Audio-driven Talking Face Generation with Stabilized Synchronization Loss

Dogucan Yaman[1], Fevziye Irem Eyiokur[1], Leonard Bärmann[1], Hazım Kemal Ekenel[2], and Alexander Waibel[1,3]

[1] Karlsruhe Institute of Technology, Karlsruhe, Germany
[2] Istanbul Technical University, Istanbul, Turkey
[3] Carnegie Mellon University, Pittsburg PA, USA
dogucan.yaman@kit.edu

A AVSyncNet

A.1 SyncNet

In audio-driven talking face generation, an explicit loss for evaluating lip synchronization is essential for achieving proper audio-lip alignment (called lip synchronization). Typically, a pretrained model for audio-visual feature extraction is employed to calculate the loss function. A commonly used method is incorporating SyncNet, as proposed by the Wav2Lip [8]. SyncNet consists of both an audio encoder and an image encoder. Each has consecutive 2D convolutional layers followed by Batch Normalization and ReLU activation function. During training, these two encoders are trained jointly. First, five consecutive face and audio sequences are provided as input. Notably, the image encoder focuses on the bottom half of the face, as only the mouth region contains relevant information for lip synchronization. After extracting audio and image features with the respective encoders, cosine similarity is computed, followed by binary cross-entropy loss, which we call as *lip-sync loss*. While aligned audio-lip pairs are treated as positive samples, non-aligned pairs serve as negative samples throughout the training process. Eq. (A) shows the cosine similarity calculation, while Eq. (B) indicates the binary cross-entropy loss.

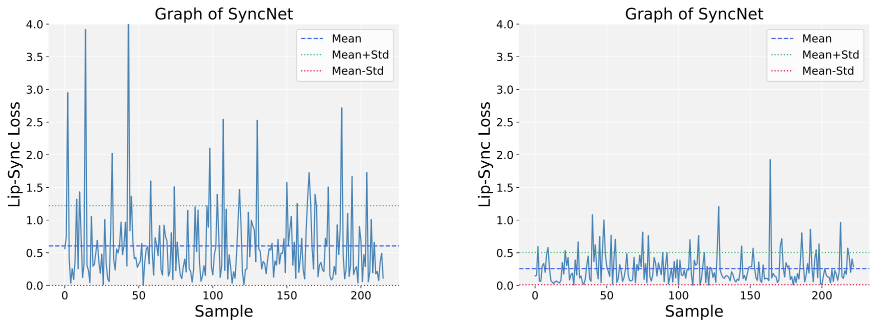
$$P_{sync} = \frac{F_I \cdot F_A}{max(||F_I||_2 \cdot ||F_A||_2, \epsilon)} \tag{A}$$

$$L_{sync} = \frac{1}{N}\sum_{1}^{N} -log(P^i_{sync}) \tag{B}$$

where $F_I \in \mathbb{R}^{1\times1\times512}$ and $F_A \in \mathbb{R}^{1\times1\times512}$ state image features and audio features extracted by SyncNet respective encoders.

A.2 AVSyncNet Architecture

Our audio encoder is based on ResNetSE34. It contains consecutive convolutional layers followed by SE block and self-attentive pooling (SAP) block. For the image



(a) SyncNet [8] **(b)** AVSyncNet (ours)

Fig. A: Lip-sync loss between GT audio-lip pairs on random LRS2 test samples, demonstrating the instability of SyncNet and more accurate AVSyncNet. This graph is similar with Fig. 1 in the main paper. After the cosine similarity, we calculate lip-sync loss as shown in Eq. (B).

encoder, We employ ResNet-50. Since we provide the bottom-half face (namely mouth region), we modify the input layer to be able to process 112×224 input.

A.3 AVSyncNet Training

In order to learn lip synchronization between the mouth region and audio snippet, it is crucial to train the image and audio encoders jointly. For the training, we follow the similar strategy as in SyncNet [8]. In the training, we first pass the bottom half of the face images (five images) and the corresponding audio sequence through their respective encoders. After each encoder, we have an additional convolutional layer to further process the features. Then, we obtain feature representations for face and audio modalities: $F_I \in \mathbb{R}^{1\times1\times512}$ and $F_A \in \mathbb{R}^{1\times1\times512}$. We calculate the cosine similarity between them. In the end, by using this cosine similarity, we compute binary cross-entropy loss, as in Eq. (A) and Eq. (B). We choose the audio-lip pairs as positive samples and non-aligned audio-lip pairs (random audio sequence selection from the video that does not overlap with the given face sequence) as negative samples throughout the training. We conduct training on the LRS2 dataset and use the same train-val-test setups as we use in the training of the talking face generation model.

