SyncTalkFace: Talking Face Generation with Precise Lip-Syncing via Audio-Lip Memory

Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, Yong Man Ro*

Image and Video Systems Lab, KAIST, South Korea
{jinny960812, ms.k, joanna2587, jeongsoo.choi, ymro}@kaist.ac.kr

Abstract

The challenge of talking face generation from speech lies in aligning two different modal information, audio and video, such that the mouth region corresponds to input audio. Previous methods either exploit audio-visual representation learning or leverage intermediate structural information such as landmarks and 3D models. However, they struggle to synthesize fine details of the lips varying at the phoneme level as they do not sufficiently provide visual information of the lips at the video synthesis step. To overcome this limitation, our work proposes Audio-Lip Memory that brings in visual information of the mouth region corresponding to input audio and enforces fine-grained audio-visual coherence. It stores lip motion features from sequential ground truth images in the value memory and aligns them with corresponding audio features so that they can be retrieved using audio input at inference time. Therefore, using the retrieved lip motion features as visual hints, it can easily correlate audio with visual dynamics in the synthesis step. By analyzing the memory, we demonstrate that unique lip features are stored in each memory slot at the phoneme level, capturing subtle lip motion based on memory addressing. In addition, we introduce visual-visual synchronization loss which can enhance lip-syncing performance when used along with audio-visual synchronization loss in our model. Extensive experiments are performed to verify that our method generates high-quality video with mouth shapes that best align with the input audio, outperforming previous state-of-the-art methods.

Introduction

Talking face generation from speech, also referred to as lip-syncing, is synthesizing a video of a target identity such that the mouth region is consistent with arbitrary audio input. It has many applications such as audio-driven photo-realistic avatars that can be employed in online classes or games, dubbing films in another language, and communication aids for the hearing-impaired who can lip-read. As the talking face generation carries various practical usage, it has received great interest for research.

The main challenge of talking face generation from speech is aligning audio and visual information so that the generated facial sequence is coherent with the input audio. Previous methods based on encoder-decoder structure have worked on improving audio and visual representations. (Zhou et al. 2019, 2021) disentangled visual input into identity and speech content space using metric learning and enhanced audio feature by embedding into a shared latent space between visual feature. (Mittal and Wang 2020) disentangled audio representation into phonetic content, emotional tone, and other factors. They have explored feature disentanglement to remove irrelevant factors in lip-syncing. However, the disentangled audio representation does not explicitly contain visual information of the mouth which can help the decoder to map visual dynamics from audio.

Recent advances utilize intermediate structural representations such as facial landmarks and 3D models to better capture facial dynamics. (Chen et al. 2019; Das et al. 2020; Zhou et al. 2020) mapped lip landmarks from audio and composited into the mouth region of a target person. (Song et al. 2020; Thies et al. 2020) learned speaker independent features in a 3D face model and rendered a talking face video of a target person with fine-tuning. However, they commonly lack sophistication in lip-syncing. This is because the facial landmarks are too sparse to provide accurate lip synchronization, and the 3D models cannot capture fine details in the mouth region including teeth (Zhang et al. 2021; Wang, Mallya, and Liu 2021). Also, they bear the limitation of having to acquire the intermediate representation separately from the generation network.

Distinct from the previous works, we introduce Audio-Lip Memory that explicitly provides visual information of the mouth region and enables more precise lip synchronization with input audio. The memory learns to align audio with corresponding lip features from sequential ground truth images during training, so that it outputs the audio-aligned lip features, when queried with audio at inference time. The recalled lip features are fused with audio features and injected into the decoder for synthesizing the talking face video. As the decoder can leverage the explicit visual hints of the mouth, it can better map audio to video both temporal- and pixel-wise. Moreover, the Audio-Lip Memory stores the representative lip features at the phoneme level and retrieves various combinations of the lip features through memory addressing, enabling sophisticated and diverse lip movements. We additionally impose visual-visual synchronization along

*Corresponding author.

with audio-visual synchronization for strong lip-syncing. Hence, the proposed model achieves high-quality video generation with fine-grained audio-visual coherence.

Our contributions are as follows: (1) We propose Audio-Lip Memory that maps audio to lip-movement intermediate representations that bridge audio with lip sync video generation. It explicitly provides visual hints of the lip movement to the decoder and enhances the sophistication of lip motion corresponding to the audio. (2) We ensure strong lip synchronization by utilizing visual-visual synchronization between ground truth face sequence and generated face sequence along with audio-visual synchronization between input audio and generated faces. (3) By analyzing learned representations inside the memory, we confirm that the representations are stored at the phoneme level in each memory slot and different combinations of addressing slots yield variational mouth shapes. Thus, direct manipulation of lip movement using memory address is possible. (4) We achieve state-of-the-art performance on LRW and LRS2 dataset in terms of visual quality and lip-sync quality.

