Supplementary material for: Self-supervised learning of audio-visual objects from video

Anonymous ECCV submission

Paper ID 2957

1 Video

Please watch our video `eccv-2957-supplementary.mp4` to see results from our model. Thank you.

We have also included a second video, `eccv-2957-pipeline.mp4`, showing in more detail the steps our it undertakes to extract audio-visual objects from a video.

2 Non-human speakers experiments

In this section, we provide more details about the dataset and evaluation on videos of non-human speakers.

Unlabeled training sets. As training data for the non-human speakers experiments we used episodes of the *The Simpsons* and *Sesame Street* shows found on YouTube. The training sets we collected consist of approximately 48 hours for *The Simpsons* (from seasons 11 to 31) and 53 hours of video for *Sesame Street* (taken from playlists of the official YouTube channel for several episode collections, as well as characters Elmo, Cookie Monster, Bert, Ernie, Abby, Grover, Rosita, Big Bird, Oscar, The Count, Kermit, and Zoe). The only processing we perform on the original clips is splitting them into scenes by using an off-the-shelf package, so as to avoid clips with scene transitions. We emphasize that no other preprocessing such as Voice Activity Detection or filtering out of title frames etc. was performed; we trained our models in this raw, potentially noisy data. Note that the training sets are entirely unlabeled; we train our models on them using self-supervision.

We observed that clearly visible talking heads appear much more often in Simpsons episodes, compared to Sesame Street. The latter also contains actual humans. Moreover, the puppets used for the show are manually moved and there is only approximate correspondence in the timing of movement with the corresponding speech, whereas the head and mouth animations in Simpsons are temporally aligned with the speech. All of these factors make the training on examples from Simpsons significantly easier.

Annotated test sets. To create the two test sets summarised in Table 3a of the main paper, we manually annotated clips from held-out subsets, using the VIA annotator [6]. There is no episode overlap between the training and test sets. We asked human labelers (three computer vision researchers) to annotate the active speaker in randomly chosen

clips, including bounding boxes around the heads, in a small number of frames per clip. We note that in this case the character is not physically generating the sound; our goal is to reproduce these human judgements about which is the speaker (e.g., the ventriloquism effect for puppets). For the *multi-speaker* we also include negative samples that can be either non-speaking faces (those are the majority and we believe harder negatives) or frames not involving any characters, title/credit sequences, etc. The ratio of positive and negative frames is approximately 1:1.

We include both *single-head* examples where only one speaker is in view (for a comparison to face detection methods on the localization task), and *multi-head* with multiple potential speakers for active speaker detection.

Evaluation protocol. We trained separate models for the *Simpsons* and the *Sesame Street* experiments, initialized from the best performing models trained on LRS2.

Using off-the-shelf detectors and SyncNet. Face detectors are a key component of many speech understanding systems, such as active speaker detection pipelines [4], as well as for curating speech datasets [2, 4, 7, 9]. Here we investigate in more depth whether these off-the-shelf methods would also apply to non-human speakers in our dataset.

As described in Table 3b of the main paper, we confirmed that an off-the-shelf face detector, RetinaFace [5], obtains poor average precision on these videos. In practice, correct face detections are poorly ranked and inconsistent frame to frame; thus it is difficult to obtain them without introducing large numbers of false positives. This behavior is expected, since these models have been trained on a different domain (human faces). Here we provide qualitative examples of the detector's behavior (please see the video results), and a comparison to our self-supervised model's results.

Likewise, we also tried using SyncNet [4] as a baseline for the active speaker detection (ASD) task. However, running this system out-of-the box failed. This is because ASD with SyncNet is based on a multi-model pipeline: first face detections are extracted with an SSD [8] detector and heuristically stitched into face tracks; SyncNet is then run to ASD on top of these face tracks. Since the face detector very rarely returns correct detections, producing virtually no face tracks, the model's later steps were consistently incorrect.

3 Extra non-human source-separation experiments.

We also trained models to perform source-separation and speech enhancement on the *Simpsons* data. For this we created synthetic videos with the mix-and-separate procedure. The separation model and training setting is the same as in the human speaker experiments as described in Section 4.3 and 5.2 of the main paper. We initialized the separation weights from the ones trained on LRS2.