A.4 Performance Comparison

In Fig. A, we share the lip-sync loss version of Fig. 1a-b in the main paper. After we calculate cosine similarity, we obtain the lip-sync loss, which is a binary cross-entropy loss as shown in Eq. (B). The lip-sync loss comparison clearly demonstrates that our AVSyncNet is much more accurate and stable than SyncNet [8].

Table A: Comparison of SyncNet [8] and our proposed AVSyncNet on LRS2 traing and test sets [1].

Model	Data	Cosine Sim ↑	Lip-sync Loss ↓
SyncNet	Train set	0.614	0.763
	Test set	0.648	0.532
AVSyncNet	Train set	0.790	0.216
	Test set	0.778	0.280

(a) Horizontal Shifting

(b) Random Shear

(c) Rotation

(d) Rotation - mouth region

Fig. B: Applied data augmentation methods to test SnyncNet and AVSyncNet against the transformation. **(a)** shows the horizontal shifting. **(b)** indicates the random shear. **(c)** demonstrates the rotation, while **(d)** is the bottom half of the images in **(c)**.

In Tab. A, we show the average cosine similarity and lip-sync loss on the LRS2 [1] train and test sets for GT audio-lip pairs. Both the cosine similarity and lip-sync loss demonstrate the superiority of our AVSyncNet over the SyncNet [8].

In Fig. 6 in the main paper, we analyse the performance of SyncNet and our AVSyncNet over the transformed data. We first apply horizontal shift as in [7] to explore the shift-invariance capacity. Then, we apply the same strategy for other transformations: random shear and rotation. We present sample images to illustrate these transformations in Fig. B.

B Stabilized Synchronization Loss

We analyze the behaviour of SyncNet [8] and utilize lip-sync loss [8] in order to show the unstable nature and poor representation capacity of them. In Fig. C, we show the cosine similarity between audio features and ground truth image features for LRS2 dataset. While Fig. Ca and Fig. Cb represent the cosine similarity between audio features and ground truth image features, Fig. Cc and Fig. Cd indicate the cosine similarity between audio features and generated image features. In Fig. Cc, we generated samples with a model that was trained with lip-sync loss [8] (Setup `A` from main paper Table 2). On the other hand, in Fig. Cd, the images are synthesized by a model that was trained with our proposed full model (Setup `H`). Fig. Cg and Fig. Ch show the distance between (a)-(c) and (b)-(d).

In order to compute these cosine similarities, we employ SyncNet [8] and extract features. Then, we calculate cosine similarity between them. This is the fundamental step of lip-sync loss [8]. The graph explicitly shows the unstable cosine similarity between audio and ground truth images (a). Therefore, calculating synchronization loss may cause unstable training as well as harm the synchronization since it sometimes cannot represent the audio-visual information. On the other hand, our proposed method(s) allows the network to have more similar similarity scores (d) with the ground truth data (b). Thus, the model has better audio-lip synchronization and the training was much more stable.

C Silent-Face Generation (G_S)

In Fig. D, we present input images and generated silent version with our silent-lip generator G_S. The results clearly indicate the quality of our G_S in terms of making lips silent (flat) as well as preserving identity along with other visual details such as texture, pose, eyes.

D Evaluation Metrics

SSIM Structural similarity index measure (SSIM) [11] is a metric to measure the perceived quality of the generated images. It requires ground truth images to calculate the score. Higher SSIM score indicates better quality.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{C}$$

PSNR Peak signal-to-noise ratio (PSNR) is a metric to measure the image quality. It benefits from the ratio between the maximum possible square of a pixel value and mean squared error (MSE) between the generated image and the ground truth one. Higher PSNR score states better quality.

$$PSNR(I', I) = 10 * log_{10} \frac{max(I')^2}{MSE(I', I)} \tag{D}$$

$$MSE(I', I) = \frac{1}{HW} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} |I'_{i,j} - I_{i,j}|^2 \tag{E}$$