Related Work

Talking Face Generation Existing works on talking face generation can be broadly categorized into reconstruction based methods and intermediate representation based methods. Reconstruction based methods (Chen et al. 2018; Song et al. 2018; Jamaludin, Chung, and Zisserman 2019; KR et al. 2019; Vougioukas, Petridis, and Pantic 2020) follow the encoder-decoder structure where identity features and speech features are extracted and fused together as an input to a decoder to synthesize talking face videos in an end-to-end manner. (Prajwal et al. 2020) took a face video as a visual input and used lower-half masked of the input video as a pose prior. It employed a pre-trained lip-sync discriminator and highlighted the importance of an accurate lip-sync discriminator that can feedback lip-sync quality to the network. (Zhou et al. 2019) disentangled speech related features and identity related features from video input through associative-and-adversarial training. In (Zhou et al. 2021), the author further disentangled visual input into identity space, pose space, and speech content space, allowing free pose-control. Although many works have explored improving visual representation by disentangling different factors in visual input, not much work has sought into improving audio representation. (Mittal and Wang 2020) attempted to improve performance from the perspective of audio representation learning. They disentangled phonetic content, emotional tone, and the rest of the other factors from audio using Variational Autoencoder with KL divergence and negative log likelihood with margin ranking loss. Instead of decoupling speech related features from the audio, our work explicitly filters out lip motion related features from the input audio. As we directly map audio to lip features before injecting the audio features to the generator, we can impose lip synchronization earlier in the generation step.

Intermediate representation based methods consist of two cascaded modules where intermediate representations such as landmarks and 3D models are leveraged to generate video from input audio. (Chen et al. 2019; Das et al. 2020) estimated facial landmarks from input audio and then generated video conditioned on the generated landmarks and a reference image. (Zhou et al. 2020; Das et al. 2020) separately considered speech content related landmarks and speaker identity related landmarks for the generation of unseen subjects. 3D model based methods commonly extract expression, geometry, and pose parameters to reconstruct 3D facial mesh (3DMM) from which a face video is generated (Song et al. 2020; Yu et al. 2020). (Thies et al. 2020) used a pretrained audio-expression network to model an expression basis in the 3D face model. (Zhang et al. 2021) proposed a style-specific generator that produces facial animation parameters that are combined with facial shape parameters to create 3D mesh points. Such intermediate representations provide structural information of facial dynamics that has limitations in containing fine details of the mouth. Moreover, acquiring the landmarks and 3D models is laborious and time-consuming. We try to overcome these limitations by leveraging recalled lip features from memory. The memory stores lip features in value memory slots at the phoneme level during training so that information about lip motion corresponding to input audio can be obtained at inference. Also, as various combinations of the lip features in each slot are possible through memory addressing, more diverse and subtle lip movements can be portrayed.

Audio-Visual Alignment Audio-visual alignment aims to find the correlation space between audio and video, and find temporal coherence between the two modality data. In the context of talking face generation task, (Prajwal et al. 2020) directly employed a pretrained embedding module (Chung and Zisserman 2016b) as a lip-sync discriminator, and (Zhu et al. 2020) presented asymmetric Mutual Information Estimator. They all relied on the audio-visual embedding module placed at the end of the generator network to give feedback to the whole network on the coherence between input audio and generated video. More recent works on cross-modal learning apply multi-way matching loss that considers intra-class pairs as well as inter-class pairs, and have shown its effectiveness (Chung, Chung, and Kang 2019; Nagrani et al. 2020; Gao and Grauman 2021). Inspired by the intra-class loss, our work additionally exploits visual-visual sync loss. As input audio and ground truth video are in sync, we can expect the complementary effect of audio-visual alignment by aligning visual lip features from generated face sequence and ground truth face sequence.

Memory Network Memory Network (Weston, Chopra, and Bordes 2014) provides a long-term memory component that can be read from and written to with inference capability. (Miller et al. 2016) introduced key-value memory structure where key memory is used to address memories with respect to a query and corresponding value is obtained from value memory using the address. Since the scheme can remember selected information, it is effective for augmenting features (Kaiser et al. 2017; Lee et al. 2018; Cai et al. 2018; Zhu et al. 2019; Pei et al. 2019; Lee et al. 2021; Kim, Park, and Ro 2021; Kim et al. 2021; Kim and Ro 2021). (Yi et al. 2020) incorporated memory to talking face generation to re-