We provide qualitative results in the video. In these, we demonstrate how our model uses the learned audio-visual objects to: i) successfully separate the voices of characters in multi-speaker clips; ii) handle challenging synthetic mixtures of the same character (e.g. Marge-Marge, Homer-Homer); iii) remove background noise and music.

4 Architecture details.

In Table 1, we provide the full architecture for the audio-visual synchronization module used for obtaining the attention maps. In Figure 2 and Table 2 we give full architecture details for the source separation module.

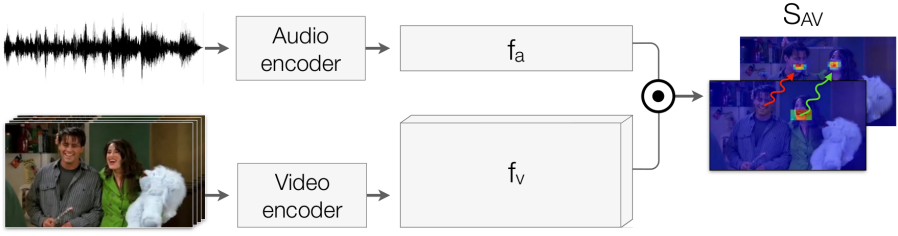

Fig. 1: Synchronization network architecture. This is part of Figure 2 of the main paper.

Table 1: Architecture details for the audio-visual synchronization network, shown on Figure 1. We use a two-stream architecture similar to [4], containing a video and audio encoder that consume their respective modality and output embeddings in the same subspace. The embeddings are used to construct the audio-visual attention map S_{av}. K denotes kernel width and S the strides (3 numbers for 3D convolutions and 2 for 2D convolutions). *mp* denotes a max-pooling layer. Batch Normalization and ReLU activation are added after every convolutional layer. **Note:** To reduce clutter, T was used in the paper instead of $T-4$ for the temporal dimension of the extracted embeddings.

(a) Audio Encoder

Layer	# filters	K	S	Output
input	1	-	-	$4T \times 80$
conv1	64	(3,3)	(1,2)	$4T \times 40$
mp1	-	(3,1)	(1,2)	$4T \times 19$
conv2	192	(3,3)	(1,1)	$4T \times 19$
mp2	-	(3,3)	(2,2)	$2T \times 9$
conv3	256	(3,3)	(1,1)	$2T \times 9$
conv4	256	(3,3)	(1,1)	$2T \times 9$
conv5	256	(3,3)	(1,1)	$2T \times 9$
mp5	-	(3,3)	(2,2)	$T \times 4$
conv6	512	(4,4)	(1,1)	$T-4 \times 1$
fc7	512	(1,1)	(1,1)	$T-4 \times 1$
fc8	1024	(1,1)	(1,1)	$T-4 \times 1$

(b) Video Encoder

Layer	# filters	K	S	Output
input	3	-	-	$T \times H \times W$
conv1	64	(5,7,7)	(1,2,2)	$T-4 \times H/2 \times W/2$
conv2	128	(5,5)	(2,2)	$T-4 \times H/4 \times W/4$
mp2	-	(3,3)	(2,2)	$T-4 \times H/8 \times W/8$
conv3	256	(3,3)	(1,1)	$T-4 \times H/8 \times W/8$
conv4	256	(3,3)	(1,1)	$T-4 \times H/8 \times W/8$
conv5	256	(3,3)	(1,1)	$T-4 \times H/8 \times W/8$
conv6	512	(5,5)	(1,1)	$T-4 \times H/8 \times W/8$
mp6	-	(3,3)	(2,2)	$T-4 \times H/16 \times W/16$
fc7	512	(1,1)	(1,1)	$T-4 \times H/16 \times W/16$
fc8	1024	(1,1)	(1,1)	$T-4 \times H/16 \times W/16$

