline's advantage). The ablations show that separation with our model is possible purely using one static face image, but it can be difficult especially when the facial attributes alone are not reliable or distinctive enough to guide separation (see Supp.). Lip motion is directly correlated with the speech content and is much more informative for speech separation when reliable. However, the performance of the lip motion-based model significantly drops when the lip motion is unreliable, as often the case in real-world videos. Our VisualVoice approach combines the complementary cues in both the lip motion and the face-voice embedding learned with cross-modal consistency, and thus is less vulnerable to unreliable lip motion. See Supp. for an ablation for the different loss terms.

4.3.2 Comparison to State-of-the-Art Methods

Table 3 compares our method to a series of state-of-the-art methods on AV speech separation and enhancement. We use the same evaluation protocols and the same metrics. Our approach improves the state-of-the-art on each of the five datasets.

Whereas Tables 1 and 2 use the exact same training sources for all methods, here we rely on the authors' reported results [19, 4, 15] in the literature to make comparisons, which draw on different sources. In Table 3a-3c, we evaluate on the Mandarin, TCD-TIMIT and CUAVE datasets using our speaker-independent model trained on VoxCeleb2 to test the cross-dataset generalization capability of our model. Note that this setting is similar to [19], where they also use a speaker-independent model trained on AV-Speech to test on these datasets. In comparison, the other prior methods require training a speaker-dependent model for each speaker in the test dataset. Our model significantly outperforms these methods, despite never seeing the speakers during training. In Table 3d, we train and test on the LRS2 dataset following [4]. Our method consistently outperforms all these prior methods. Notably, in Table 3e, our ablated static face model trained with cross-modal consistency significantly improves the prior static image-based model FaceFilter [15] by 4.68 in SDR. This shows that the cross-modal speaker embeddings learned through our VisualVoice framework can provide sufficient cues for separation, even without using any information on lip movements. This is important for a wide range of scenarios (*e.g.*, online social network platforms) where videos containing lip motion are absent, but a user's profile image is available to use for separation.

	Gabbay *et al.* [21]	Hou *et al.* [35]	Ephrat *et al.* [19]	Ours
PESQ	2.25	2.42	2.50	**2.51**
STOI	–	0.66	0.71	**0.75**
SDR	–	2.80	6.10	**6.69**

(a) Results on Mandarin dataset.

	Gabbay *et al.* [21]	Ephrat *et al.* [19]	Ours
SDR	0.40	4.10	**10.9**
PESQ	2.03	2.42	**2.91**

(b) Results on TCD-TIMIT dataset.

	Casanovas *et al.* [12]	Pu *et al.* [60]	Ephrat *et al.* [19]	Ours
SDR	7.0	6.2	12.6	**13.3**

(c) Results on CUAVE dataset.

	Afouras *et al.* [2]	Afouras *et al.* [4]	Ours
SDR	11.3	10.8	**11.8**
PESQ	3.0	3.0	**3.0**

(d) Results on LRS2 dataset.

	Chung *et al.* [15]	Ours (static face)	Ours
SDR	2.53	7.21	**10.2**

(e) Results on VoxCeleb2 dataset.

Table 3: Comparing to prior state-of-the-art methods on audio-visual speech separation and enhancement. Baseline results are quoted from [19, 4, 15].

4.3.3 Qualitative Results

Real-World Speech Separation. To further test our method's success on real-world videos with mixed speech, we run our model on a variety of test videos in various challenging scenarios including presidential debates, zoom calls, interviews, noisy restaurants, *etc.* Note that these videos lack ground-truth, but can be manually checked for quality as shown in the Supp. video.

Best/Worst Performing Pairs. We illustrate the best and worst performing pairs for speech separation using synthetic pairs for our static face model in the Supp. Pairs that perform best tend to be very different in terms of facial attributes like gender, age, and nationality. Speech separation can be hard if the two mixed identities are visually similar or the facial attributes are hard to obtain from only a static face image due to occlusion or irregular pose.