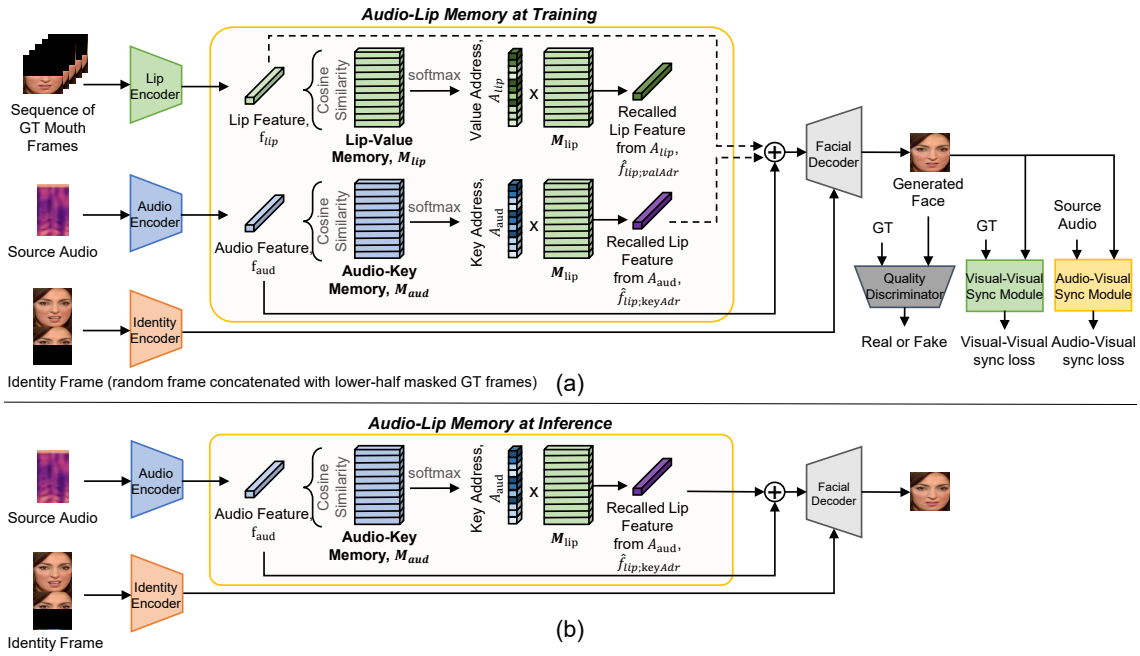

Figure 1: Overview of our proposed model. During training, Audio-Lip Memory learns to store lip feature f_{lip} in the lip-value memory and to align key address A_{aud} with value address A_{lip} as depicted in (a). During inference, the model utilizes recalled lip feature from key address, $\hat{f}_{lip;keyAdr}$, obtained from audio input as a query as shown in (b).

fine roughly rendered frames into realistic frames. It stores spatial features and identity features as key-value pairs and retrieves the best-matching identity feature using the spatial feature as a query. Unlike previous works that use memory only to remember critical information, we employ the key-value memory to align and store two different modality features. We map audio to lip features through key-value memory addressing so that the lip features not available at inference can be utilized with audio input as a query. The recalled lip features from the value memory is used as an intermediate representation to bridge between audio and video.

Methods

We propose Audio-Lip Memory that explicitly maps audio to lip features. Our whole pipeline is depicted in Fig.1. We take a frame of target identity, 0.2 seconds of source audio, and upper half masked face sequence (5 frames) corresponding to the source audio as input. We aim to lip-sync the input video of the target identity such that the mouth region is consistent with the input audio, altering only the mouth while preserving all the other elements (i.e., pose, identity, and etc). We align and store encoded audio features and encoded lip features in the Audio-Lip Memory so that lip features can be obtained when queried with audio features. The recalled lip features from the memory are fused with audio features and injected into the decoder network for video synthesis. In addition, strong lip synchronization is imposed with audio-visual synchronization loss and visual-visual synchronization loss. We explain the details of the Audio-Lip memory in Sec. 3.1 and video synthesis in Sec. 3.2.

Audio-Lip Memory

Audio-Lip Memory maps audio features to lip motion related features. We firstly encode a spectrogram of source audio (0.2 seconds) into audio feature $f_{aud} \in \mathbb{R}^C$ using an audio encoder, and corresponding 5 sequential frames with upper half masked to lip feature $f_{lip} \in \mathbb{R}^C$ using a lip encoder. Audio-Lip Memory is composed of an audio-key memory $\mathbf{M}_{aud} \in \mathbb{R}^{S\times C}$ and a lip-value memory $\mathbf{M}_{lip} \in \mathbb{R}^{S\times C}$, where S denotes slot size and C channel. Note that we universally set C to 512. The memory learns to store representative lip features in the lip-value memory through reconstruction loss between recalled lip features from the key address and lip features extracted from the lip encoder. It simultaneously learns to align lip features with audio features through key-value address alignment loss so that the corresponding lip feature can be retrieved using an audio feature as a query.

Storing lip features in lip-value memory The lip-value memory $\mathbf{M}_{lip} = \{m^i_{lip}\}^S_{i=1}$ where $m^i_{lip} \in \mathbb{R}^C$ is a unique lip feature in the i-th slot. When a lip feature of 5 sequential mouth frames from the lip encoder is given as a query, distance between the lip feature and each of the slots is computed using cosine similarity:

$$d^i_{lip} = \frac{m^i_{lip} \cdot f_{lip}}{\|m^i_{lip}\|_2 \cdot \|f_{lip}\|_2}. \quad (1)$$