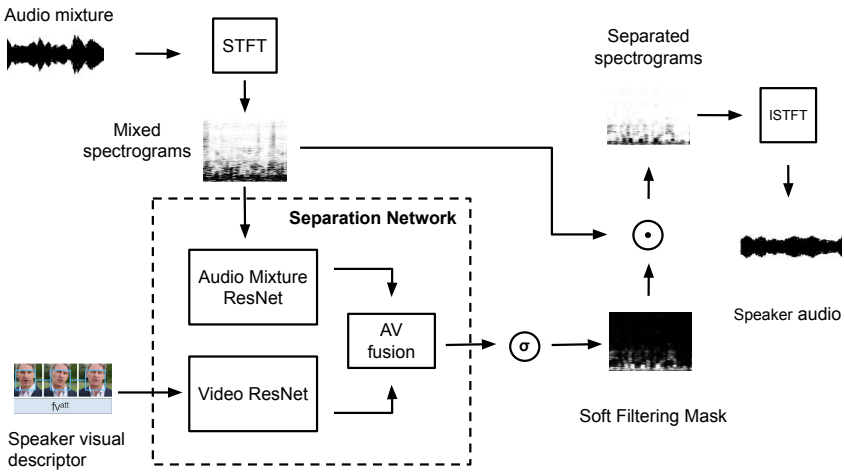

Fig. 2: Separation network architecture. This is a detail of Figure 4 of the main paper.

Table 2: Architecture details for the Separation Network, shown on Figure 2. The modules are described in detail in [3] and include: a) A 1D ResNet that processes the local descriptors extracted for each speaker-object. In particular the descriptors are pooled from the conv6 layer of the Video Encoder shown on Table 1. b) A 1D ResNet that processes the spectrogram of the audio mixture. c) A BLSTM and two fully-connected layers that perform the modality fusion. Notation: K: Kernel width; S: Stride – fractional strides denote transposed convolutions; All convolutional layers are depth-wise separable. Batch Normalization, ReLU activation and a shortcut connection are added after every convolutional layer. **Note:** We also use the phase refining network described in [1] for enhancing the phase of the audio signal, which we omit here for simplicity. For details please refer to the original paper.

(a) Video ResNet

Layer	# filters	K	S	Output
input	512	-	-	$T \times 1$
fc0	1536	(1,1)	(1,1)	$T \times 1$
conv1-2	1536	(5,1)	(2,1)	$T \times 1$
conv3	1536	(5,1)	(½,1)	$2T \times 1$
conv4-6	1536	(5,1)	(1,1)	$2T \times 1$
conv7	1536	(5,1)	(½,1)	$4T \times 1$
conv8-9	1536	(5,1)	(1,1)	$4T \times 1$
fc10	256	(1,1)	(1,1)	$4T \times 1$

(b) Audio Mixture ResNet

Layer	# filters	K	S	Out
input	80	-	-	$T \times 1$
fc0	1536	(1,1)	(1,1)	$4T \times 1$
conv1-5	1536	(5,1)	(1,1)	$4T \times 1$
fc6	256	(1,1)	(1,1)	$4T \times 1$

(c) AV Fusion Network

Layer	# filters	Out
input	512	$4T \times 1$
BLSTM	400	$4T \times 1$
fc1	600	$4T \times 1$
fc2	600	$4T \times 1$
fc_mask	F	$4T \times F$

Bibliography

[1] Afouras, T., Chung, J.S., Zisserman, A.: The conversation: Deep audio-visual speech enhancement. In: INTERSPEECH (2018) 4

[2] Afouras, T., Chung, J.S., Zisserman, A.: LRS3-TED: a large-scale dataset for visual speech recognition. In: arXiv preprint arXiv:1809.00496 (2018) 2

[3] Afouras, T., Chung, J.S., Zisserman, A.: My lips are concealed: Audio-visual speech enhancement through obstructions. In: INTERSPEECH (2019) 4

[4] Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Workshop on Multi-view Lip-reading, ACCV (2016) 2, 3

[5] Deng, J., Guo, J., Yuxiang, Z., Yu, J., Kotsia, I., Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild. In: arxiv (2019) 2

[6] Dutta, A., Zisserman, A.: The VIA annotation software for images, audio and video. In: Proceedings of the 27th ACM International Conference on Multimedia. MM '19, ACM, New York, NY, USA (2019) 1

[7] Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. ACM Transactions on Graphics (TOG) **37**(4), 112 (2018) 2

[8] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Proc. ECCV. pp. 21–37. Springer (2016) 2

[9] Nagrani, A., Chung, J.S., Zisserman, A.: VoxCeleb: a large-scale speaker identification dataset. In: INTERSPEECH (2017) 2