To further understand when the cross-modal face-voice embeddings help the most, we also compare the per-pair performance of our model with only lip motion and our full model. The pairs with the largest improvement from the cross-modal face-voice embeddings tend to be those that either have very different facial appearances or whose lip motion cues are difficult to extract (*e.g.*, non-frontal views). See Supp. for examples.

4.4. Learned Cross-Modal Embeddings

Our results thus far show how the cross-modal embedding learning enhances speech separation, our primary goal. As a byproduct of our AV speech separation framework, cross-modal embedding learning may also benefit from our model's joint learning. Thus we next evaluate the cross-modal verification task, in which the system must decide if a given face and voice belong to the same person.

To compare with prior cross-modal learning work, we train and evaluate on the VoxCeleb1 dataset and compare with the following baselines: 1) **Learnable-Pins [49]**: A state-of-the-art cross-modal embedding learning method. We directly quote their reported results and follow the same evaluation protocols and data splits to compare with our method; 2) **Random**: Embeddings extracted from a randomly initialized network of the same architecture as our method; 3) **Ours (single-task)**: Our cross-modal embedding network without jointly training for speech separation.

Table 4 shows the results. We use standard metrics for verification. Our cross-modal embedding network alone compares favorably with [49] on seen-heard speakers and generalizes much better to unseen-unheard speakers. When trained with speech separation in a multi-task setting, our method achieves large gains, demonstrating that our idea to jointly train for these two tasks is beneficial to learn more reliable cross-modal face-voice embeddings. Table 4 provides an apples-to-apples comparison, whereas methods following [49] evaluate under several different protocols. Other new loss function designs and the extra supervision used in the other prior work are orthogonal to our idea and could also augment our model.

To *visualize* that our VISUALVOICE framework has indeed learned useful cross-modal face-voice embeddings, Figure 4 shows the t-SNE [45] embeddings of the voices for 15 random speakers from the VoxCeleb1 test set. The embeddings are extracted from our vocal attributes analysis network jointly trained with speech separation. The two sub-figures are color-coded with gender and identity, respectively. Our method's learned voice embeddings tend to cluster speakers of the same cross-modal attributes together despite having access to no identity labels and no attribute labels during training.

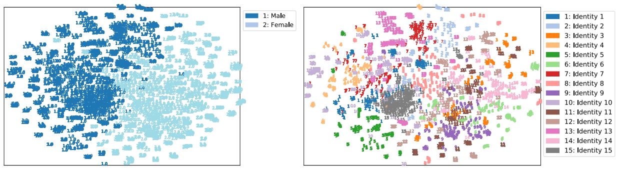

Figure 4: Our learned cross-modal embeddings of voices for 15 speakers from the VoxCeleb1 test set visualized with t-SNE. The two figures are color coded with gender and identity, respectively.

	Seen-Heard		Unseen-Unheard	
	AUC ↑	EER ↓	AUC ↑	EER ↓
Random	50.8	49.6	49.7	50.1
Learnable Pins [49]	73.8	34.1	63.5	39.2
Ours (single-task)	75.0	32.2	72.4	34.7
Ours	**84.9**	**23.6**	**74.2**	**32.3**

Table 4: Cross-modal verification results on the VoxCeleb1 dataset. ↓ lower better, ↑ higher better.

5. Conclusion

We presented an audio-visual speech separation framework that simultaneously learns cross-modal speaker embeddings and speech separation in a multi-task setting. Our VISUALVOICE approach exploits the complementary cues between the lip motion and cross-modal facial attributes. It achieves state-of-the-art results on audio-visual speech separation and generalizes well to challenging real-world videos. Our design for the cross-modal matching and speaker consistency losses is not restricted to the speech separation task, and can be potentially useful for other audio-visual applications, such as learning intermediate features for speaker identification and sound source localization. As future work, we plan to explicitly model the fine-grained cross-modal attributes of faces and voices, and leverage them to further enhance speech separation.

Acknowledgements: Thanks to Ariel Ephrat, Triantafyllos Afouras, and Soo-Whan Chung for help with experiments setup and to Lorenzo Torresani, Laurens van der Maaten, and Sagnik Majumder for feedback on paper drafts. RG is supported by a Google PhD Fellowship and a Adobe Research Fellowship. UT Austin is supported in part by NSF IIS-1514118 and the IFML NSF AI Institute.