Then, we take softmax of the similarity distance computed on individual slots as follows:

$$\alpha^i_{lip} = \frac{\exp(\kappa \cdot d^i_{lip})}{\sum^S_{j=1} \exp(\kappa \cdot d^j_{lip})}, \quad (2)$$

where κ is a scaling term, and α^i_{lip} is an attention weight on the i-th slot of the lip-value memory with respect to the lip feature. By computing the attention weights for all slots, we get a value address $A_{lip} = \{\alpha^1_{lip}, \alpha^2_{lip}, \dots \alpha^S_{lip}\} \in \mathbb{R}^S$. It is used to locate relevant slots in the lip-value memory associated with the lip feature. Finally, we can retrieve the lip feature associated with the query by taking dot product between the value address and the lip-value memory:

$$\hat{f}_{lip;valAdr} = A_{lip} \cdot \mathbf{M}_{lip}. \tag{3}$$

We denote $\hat{f}_{lip;valAdr} \in \mathbb{R}^C$ as recalled lip feature from the value address. By taking the weighted sum of the different lip features stored in individual slots, we can utilize various combinations of the lip features and generate more diverse lip motions. In order to save the lip feature in the lip-value memory, we employ reconstruction loss between the recalled lip feature $\hat{f}_{lip;valAdr}$ and the lip feature given as a query f_{lip} as follows:

$$\mathcal{L}_{store} = \|f_{lip} - \hat{f}_{lip;valAdr}\|_2^2. \tag{4}$$

Through $\mathcal{L}_{store}$, the model learns to embed representative lip features of 5 sequential ground truth frames in the slots attended by value addresses.

Aligning key address with value address After storing the lip features in the lip-value memory, we should be able to retrieve the corresponding lip feature when an audio feature is given as a query. This is how the memory network works at inference time when there are no matching ground truth images to extract lip features from. We obtain key address in the same way as the value address, replacing lip feature f_{lip} with the audio feature f_{aud} and lip-value memory $\mathbf{M}_{lip}$ with audio-key memory $\mathbf{M}_{aud}$ as follows:

$$d^i_{aud} = \frac{m^i_{aud} \cdot f_{aud}}{\|m^i_{aud}\|_2 \cdot \|f_{aud}\|_2}, \tag{5}$$

$$\alpha^i_{aud} = \frac{\exp(\kappa \cdot d^i_{aud})}{\sum_{j=1}^{S} \exp(\kappa \cdot d^j_{aud})}, \tag{6}$$

$$A_{aud} = \{\alpha^1_{aud}, \alpha^2_{aud}, \dots \alpha^S_{aud}\}. \tag{7}$$

We align key address with value address through key-value address alignment loss:

$$\mathcal{L}_{align} = D_{KL}(A_{lip} \| A_{aud}), \tag{8}$$

which is KL divergence between the two address vectors. By aligning key address and value address obtained from audio and video pairs that are in sync, both of them point to equivalent slots in the lip-value memory. Therefore, we can obtain lip features by using key addresses to retrieve information saved in the lip-value memory:

$$\hat{f}_{lip;keyAdr} = A_{aud} \cdot \mathbf{M}_{lip}, \tag{9}$$

where $\hat{f}_{lip;keyAdr} \in \mathbb{R}^C$ is the recalled lip feature from key address. It contains lip movement related features corresponding to the input audio, acting as a strong bridge between audio and video in synthesizing the mouth region. As the decoder can take advantage of the additional visual hints on the lip movements, both visual quality and lip-sync quality can be enhanced. Also, learning audio-visual alignment earlier in the generation step imposes a stronger lip synchronization.

Video Synthesis

Identity encoder extracts identity feature f_I from a random reference frame concatenated with a pose-prior (target face with lower-half masked) along the channel axis. The pose-prior is crucial as it guides the model to generate the lower half mouth region that fits the upper half pose, reducing artifacts when pasting back to the original video (KR et al. 2019). The recalled lip feature from the key address is channel-wise concatenated with the audio feature and injected into the decoder G. The decoder has a U-Net-like architecture (Ronneberger, Fischer, and Brox 2015) with multi-scale intermediate features concatenated with those from the identity encoder, one after every up-sampling operation. This skip-connection is to ensure that the input identity and pose features are preserved.

At the inference time, we take recalled lip features from key addresses as shown in Fig.1 (b). At training, we additionally use lip features extracted directly from the lip encoder as shown in Fig.1 (a),

$$\hat{I}_g = G(\hat{f}_{lip;keyAdr} \oplus f_{aud}, f_I), \tag{10}$$

$$\hat{I}_G = G(f_{lip} \oplus f_{aud}, f_I), \tag{11}$$

where $\hat{I}_g$ is a frame generated with a recalled lip feature from a key address and $\hat{I}_G$ is generated with a lip feature directly from the lip encoder. Although only $\hat{I}_g$ is used at inference, we additionally adopt $\hat{I}_G$ during training in loss computation so that the lip encoder learns to extract meaningful features related to the lip movement from the face sequence.

We design our generation loss functions to increase visual quality and lip-sync quality. Reconstruction loss and perceptual loss are pertinent to visual quality, and audio-visual sync loss and visual-visual sync loss are related to lip-sync quality. Note that we compute generation loss with regards to both $\hat{I}_g$ and $\hat{I}_G$.