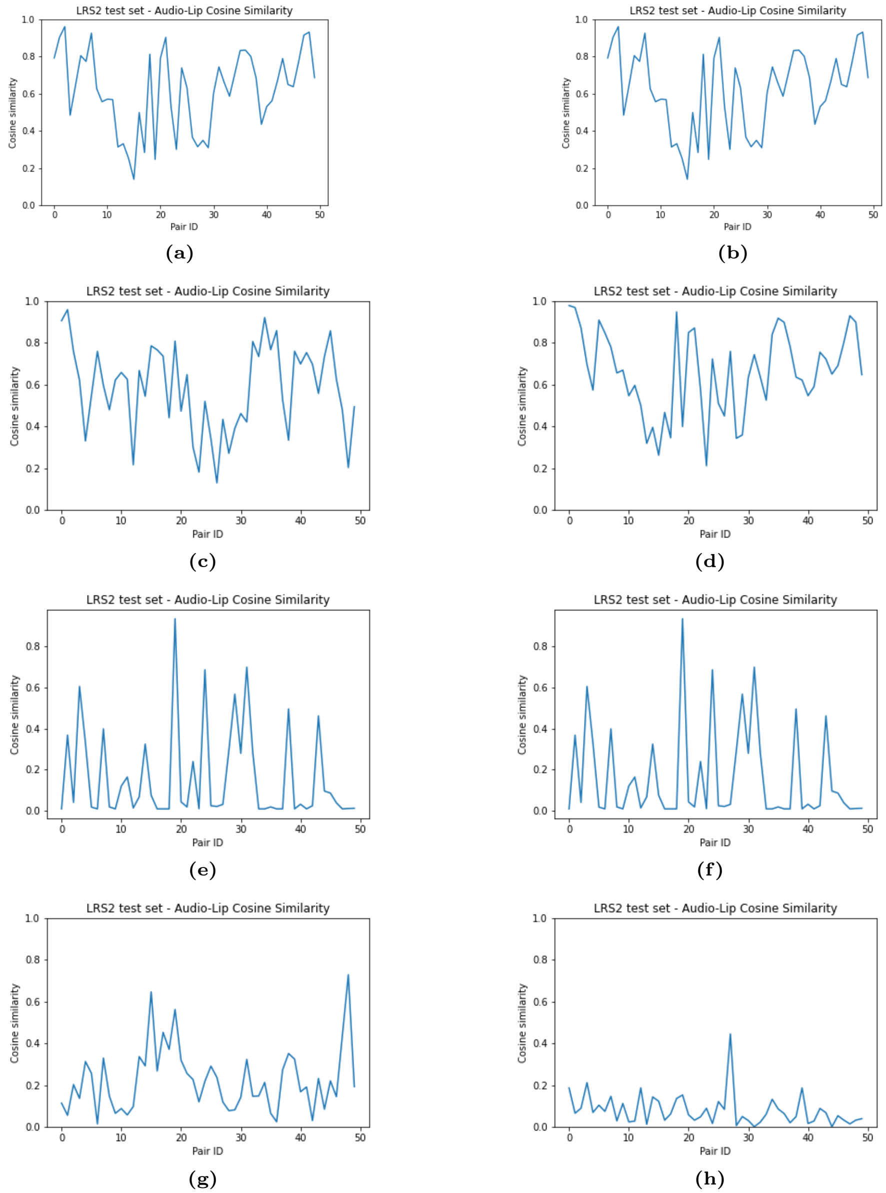


Fig. C: Cosine similarity analyses on LRS2 test set. (a) and (b) are identical and provided for visual reference to the vertically aligned figures below. They show cosine similarity between audio and ground truth image pairs. (c) shows the cosine similarity between audio and generated images. These images are synthesized by using lip-sync loss [8]. (d) presents cosine similarity between audio and generated images by using our proposed stabilized synchronization loss. (e) and (f) are the same, they show the cosine similarity between audio and randomly selected reference images. (g) and (h) are the distance between (a)-(c) and (b)-(d) respectively.