References

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. *TPAMI*, 2018.

[2] T. Afouras, J. S. Chung, and A. Zisserman. The conversation: Deep audio-visual speech enhancement. In *INTERSPEECH*, 2018.

[3] T. Afouras, J. S. Chung, and A. Zisserman. My lips are concealed: Audio-visual speech enhancement through obstructions. In *ICASSP*, 2019.

[4] T. Afouras, A. Owens, J.-S. Chung, and A. Zisserman. Self-supervised learning of audio-visual objects from video. In *ECCV*, 2020.

[5] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman. Emotion recognition in speech using cross-modal transfer in the wild. In *ACMMM*, 2018.

[6] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *ICCV*, 2017.

[7] R. Arandjelović and A. Zisserman. Objects that sound. In *ECCV*, 2018.

[8] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *NeurIPS*, 2016.

[9] B. Barsties, R. Verfaillie, N. Roy, and Y. Maryn. Do body mass index and fat volume influence vocal quality, phonatory range, and aerodynamics in females? In *CoDAS*, 2013.

[10] A. S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. 1994.

[11] V. Bruce and A. Young. Understanding face recognition. *British journal of psychology*, 1986.

[12] A. L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval. Blind audiovisual source separation based on sparse redundant representations. *IEEE Transactions on Multimedia*, 2010.

[13] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman. Spot the conversation: speaker diarisation in the wild. In *INTERSPEECH*, 2020.

[14] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

[15] S.-W. Chung, S. Choe, J. S. Chung, and H.-G. Kang. Facefilter: Audio-visual speech separation using still images. In *INTERSPEECH*, 2020.

[16] S.-W. Chung, J. S. Chung, and H.-G. Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *ICASSP*, 2019.

[17] S.-W. Chung, H. G. Kang, and J. S. Chung. Seeing voices and hearing voices: learning discriminative embeddings using cross-modal self-supervision. In *INTERSPEECH*, 2020.

[18] N. Q. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[19] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. In *SIGGRAPH*, 2018.

[20] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. A. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *NeurIPS*, 2001.

[21] A. Gabbay, A. Shamir, and S. Peleg. Visual speech enhancement. In *INTERSPEECH*, 2018.

[22] C. Gan, D. Huang, H. Zhao, J. B. Tenenbaum, and A. Torralba. Music gesture for visual sound separation. In *CVPR*, 2020.

[23] R. Gao, C. Chen, Z. Al-Halab, C. Schissler, and K. Grauman. Visualechoes: Spatial image representation learning through echolocation. In *ECCV*, 2020.

[24] R. Gao, R. Feris, and K. Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018.

[25] R. Gao and K. Grauman. 2.5d visual sound. In *CVPR*, 2019.

[26] R. Gao and K. Grauman. Co-separating sounds of visual objects. In *ICCV*, 2019.

[27] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, 2020.

[28] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017.

[29] 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". *Journal of Neuroscience*, 2013.

[30] D. Griffin and J. Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984.

[31] N. Harte and E. Gillen. Tcd-timit: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 2015.

[32] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[33] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *ICASSP*, 2016.

[34] J. R. Hershey and J. R. Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *NeurIPS*, 2000.

[35] 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 Transactions on Emerging Topics in Computational Intelligence*, 2018.

[36] D. Hu, R. Qian, M. Jiang, X. Tan, S. Wen, E. Ding, W. Lin, and D. Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. In *NeurIPS*, 2020.

[37] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Deep learning for monaural speech separation. In *ICASSP*, 2014.

[38] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*, 2019.

[39] C. Kim, H. V. Shin, T.-H. Oh, A. Kaspar, M. Elgharib, and W. Matusik. On learning associations of faces and voices. In *ACCV*, 2018.

[40] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

[41] B. Korbar, D. Tran, and L. Torresani. Co-training of audio and video representations from self-supervised temporal synchronization. In *NeurIPS*, 2018.

[42] A. Kumar and D. Florencio. Speech enhancement in multiple-noise conditions using deep neural networks. In *INTERSPEECH*, 2016.

