VISUALVOICE: Audio-Visual Speech Separation with Cross-Modal Consistency (Supplementary Materials)

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

The supplementary materials for [7] consist of:

A. Broader impact.

B. Supplementary video.

C. General single-speaker model VS. dedicated two-speaker model.

D. Separation results for seen-heard speakers VS. unseen-unheard speakers.

E. Ablation study on the loss terms.

F. Best/worst performing pairs.

G. Network and implementation details.

H. Dataset details.

A. Broader Impact

We are conscious of possible undesirable effects that can arise when working with data-driven approaches to human understanding in images and video. Specifically, a method's training data will guide the extent to which the model can generalize well and fairly to arbitrary inputs. To mitigate risks in this regard, we have taken several steps. First, we learn the cross-modal face-voice embeddings from the VoxCeleb2 dataset, which to our knowledge is the largest relevant available dataset with over 6,000 speakers spanning a range of different ethnicities, accents, professions, and ages. Second, we examine the speech separation results separately for a seen-heard test set and an unseen-unheard test set from VoxCeleb2. The results show our method achieves similar performance for seen-heard and unseen-unheard speakers. This shows that our model generalizes well to unseen-unheard speakers in VoxCeleb2 and is not limited to handling seen-heard speakers in the training data.

Finally, the output of our model consists of voices separated from the the original test video—in terms of masking the input spectrogram—as opposed to being generated or machine synthesized. This is important because it means our model is not free to hallucinate arbitrary voice sounds for the input speakers, e.g., the model cannot artificially conjure sounds or words often associated with training faces that happen to look like the input speaker unless they are consistent with the input sounds. Indeed, as shown in results in the main paper, lip motion continues to play a key role during speech separation, isolating words based on their visual agreement with what was physically spoken. The learned cross-modal face-voice embeddings complement lip motion cues to further enhance the separation results, particularly when lip motion is harder to read or the two input faces are very different in appearance.

To further explore the model's performance as a function of a person's race, gender, ethnicity, or other identity data, it would be interesting to sort results by the relative impact of our model along each dimension independently. However, existing meta-data does not permit this study (VoxCeleb2 only provides identity and gender labels). We hope to analyze the per-category performance of our models for these cross-modal speaker attributes when datasets as such meta-data and/or new dataset efforts become available.

B. Supplementary Video

In the supplementary video[1], we show example separation results. We first show audio-visual speech separation and enhancement results on **real-world test videos** of multiple speakers in various challenging scenarios including presidential debates, zoom calls, interviews, and noisy restaurants. Next, we show some qualitative results on synthetic mixtures from the VoxCeleb2 dataset and compare with the AV-Conv [2] baseline and our static face based model. Finally, we show some failure cases of our model.

C. General Single-Speaker Model VS. Dedicated Two-Speaker Model

As mentioned in Sec. 3.2 in the main paper, we can either build an audio-visual feature map for each speaker in

[1] http://vision.cs.utexas.edu/projects/VisualVoice/

the mixture to separate their respective voices or build a model tailored to two-speaker speech separation. For the former case, the model can be generally applicable in testing scenarios where the number of speakers is unknown, while the latter only applies to two-speaker speech separation but can benefit from the contextual visual information of the other speaker in the mixture. For audio-visual speech separation on VoxCeleb2, the model tailored to two-speaker speech separation achieves SDR of 10.2, while the general single-speaker model achieves SDR of 9.88. Most of our experiments are on two-speaker speech mixtures due to its wide applications in real-world. Note that a model tailored to more than two speakers can be similarly built by concatenating the visual features of all the speakers in the mixture.

D. Separation Results for Seen-Heard Speakers VS. Unseen-Unheard Speakers

Table 1 shows the speech separation results separately for the seen-heard test set and the unseen-unheard test set on the VoxCeleb2 dataset. We can see that the methods purely based on lip motion tend to have similar performance for seen-heard and unseen-unheard speakers. For models that reply on facial appearance, the separation performance is slightly better on the seen-heard test set because the learned cross-model face-voice embeddings are more reliable for seen-heard speakers. Our method leverages both the lip motion and the cross-modal facial attributes, generalizing well to unseen-unheard speakers.

To further verify that it is beneficial to disentangle lip motion and cross-modal facial attributes, we show a baseline called Face-Track, which directly processes the full face track to extract visual features similar to prior work [4]. The gain of our method demonstrates that it is helpful to focus specifically on the lip regions (mouth ROIs) when analyzing the lip movements for separation, and the cross-modal face-voice embeddings learned through our multi-task learning framework can better exploit the complementary facial appearance cues to enhance separation.