Reconstruction Loss The network is trained to minimize L1 reconstruction loss between the generated frames and ground truth frames I as follows:

$$\mathcal{L}_{recon} = \frac{1}{N} \sum_{i=1}^{N} (\|\hat{I}^i_g - I^i\|_1 + \|\hat{I}^i_G - I^i\|_1). \tag{12}$$

Generative Adversarial Loss We employ GAN loss (Goodfellow et al. 2014) to evaluate image realism. L1 reconstruction alone can yield blurry images or slight artifacts as it is a pixel-level loss.

$$\mathcal{L}_{gan} = \mathbb{E}_{\hat{I} \in [\hat{\mathbf{I}}_G, \hat{\mathbf{I}}_g]}[\log(1 - D(\hat{I}))], \tag{13}$$

$$\mathcal{L}_{disc} = \mathbb{E}_I[\log(1 - D(I))] + \mathbb{E}_{\hat{I} \in [\hat{\mathbf{I}}_G, \hat{\mathbf{I}}_g]}[\log D(\hat{I})]. \tag{14}$$

D is a quality discriminator trained on $\mathcal{L}_{disc}$, penalizing on unrealistic face generation. We adopt its architecture from (Prajwal et al. 2020). $\hat{I}$ is an image from a set of generated images with Eq. 10 and 11.

Method	LRW					LRS2				
	PSNR	SSIM	LMD	LSE-D	LSE-C	PSNR	SSIM	LMD	LSE-D	LSE-C
ATVGnet	31.409	0.781	1.894	7.664	5.735	30.427	0.735	2.549	8.223	5.584
3D Identity Mem	30.725	0.745	1.659	8.991	3.963	29.867	0.696	2.170	9.263	4.182
Wav2Lip	32.147	0.875	1.371	**6.617**	**7.237**	31.274	0.837	1.940	**5.995**	**8.797**
PC-AVS	30.440	0.778	1.462	7.344	6.420	29.887	0.747	1.963	7.301	6.728
Ground Truth	N/A	1.000	0.000	6.968	6.876	N/A	1.000	0.000	6.259	8.247
Ours ($\mathcal{L}_{a\text{-}v}$)	33.099	0.886	1.276	7.375	6.162	**32.681**	0.875	1.440	6.392	7.835
Ours ($\mathcal{L}_{v\text{-}v}$)	33.112	**0.893**	1.262	7.394	6.131	32.611	0.875	1.433	6.787	7.363
Ours ($\mathcal{L}_{a\text{-}v} + \mathcal{L}_{v\text{-}v}$)	**33.126**	**0.893**	**1.253**	7.013	6.619	32.529	**0.876**	**1.387**	6.352	7.925

Table 1: Quantitative results on LRW and LRS2 test sets. The best scores in each metric are highlighted in bold.

Audio-Visual Sync Loss We use the audio-visual sync module proposed in (Prajwal et al. 2020; Chung and Zisserman 2016b). We train the audio-visual sync module, $\mathcal{F}_a$ and $\mathcal{F}_v$, separately and do not fine-tune further on the generated frames so that it learns from clean pairs of audio and video segments. It takes a sequence of 5 generated frames (lower half only) and an audio segment a corresponding to the frame sequence. It outputs audio feature f_a and video feature f_v from which binary cross-entropy of cosine similarity is computed as follows:

$$d_{\text{sync}}(f_a, f_v) = \frac{f_a \cdot f_v}{\|f_a\|_2 \cdot \|f_v\|_2}, \tag{15}$$

$$\mathcal{L}_{\text{a-v}} = -\frac{1}{N}\sum_{i=1}^{N}(\log d_{\text{sync}}(\mathcal{F}_a(a_i), \mathcal{F}_v(\hat{\mathbf{I}}^{\mathbf{i}}_{\mathbf{g}})) \tag{16}$$

$$+ \log d_{\text{sync}}(\mathcal{F}_a(a_i), \mathcal{F}_v(\hat{\mathbf{I}}^{\mathbf{i}}_{\mathbf{G}}))), \tag{17}$$

where $\hat{\mathbf{I}}^{\mathbf{i}}_{\mathbf{g}} = \{\hat{I}^n_g\}^{i+2}_{n=i-2}$, and $\hat{\mathbf{I}}^{\mathbf{i}}_{\mathbf{G}} = \{\hat{I}^n_G\}^{i+2}_{n=i-2}$.

Visual-Visual Sync Loss We present visual-visual sync loss that can complement audio-visual sync loss by encouraging coherence in visual domain. Lip features from a sequence of generated frames and ground truth frames can be obtained from a lip encoder E_{lip}. As the lip encoder is trained to extract lip motion related features that can be aligned with audio features through $\mathcal{L}_{store}$ and $\mathcal{L}_{align}$, we can expect the lip encoder to act as a strong visual-visual sync module. We define visual-visual sync loss as L1 distance between the two features as follows:

$$\mathcal{L}_{v\text{-}v} = \frac{1}{N}\sum_{i=1}^{N}(\|E_{lip}(\hat{\mathbf{I}}^{\mathbf{i}}_g) - E_{lip}(\mathbf{I}^{\mathbf{i}})\|_1 \tag{18}$$

$$+ \|E_{lip}(\hat{\mathbf{I}}^{\mathbf{i}}_G) - E_{lip}(\mathbf{I}^{\mathbf{i}})\|_1). \tag{19}$$

We freeze the lip encoder to exclude the loss, $\mathcal{L}_{v\text{-}v}$, from training the lip encoder. By further aligning generated frames with ground truth frames, sophisticated synchronization in pixel level can be achieved.