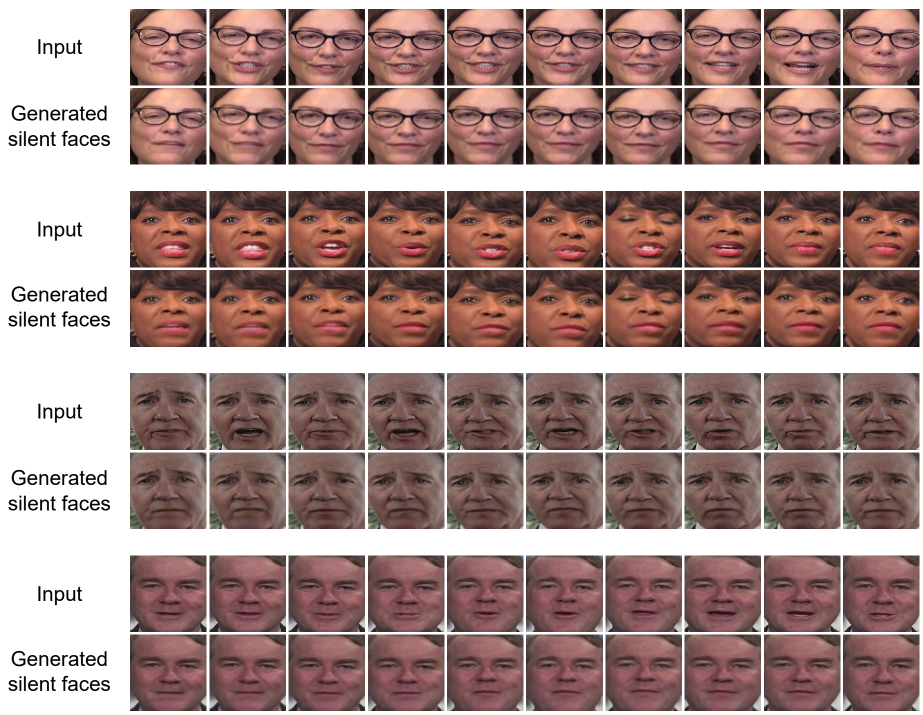


Fig. D: Generated silent face samples with our silent-lip generator G_S.

Although it is one of the common image quality metrics, it is not as informative and robust as FID and SSIM.

FID Fréchet Inception Distance (FID) metric [6] is a quality metric and indicates the feature level distance between generated images and ground truth images. Lower FID score means higher similarity between these real and generated samples and least possible score is 0. In order to calculate this metric, first, features from real and generated samples are extracted by pretrained inception v3 model [9] that was trained on ImageNet dataset [5] for image classification. The features are extracted from the last pooling layer of this network. Then, the following formula is employed to find FID score.

$$FID(F', F) = |\mu_{F'} - \mu_F| + TR(\Sigma_{F'} + \Sigma_F - 2(\Sigma_{F'}\Sigma_F)^{\frac{1}{2}}) \tag{F}$$

IFC Inter-frame consistency, also known as motion constraint loss, is proposed in [15] as a loss function. In that paper, the authors tried to synthesize a video that contains the response of a subject for the given audio. Since the authors generated the video, temporal consistency is crucial to have a natural video. Therefore, this is used as a loss function to evaluate the inter-frame consistency. The idea is rather simple. First, the difference between consecutive frames of the generated video is calculated in the pixel space. Then, it is done for the ground-truth video in the same manner. Afterward, the difference between these two distance sets is calculated. We propose to use this loss function as an evaluation metric in the same manner in order to measure the inter-frame consistency of the generated video and ground-truth video. The utilized formula is as below:

$$IFC = \sum_{t=2}^{T} ||\mu(\beta_t) - \mu(\hat{\beta}_t)||_2 \tag{G}$$

where $\mu(\beta_t^l) = \beta_t^l - \beta_{t-1}^l$ measures the changes between current frame t and previous frame $t-1$ in pixel space. While t is an index to represent frame time step, β and $\hat{\beta}$ indicate the ground truth video and generated video, respectively.

LMD We extract landmarks from generated and ground truth images [2][4] and calculate distance between them [3]. Please note that we only consider the landmarks in the mouth region, not the entire face. Although it compares the distance between landmark points of the generated image and the ground-truth one, it is hard to say that it evaluates synchronization. Because the different landmark point does not only depend on the lip movement. Even if the generated face has the correct lip movement as the ground-truth image has, the landmark points on the generated image might be quite different than the ground-truth one. For instance, the pose differences due to generation error or shifting on the image

[4] https://github.com/1adrianb/face-alignment

Table B: Quantitative results on HDTF dataset [13].