E. Ablation Study on Loss Terms

We perform an ablation study to examine the impact of the key components of our VISUALVOICE framework. We empirically set λ_1 and λ_2 by tuning on validation data. Note that the loss terms are not normalized for similar scales, so the absolute values of the loss weights do not directly indicate their impact on learning. Table 2 compares the speech separation performance of several variants of our model on the VoxCeleb2 dataset. We compare our model with one variant that only uses the mask-prediction loss; one variant without using the cross-modal matching loss; one variant without using the speaker consistency loss. We can see that the mask-prediction loss provides the main supervision for learning speech separation. Together with the cross-modal matching loss and speaker consistency loss, we achieve the best results both for our full model and our static face based model. It's possible that a more rigorous hyperparameter search of the loss weights can lead to better performance.

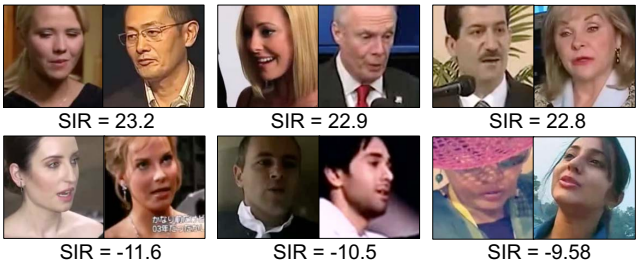

Figure 1: Qualitative examples of the best performing pairs (first row) and worst performing pairs (second row) for our static-face image based model.

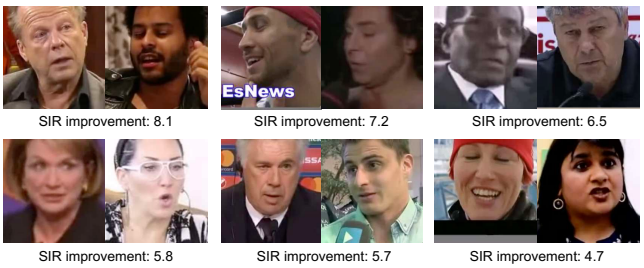

Figure 2: Qualitative examples of the pairs with the largest improvement from cross-modal face-voice embeddings.

F. Best/Worst Performing Pairs

Fig. 1 illustrates the best and worst performing pairs for speech separation using synthetic pairs for our static face model. Pairs that perform the best tend to be very different in terms of facial attributes like gender, age, and nationality (first row). 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 (second row).

To further understand when the cross-modal face-voice embeddings help the most, we compare the per-pair performance of our model with only lip motion and our full model in Fig. 2. 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).

G. Network and Implementation Details

Our audio-visual speech separator network uses the visual cues in the face track to guide the speech separation for each speaker. The visual stream of our network con-

	Seen-Heard					Unseen Unheard				
	SDR	SIR	SAR	PESQ	STOI	SDR	SIR	SAR	PESQ	STOI
Audio-Only [14]	8.00	13.9	10.1	2.63	0.82	7.70	13.6	9.85	2.59	0.82
AV-Conv [2]	8.94	14.8	11.2	2.73	0.84	8.89	14.8	11.1	2.72	0.84
Face-Track	9.28	15.8	11.2	2.76	0.85	9.20	15.6	11.0	2.75	0.84
Ours (static face)	7.37	12.2	10.7	2.54	0.80	7.06	11.8	10.6	2.49	0.80
Ours (lip motion)	9.96	16.9	11.2	2.81	0.86	9.94	17.0	11.1	2.80	0.87
Ours	**10.3**	**17.2**	**11.4**	**2.83**	**0.87**	**10.1**	**17.2**	**11.2**	**2.82**	**0.87**

Table 1: Audio-visual speech separation results on the VoxCeleb2 dataset. We show the performance separately for seen-heard test set and unseen-unheard test set. Higher is better for all metrics.

	Ours (static face)					Ours				
	SDR	SIR	SAR	PESQ	STOI	SDR	SIR	SAR	PESQ	STOI
Only mask prediction loss	6.69	10.9	10.3	2.43	0.75	9.81	16.4	10.9	2.76	0.82
Without cross-modal matching loss	6.95	11.4	10.4	2.48	0.77	9.93	16.7	11.1	2.79	0.84
Without speaker consistency loss	7.10	11.7	10.4	2.50	0.79	10.0	17.0	11.1	2.81	0.86
All losses	**7.21**	**12.0**	**10.6**	**2.52**	**0.80**	**10.2**	**17.2**	**11.3**	**2.83**	**0.87**

Table 2: Ablation study on the loss terms. Higher is better for all metrics.

sists of two parts: a lip motion analysis network and a facial attributes analysis network.