Total Loss The final objective is as follow:

$$\mathcal{L} = \lambda_1\mathcal{L}_{recon} + \lambda_2\mathcal{L}_{a\text{-}v} + \lambda_3\mathcal{L}_{v\text{-}v} \tag{20}$$

$$+ \lambda_4\mathcal{L}_{gan} + \lambda_5\mathcal{L}_{store} + \lambda_6\mathcal{L}_{align}, \tag{21}$$

where λ_n is hyper-parameter weight.

Experiment

Experimental Settings

Dataset We train and evaluate on LRW (Chung and Zisserman 2016a) and LRS2 (Afouras et al. 2018) datasets. LRW is a word-level dataset with over 1000 utterances of 500 words. LRS2 is a sentence-level dataset with over 140,000 utterances. Both are from BBC News in the wild.

Metrics We evaluate results using PSNR, SSIM, LMD, LSE-D, and LSE-C. PSNR and SSIM measure visual quality and LMD, LSE-D, and LSE-C measure lip-sync quality. LMD is the distance between lip landmarks (detected using dlib (King 2009)) of ground truth frames and those of generated frames. LSE-C and LSE-D proposed by (Prajwal et al. 2020) are confidence score (higher the better) and distance score (lower the better) between audio and video features from SyncNet (Chung and Zisserman 2016b), respectively. LSE-C and LSE-D measure correspondence between audio and visual features while LMD directly measures visual to visual coherence. For a fair comparison, we evaluate the cropped region of the face based on the face detector used in ATVGnet (Chen et al. 2019).

Comparison Methods We compare our work with 4 state-of-the-art methods on talking face generation: ATVGnet (Chen et al. 2019), Wav2Lip (Prajwal et al. 2020), PC-AVS (Zhou et al. 2021) and 3D Identity Mem (Yi et al. 2020). ATVGnet generates frames conditioned on landmarks with an attention mechanism. Wav2Lip, utilized as a baseline, is a reconstruction-based method. PC-AVS employs modularized audio-visual representations of identity, pose, and speech content. 3D Identity Mem is a 3D model based method augmented with identity memory. We use open-source codes to train on the target dataset.

Implementation Details We process video frames to face-centered crops of size 128×128 at 25 fps and audio to mel-spectrogram of size 16×80. Mel-spectrograms are constructed from 16kHz audio, window size 800, and hop size 200. At the inference, we use the first frame as a reference frame and the upper half of the target frame as a pose-prior. Hyper-parameters are empirically set: λ_1 to 10, λ_2, λ_3, λ_4, λ_5, λ_6 all to 0.01, and κ to 16. We take Wav2Lip as a baseline model and add Audio-Lip Memory and lip encoder

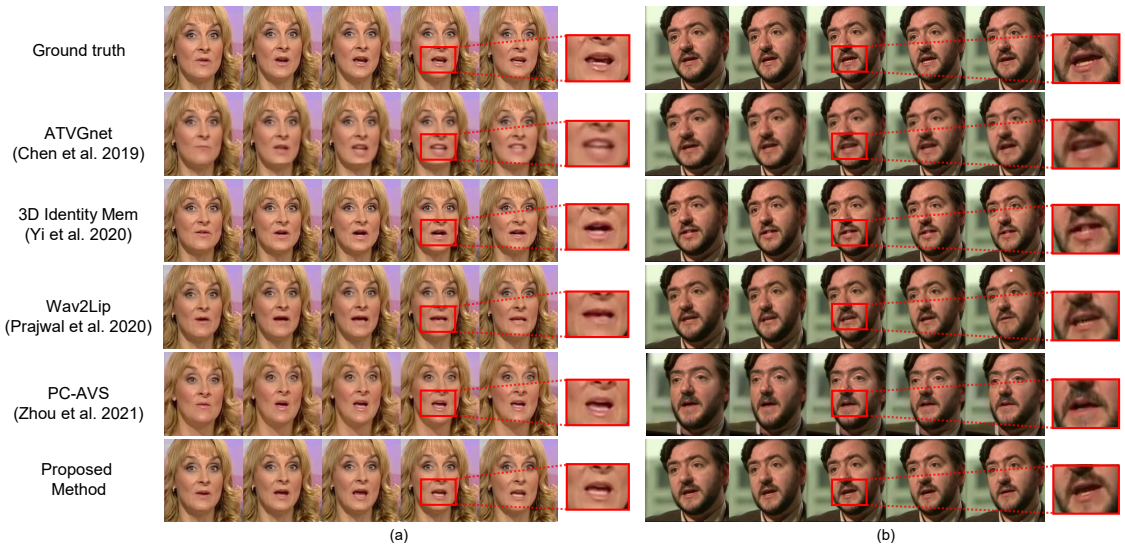

Figure 2: Comparison with state-of-the-art methods for talking face generation. Focusing on the red boxed regions, our method generates mouth that best aligns with the ground truth.