	HDTF						
Method	SSIM ↑	PSNR ↑	FID ↓	IFC ↓	LMD ↓	LSE-C ↑	LSE-D ↓
Wav2Lip [8]	0.841	24.812	35.41	0.248	1.341	9.054	6.141
VideoReTalking w/ FR [4]	0.830	24.551	29.77	0.287	3.085	6.121	7.368
TalkLip [10]	0.820	25.229	25.10	0.305	2.981	6.189	7.276
IPLAP [14]	0.869	27.801	22.09	0.263	2.217	5.563	8.495
Ours w/o FR	0.893	28.602	21.46	0.201	1.296	8.304	6.366
Ours w/ FR	0.885	26.454	24.25	0.346	1.688	8.155	6.347

can affect the LMD. Therefore, LMD considers the pose difference, shifting, generation error, and lip synchronization altogether although it is mostly affected by lip movements.

LSE-C and LSE-D [8] LSE-C is the average confidence score and benefits from audio and lip representations from SyncNet. Higher confidence scores insicate better audio-lip synchronization. On the other hand LSE-D metric measures the distance between audio and lip representations by using SyncNet. Lower LSE-D means higher audio-lip synchronization.

E Additional Quantitative Results

E.1 HDTF

We conduct experiments on HDTF dataset [13] with our model and also baseline models. The results are presented in Tab. B. Please note that DINet was trained on HDTF dataset. In their paper, the authors mentioned that they selected 20 videos from HDTF dataset randomly for test set. As we do not know which videos were selected for test, we did not present the results for DINet as it would be a test on a training data, harming fair comparison. On the other hand, all other models were not trained (or finetuned) on HDTF dataset.

F Additional Qualitative Results

F.1 HDTF

We present more sample images generated by our model and other models along with GT images in Fig. E.

G User Study

We conducted a user study with videos from HDTF dataset [13] and present the results in Tab. C. We chose 10 different videos for each model. In total, 15 participants joined our user study.

Fig. E: Comparison of different models. The videos are selected from HDTF dataset [13].

References

1. Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Deep audio-visual speech recognition. IEEE transactions on pattern analysis and machine intelligence **44**(12), 8717–8727 (2018) 3
2. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: International Conference on Computer Vision (2017) 7
3. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7832–7841 (2019) 7
4. Cheng, K., Cun, X., Zhang, Y., Xia, M., Yin, F., Zhu, M., Wang, X., Wang, J., Wang, N.: Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022) 8, 10
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 7

Table C: User study on randomly selected HDTF videos [13].

Method	Sync ↑	Visual ↑	Overall ↑
Wav2Lip [8]	2.71	2.85	1.81
PC-AVS	1.15	2.14	1.42
VideoReTalking w/ FR [4]	3.72	3.26	3.37
DINet [12]	2.28	2.42	2.34
TalkLip [10]	3.16	1.90	2.15
IPLAP [14]	3.02	3.75	3.66
Ours	3.85	3.64	3.71
Ours w/ FR	**3.88**	**3.92**	**4.11**

6. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 7
7. Muaz, U., Jang, W., Tripathi, R., Mani, S., Ouyang, W., Gadde, R.T., Gecer, B., Elizondo, S., Madad, R., Nair, N.: Sidgan: High-resolution dubbed video generation via shift-invariant learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7833–7842 (2023) 3
8. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 484–492 (2020) 1, 2, 3, 4, 5, 8, 10
9. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016) 7
10. Wang, J., Qian, X., Zhang, M., Tan, R.T., Li, H.: Seeing what you said: Talking face generation guided by a lip reading expert. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14653–14662 (2023) 8, 10
11. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004) 4
12. Zhang, Z., Hu, Z., Deng, W., Fan, C., Lv, T., Ding, Y.: Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. arXiv preprint arXiv:2303.03988 (2023) 10
13. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3661–3670 (2021) 8, 9, 10
14. Zhong, W., Fang, C., Cai, Y., Wei, P., Zhao, G., Lin, L., Li, G.: Identity-preserving talking face generation with landmark and appearance priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2023) 8, 10
15. Zhou, M., Bai, Y., Zhang, W., Yao, T., Zhao, T., Mei, T.: Responsive listening head generation: A benchmark dataset and baseline (2022) 7