The lip motion analysis network takes 64 mouth regions of interest (ROIs) as input. To obtain the ROIs from the face track, we use an SFD face detector [15] to detect 68 facial landmarks. Following [11], the faces are then aligned to a mean reference face to remove differences related to rotation and scale using a similarity transformation. A 96×96 ROI is cropped once the center of the mouth is located for each frame. During training, we use random horizontal flippling, random croping of size of 88×88. During testing, the center patch is used. We use gray-scale images for the mouth ROIs as the input to the lip motion analysis network, which consists of a 3D convolutional layer with kernel $5 \times 7 \times 7$ followed by a ShuffleNet-V2 network and a temporal convolutional network. The temporal convolutional network takes the time-indexed sequence of feature vectors extracted from the ShuffleNet-V2 network, and maps it into another such sequency through the use of a 1D temporal convolution. See [11] for details. Finally, we obtain a feature map of lip motion of dimension 512×64.

The face attributes analysis network is a ResNet-18 network that takes an image of size 224×224 as input, and the feature map after the final pooling layer is downsampled to dimension of $V_f = 128$ through a fully-connected layer. 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 640×64.

On the audio side, we use a U-Net [13] style network similar to [6, 16, 5], but here we tailor the network 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 $2 \times 257 \times 256$. The input is first passed through 2 convolutional layers (kernel size = 4, stride = 2, padding = 1) that downsample the frequency and time dimension until the time dimension is equal to $N = 64$. Then we use 6 conv-blocks that each consist of two convolutional layers (kernel size = 3, stride = 1, padding = 1) followed by a frequency pooling layer. The first two convolutional layers in each conv-block preserve the spatial dimension and the frequency pooling layer after it reduces the frequency dimension by a factor of 2 while preserving the time dimension. Next, we use a `Tanh` layer to map the output feature map values to the range of [-1, 1]. Because the real and imaginary parts of the ground-truth complex mask typically lie between -5 and 5, we further use a `Scaling` operation to scale the output by 5. Finally, we obtain a bounded predicted complex mask of the same dimension as the input spectrogram for the speaker. In source separation tasks, spectrogram masks have proven better than alternatives such as direct prediction of spectrograms or raw waveforms [4, 6]. The 1D time series audio signal varies widely with small distortions, and perceptual information is difficult to extract directly. STFT separates the frequencies and amplitudes, generating a spectrogram that is more structured and can be analyzed similarly to an image using a CNN. The mask makes the prediction target bounded and further regularizes the learning process.

The voice attributes analysis network has the same configuration as the face attributes analysis network except the first layer of the ResNet-18 network takes the separated complex spectrogram of dimension $2 \times 256 \times 256$ as input. Similarly, the feature map after the final pooling

layer is downsampled to dimension of 128 through a fully-connected layer.

H. Dataset Details

Our experiments on audio-visual speech separation and enhancement are mainly on the VoxCeleb2 dataset due to the availability of the pre-computed face tracks and identity labels, which allow us to explicitly test for speaker-independent performance. For audio-visual enhancement experiments, we additionally mix the speech mixture with non-speech audios in AudioSet [8] (excluding the speech category from the ontology) as background noise. We mix with audios from the official training split / evaluation split during training and testing, respectively.

We also evaluate on four standard benchmark datasets below to compare our model with a series of state-of-the-art audio-visual speech separation and enhancement methods in Sec.4.3.2 in the main paper: 1) **Mandarin [10]** is an audio-visual dataset prepared by Hou *et al*. [10] containing video recordings of Mandarin sentences spoken by a native speaker. Each sentence is approximately 3-4 seconds and contains 10 Chinese characters with the phonemes designed to distribute equally. We use the official test set that contains 40 clean utterances, mixed with the 10 noise types (*e.g*., crying, music, offscreen speakers, *etc*.) at 5 dB, 0 dB, and -5 dB SIRs and car engine ambient noise; 2) **TCD-TIMIT [9]** consists of 59 volunteer speakers with around 200 videos each. The speakers are recorded saying various sentences from the TIMIT dataset. We mix every clip with another clip of a random speaker to evaluate the speech separation performance; 3) **CUAVE [12]** is an audio-visual speech database consisting of videos of connected and continuous digits spoken in different situations by various speakers. The ground-truth audio for one speaker in the mixture is avaialble to evaluate the separation performance; 4) **LRS2 [1]** is a dataset for lip reading that consists of 224 hours of videos long with pre-computed face tracks of the speakers. We follow the setting of [3] and evaluate our model using only videos that are between 2 - 5 seconds long.

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

[4] 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.

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

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

[7] R. Gao and K. Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *CVPR*, 2021.

[8] 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.

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S3fd: Single shot scale-invariant face detector. In *ICCV*, 2017.

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