Method	Visual Quality	Lip-Sync Quality	Realness
Ground Truth	4.713 ±0.091	4.871 ±0.052	4.876 ±0.041
ATVGnet	2.059 ±0.284	2.515 ±0.448	1.803 ±0.473
3D Identity Mem	2.132 ±0.399	1.829 ±0.490	1.400 ±0.505
Wav2Lip	3.239 ±0.446	3.929 ±0.506	3.679 ±0.592
PC-AVS	3.108 ±0.444	3.471 ±0.491	3.095 ±0.541
Ours	**3.582** ±0.338	**4.226** ±0.401	**3.934** ±0.480

Table 2: Human evaluation by mean opinion scores with 95% confidence interval on visual quality, lip-sync quality, and video realness.

which consists of a 3D convolutional layer followed by 2D convolutional layers to encode lip motion feature. We empirically find the optimum slot size to be 96. We first pre-train SyncNet on the target dataset and then train the framework with total loss $\mathcal{L}$ with the Adam optimizer using PyTorch. The learning rate is set to 1×10^{-4}, except for the discriminator, whose is 5×10^{-4}. We train on 8 RTX 3090 GPUs and Intel Xeon Gold CPU.

Experimental Results

Quantitative Results Table 1 shows the quantitative comparison between other methods on LRW and LRS2 datasets. Our model generates faces with the highest PSNR, SSIM, and LMD on both datasets. Wav2Lip performs better on LSE-D and LSE-C metrics and even outperforms those of ground truth. However, as noted in (Zhou et al. 2021), it only proves that their lip-sync results are nearly comparable to the ground truth, not better. Our LSE-D and LSE-C scores are indeed closer to the ground truth scores and we perform better on the LMD metric which is another sync metric that measures correspondence in the visual domain. To quantify the effect of our visual-visual sync loss, we have conducted experiments using different combinations of the sync loss. As shown in Table 1, $\mathcal{L}_{a\text{-}v}$ has better LSE-D and LSE-C than $\mathcal{L}_{v\text{-}v}$ while $\mathcal{L}_{v\text{-}v}$ is better on PSNR, SSIM, and LMD

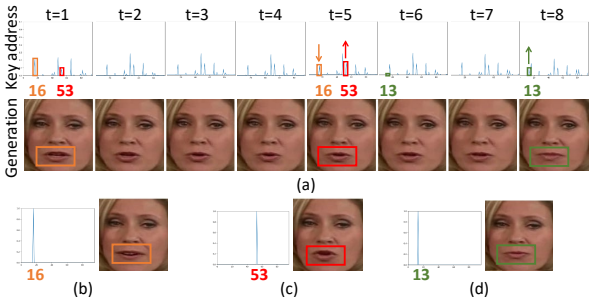

Figure 3: (a) Key addresses from audio input and corresponding generated frames in a sequence. (b), (c), and (d) Generated frames using slots 16, 53, and 13 respectively that noticeably changed its values in (a).

in overall on both datasets. Such result makes sense as $\mathcal{L}_{a\text{-}v}$ is relevant to audio-visual synchronization that LSE-D and LSE-C measure while $\mathcal{L}_{v\text{-}v}$ indicates visual-visual synchronization that PSNR, SSIM, and LMD measure. It is more important to note that the two sync losses combined together yield the best performance overall on both datasets as they have complementary effects aligning different pairs of domains. Regardless of which sync loss was used, applying the memory always outperforms other methods on PSNR, SSIM, and LMD, because the memory explicitly provides visual information of the lip motion to the decoder to take advantage of.

Qualitative Results We compare our generation results against previous state-of-the-art methods in Fig.2. It shows that our method generates the highest quality video with mouth shapes that best match the ground truth. As ATVGnet and Identity Mem produce given one identity reference, there are restrictions to pose and expression variance so naturalness is seemingly low. PC-AVS fails to preserve the identity features of the target frame. Wav2Lip produces mouth shapes that do not exactly align with the ground truth,

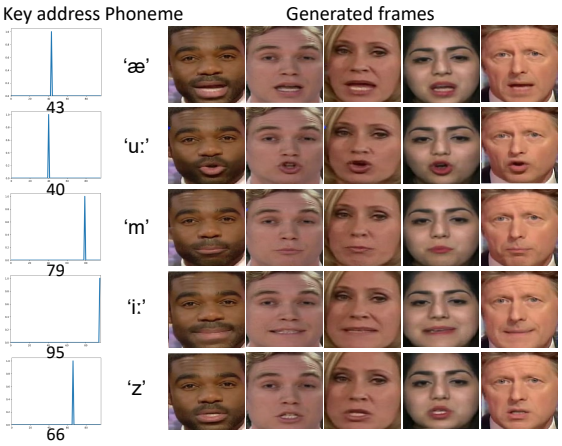

Figure 4: Generated frames using a single slot of Lip-Value Memory. Each slot contains a unique lip feature that can be associated with a phoneme.

