

SyncTalklip: Highly Synchronized Lip-Readable Speaker Generation with Multi-Task Learning

Anonymous Authors

1 BASELINE

ATVNet employs an intermediate representation-based approach, initially predicting facial landmarks before generating realistic images. Wav2Lip, a representative of reconstruction methods, claims state-of-the-art (SOTA) performance in lip-sync synchronization. Talklip is the first method to prioritize intelligibility, but at the expense of synchronization. It uses AV-HuBERT as a supervisor. Faceformer represents an emerging approach, namely the sequence-to-sequence manner, which employs a transformer to process the entire audio and output a video sequence. Originally, FaceFormer was designed to generate a vertex sequence (3D scan data) from an audio sequence. It does not work if we simply replace the vertex sequence with the image sequence. Thus, we adopt the video encoder and corresponding visual input to replace the style encoder in FaceFormer and add a skip connection between the visual input and the output image. PC-AVS is also a reconstruction method that disentangles identity, speech content, and poses. The latter methods do not publicly share their code or training materials. Thus, their performances are documented based on available data.

2 THE MAIN IDEA OF SYNC TALKLIP (FREEZE)

During the training process of AV-SyncNet, the audio-video embeddings are aligned. However, during the fine-tuning process of SyncTalklip, the alignment of these embeddings is adjusted to enhance lip readability. As mentioned in Sec. 3.4, the AV-SyncNet audio encoder is frozen during the training of the SyncTalklip. Instead, an additional audio encoder is employed to participate in the update process. The output from the frozen AV-SyncNet encoder serves as the input for computing cross-modality contrast loss, but it does not participate in gradient updates. In this way, the alignment between audio and video embeddings is maintained, preserving the initial synchronization established by AV-SyncNet.

3 MORE EXAMPLES

There are some examples in Fig. 3. We strongly encourage the reader to check out the demo video on our website: <https://sync-talklip.github.io>. It presents the results of ablation studies, extended model outcomes, and outcomes from other models.

4 LIP-READING EXPERTS

AV-HuBERT trained on LRS2 (224h) is used to test WER_1 , AV-HuBERT trained on LRS3 (433h) is used to test WER_2 . Conformer is used to test WER_3 . Their abilities are shown in Tab. 1. AV-SyncNet serves as the supervisor, and lip-reading experts are employed to assess the capabilities of our model.

5 MODULE ARCHITECTURE

The details of the generator to synthesize a face image are provided in Tab. 2. The details of the discriminator to penalize unrealistic synthesized face images are provided in Tab. 3. The discriminator

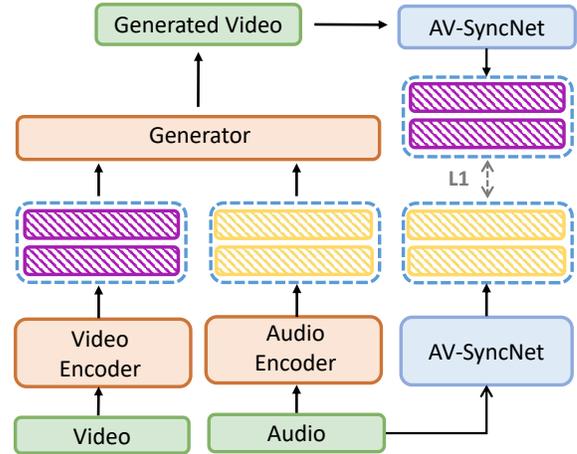


Figure 1: The main framework of SyncTalklip (freeze). The figure does not depict all the losses but conveys the main idea. AV-SyncNet remains frozen from start to finish, and the drawing-in process is consistent with that of AV-SyncNet.



Figure 2: The examples of the SyncTalklip. The related audio is from the LRS2 dataset.

only takes the lower half of faces as inputs. We use a video encoder to extract the identity and pose information to a united visual

Table 1: The ability of all lip-reading experts mentioned in our work. AV-SyncNet(conv) denotes the convolution operation applied before calculating the audio and video loss.

Method	Utt(hrs)		WER(%)
	Train	Test	
AV-HuBERT	LRS2(224h)	LRS2(224h)	25.0
	LRS3(433h)	LRS3(433h)	28.6
AV-SyncNet	LRS2(224h)	LRS2(224h)	26.0
AV-SyncNet (conv)	LRS2(224h)	LRS2(224h)	26.1
Conformer	LRS2+LRS3	LRS2(224h)	40.9

embedding from a concatenation ($6 \times 96 \times 96$) of an identity and a pose image.

Table 2: Generator architecture. All parameters listed in the column of Filters for Conv2D are kernel sizes, output channels, strides, padding, and repetition of layers. Conv 2D T. means 2D transposed convolutional layers which has an extra parameter called output padding, placed after the padding parameter.