[43] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, 2018.

[44] P. Ma, B. Martinez, S. Petridis, and M. Pantic. Towards practical lipreading with distilled and efficient models. *arXiv preprint arXiv:2007.06504*, 2020.

[45] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 2008.

[46] B. Martinez, P. Ma, S. Petridis, and M. Pantic. Lipreading using temporal convolutional networks. In *ICASSP*, 2020.

[47] P. Morgado, N. Vasconcelos, T. Langlois, and O. Wang. Self-supervised generation of spatial audio for 360° video. In *NeurIPS*, 2018.

[48] E. Nachmani, Y. Adi, and L. Wolf. Voice separation with an unknown number of multiple speakers. In *ICML*, 2020.

[49] A. Nagrani, S. Albanie, and A. Zisserman. Learnable pins: Cross-modal embeddings for person identity. In *ECCV*, 2018.

[50] A. Nagrani, S. Albanie, and A. Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *CVPR*, 2018.

[51] A. Nagrani, J. S. Chung, S. Albanie, and A. Zisserman. Disentangled speech embeddings using cross-modal self-supervision. In *ICASSP*, 2020.

[52] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017.

[53] K. Nakadai, K.-i. Hidai, H. G. Okuno, and H. Kitano. Real-time speaker localization and speech separation by audio-visual integration. In *ICRA*, 2002.

[54] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik. Speech2face: Learning the face behind a voice. In *CVPR*, 2019.

[55] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018.

[56] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *CVPR*, 2016.

[57] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, 2016.

[58] S. Parekh, S. Essid, A. Ozerov, N. Q. Duong, P. Pérez, and G. Richard. Motion informed audio source separation. In *ICASSP*, 2017.

[59] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy. Cuave: A new audio-visual database for multimodal human-computer interface research. In *IEEE International conference on acoustics, speech, and signal processing*, 2002.

[60] J. Pu, Y. Panagakis, S. Petridis, and M. Pantic. Audio-visual object localization and separation using low-rank and sparsity. In *ICASSP*, 2017.

[61] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel. mir_eval: A transparent implementation of common mir metrics. In *ISMIR*, 2014.

[62] T. Rahne, M. Böckmann, H. von Specht, and E. S. Sussman. Visual cues can modulate integration and segregation of objects in auditory scene analysis. *Brain research*, 2007.

[63] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *ICASSP*, 2001.

[64] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015.

[65] S. T. Roweis. One microphone source separation. In *NeurIPS*, 2001.

[66] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018.

[67] P. Smaragdis and M. Casey. Audio/visual independent components. In *International Conference on Independent Component Analysis and Signal Separation*, 2003.

[68] P. Smaragdis, B. Raj, and M. Shashanka. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *International Conference on Independent Component Analysis and Signal Separation*, 2007.

[69] M. Spiertz and G. Volker. Source-filter based clustering for monaural blind source separation. In *12th International Conference on Digital Audio Effects*, 2009.

[70] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

[71] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018.

[72] E. Tzinis, S. Wisdom, A. Jansen, S. Hershey, T. Remez, D. P. Ellis, and J. R. Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. In *ICLR*, 2021.

[73] Y. Wang, A. Narayanan, and D. Wang. On training targets for supervised speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 2014.

[74] Y. Wen, M. A. Ismail, W. Liu, B. Raj, and R. Singh. Disjoint mapping network for cross-modal matching of voices and faces. In *ICLR*, 2019.

[75] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *International Conference on Latent Variable Analysis and Signal Separation*, 2015.

[76] D. S. Williamson, Y. Wang, and D. Wang. Complex ratio masking for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 2015.

[77] X. Xu, B. Dai, and D. Lin. Recursive visual sound separation using minus-plus net. In *ICCV*, 2019.

[78] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on signal processing*, 2004.

[79] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *ICASSP*, 2017.

[80] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba. The sound of motions. In *ICCV*, 2019.

[81] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In *ECCV*, 2018.

[82] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, 2019.

[83] H. Zhou, X. Xu, D. Lin, X. Wang, and Z. Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *ECCV*, 2020.

[84] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*, 2018.