Layer Type	Filters	Output dim.
Conv 2D	{[1, 1], 512, [1, 1], 0} \times 1	$512 \times 1 \times 1$
Conv 2D T.	{[3, 3], 512, [2, 2], 0} \times 1	$512 \times 3 \times 3$
Conv 2D	{[3, 3], 512, [1, 1], 1} \times 1	$512 \times 3 \times 3$
Conv 2D T.	{[3, 3], 512, [2, 1], 1} \times 1	$512 \times 6 \times 6$
Conv 2D	{[3, 3], 512, [1, 1], 1} \times 2	$512 \times 6 \times 6$
Conv 2D T.	{[3, 3], 384, [2, 1], 1} \times 1	$384 \times 12 \times 12$
Conv 2D	{[3, 3], 384, [1, 1], 1} \times 2	$384 \times 12 \times 12$
Conv 2D T.	{[3, 3], 256, [2, 1], 1} \times 1	$256 \times 24 \times 24$
Conv 2D	{[3, 3], 256, [1, 1], 1} \times 2	$256 \times 24 \times 24$
Conv 2D T.	{[3, 3], 128, [2, 1], 1} \times 1	$128 \times 48 \times 48$
Conv 2D	{[3, 3], 128, [1, 1], 1} \times 2	$128 \times 48 \times 48$
Conv 2D T.	{[3, 3], 64, [2, 1], 1} \times 1	$64 \times 96 \times 96$
Conv 2D	{[1, 1], 64, [1, 1], 1} \times 2	$64 \times 96 \times 96$
Conv 2D	{[3, 3], 32, [1, 1], 1} \times 1	$32 \times 96 \times 96$
Conv 2D	{[1, 1], 3, [1, 1], 0} \times 1	$3 \times 96 \times 96$

6 SYNCHRONIZATION METRIC

6.1 The definition of the PRI

To evaluate the alignment ability of AV-SyncNet from another angle, we propose the primacy index (PRI), which is a novel metric. This metric can use embeddings as input, which is different from the LSE-C and LSE-D mentioned in the Sec. 4.3. Given N audio-video embeddings, the dimension of each is D . Denote f_i^a and f_i^v as the

Table 3: Discriminator architecture. All parameters listed in the column of Filters are kernel sizes, output channels, strides, padding, and repetition of layers.

Layer Type	Filters	Output dim.
Conv 2D	[7, 7], 32, [1, 1], 1 \times 1	$32 \times 48 \times 96$
Conv 2D	[5, 5], 64, [1, 2], 1 \times 1	$64 \times 48 \times 48$
Conv 2D	[5, 5], 64, [1, 1], 1 \times 1	$64 \times 48 \times 48$
Conv 2D	[5, 5], 128, [2, 2], 1 \times 1	$128 \times 24 \times 24$
Conv 2D	[5, 5], 128, [1, 1], 1 \times 1	$128 \times 24 \times 24$
Conv 2D	[5, 5], 256, [2, 2], 1 \times 1	$256 \times 12 \times 12$
Conv 2D	[5, 5], 256, [1, 1], 1 \times 1	$256 \times 12 \times 12$
Conv 2D	[5, 5], 512, [2, 2], 1 \times 1	$512 \times 6 \times 6$
Conv 2D	[5, 5], 512, [1, 1], 1 \times 1	$512 \times 6 \times 6$
Conv 2D	[3, 3], 512, [2, 2], 1 \times 1	$512 \times 3 \times 3$
Conv 2D	[3, 3], 512, [1, 1], 1 \times 1	$512 \times 3 \times 3$
Conv 2D	[3, 3], 512, [1, 1], 1 \times 1	$512 \times 1 \times 1$
Conv 2D	[1, 1], 512, [1, 1], 1 \times 1	$512 \times 1 \times 1$

i -th embeddings of audio and video, respectively. First, calculate the distance matrix $M \in \mathbb{R}^{N \times N}$, where M_{ij} represents the distance between a_i and v_j . Denote t_i as the rank of the M_{ii} in the i -th row, so the t_i/N is the importance of the M_{ii} . Then the PRI is defined as follows:

$$PRI = \frac{\sum_{i=1}^N t_i}{N \times N} \in (0, 1)$$

The distance of the a_i and v_j is calculated as follows:

$$M_{ij} = \|a_i - v_j\|_1$$

6.2 The relationship between LSE-D and PRI

LSE-D employs a sliding window strategy, calculating the distance between a given frame and other frames within the window. Denote the window size as S , where the interval $[i, i+S]$ is considered as within the sliding window. Calculate the distance between a_i and v_j , where $j \in [i, i+S]$. This yields the distance metrics: $M \in \mathbb{R}^{N \times S}$. Average across the S dimension and calculate the minimum of the resulting values, which forms the basis of the LSE-D algorithm.

The distance of a_i and v_j , is calculated as follows:

$$M_{ij} = \frac{f_i^a \cdot f_j^v}{\|f_i^a\|_2 \cdot \|f_j^v\|_2} \in (-1, 1)$$

Compared to LSE-D, PRI has a fixed window size and utilizes different distance metrics. What's more, LSE-D is generated by SyncNet, which takes videos as input, whereas PRI utilizes embeddings as input.

6.3 Performance

As shown in the Tab. 4, the alignment of AV-SyncNet is on par with SyncNet at the embedding level. The outcome for AV-HuBERT

is less than 0.5, which demonstrates that its alignment has been corrupted during the fine-tuning process.

Table 4: Your table caption here.

Method	AV-HuBERT	SyncNet	AV-SyncNet
PRI	0.51	0.45	0.47

7 PRIMARY HYPERPARAMETER

The learning rate is set to 1×10^{-4} , and the batch size is configured at 8. In AV-SyncNet, the parameters λ_{av} and λ_{lip} are empirically set to 1 each. In SyncTalklip, the values are set as follows: $\lambda_{cav} = 1 \times 10^{-3}$, $\lambda_{gan} = 7 \times 10^{-2}$, $\lambda_{rec} = 1 \times 10^{-5}$, and $\lambda_{lip} = (1 - \lambda_{cav} - \lambda_{gan} - \lambda_{rec})$ respectively.