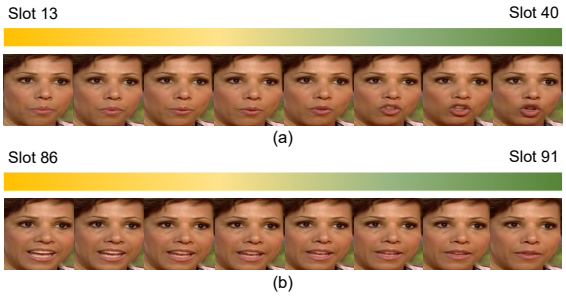

Figure 5: Interpolation of frames generated between two different slots of Lip-Value Memory. It shows that the memory address can specifically manipulate only mouth region in fine-grained level.

and there exist some artifacts. On the other hand, our method accurately captures the mouth shape including the teeth with high visual quality, as demonstrated by PSNR, SSIM, and LMD scores in Table 1. Such results can be contributed to the memory network allowing sophisticated and subtle lip generation and the two complementary sync losses aligning on audio and visual domains.

User study We conduct a user study to compare generation results. 20 videos are generated using each method, 10 from the LRW test set and 10 from the LRS2 test set. 20 participants were asked to rate generated videos including ground truths to evaluate visual quality, lip-sync quality, and realness in the range of 1 to 5. As shown in Table 2, the scores are consistent with the quantitative results. Our method outperforms all other methods on all three criteria, especially the lip-sync quality scores. Especially the lip sync quality scores high, demonstrating effectiveness of the audio-lip memory and visual-visual synchronization loss in improving temporal coherence.

Memory Analysis

We analyze elements stored in each slot in lip-value memory. Fig.3 shows key addresses and corresponding generated frames in a sequence. The key address is generated from an audio segment pertaining to the word 'North'. We can see that the address smoothly varies as lips move. Focusing on the slots that noticeably change their address value, from t=1 to t=5, the address on the 16th slot decreases from 0.218 to 0.164 while the 53rd slot address increases from 0.097 to 0.186. To visualize the lip feature stored at each slot, we generate with *silent audio* and a single slot addressed to the max as shown in Fig.3 (b), (c), and (d). We can see that a frame from slot 16 has lips drawn to the sides similar to the lip shape in t=1 and a frame from slot 53 has pursed lips as in t=5. Also, a frame from slot 13 has closed lips as in t=8 frame when the address on slot 13 suddenly increased. This result indicates that the memory well decouples lip features associated with speech sound and bestows memory with explicit control over the lip movement while keeping all other factors such as identity and pose unchanged.

Slots	PSNR	SSIM	LMD	LSE-D	LSE-C
24	32.522	0.873	1.458	6.442	7.831
48	32.373	0.873	1.431	6.379	7.838
96	32.529	**0.876**	**1.387**	**6.352**	**7.925**
120	**32.655**	0.873	1.469	6.475	7.785

Table 3: Ablation study on the number of slots

We further generate frames using a single slot in Fig.4. It is possible to assign each slot with a phoneme. For example, slot 43 closely aligns with 'æ', slot 40 'u:', slot 79 'm', slot 95 'i:' and slot 66 'z'. It demonstrates that each slot contains a unique lip feature at the phoneme level and that by taking combinations of the lip features in each slot through address, diverse lip movements can be generated.

We verify that lip shape can be smoothly interpolated between addresses on two different slots. As shown in Fig.5, as the ratio of the address varies from 13 to 40 in (a) and from 86 to 91 in (b), the lips change accordingly. Since the generation is very sensitive to the address value concerning the input audio, our method can generate subtle lip movements.

Lastly, we perform ablation study on using a different number of slots as shown in Table 3. The performance gradually increases from using 24 slots to 96 slots but decreases when the slot size is further increased to 120. This indicates that a large number of slots to hold many lip features does not linearly increase the performance because it may complicate the model in aligning key and value addresses. Thus, we empirically set the optimum slot size to 96.

Conclusion

Our proposed Audio-Lip Memory extracts the lip motion related features to bridge from audio to video generation. The lip synchronization is achieved during the memory learning that aligns the audio and visual lip features, and it is further enforced by the audio-visual and the visual-visual synchronization losses. We have verified that the lip features are stored in each memory slot at the phoneme level, and different combinations of the slots through memory addressing can yield diverse and subtle lip motions. Therefore, our work effectively exploits visual information of the mouth region to simultaneously achieve high visual quality and lip synchronization for talking face generation.

Acknowledgements

This work was partially supported by Genesis Lab under a research project (G01210312).

References

Afouras, T.; Chung, J. S.; Senior, A.; Vinyals, O.; and Zisserman, A. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*.

Cai, Q.; Pan, Y.; Yao, T.; Yan, C.; and Mei, T. 2018. Memory matching networks for one-shot image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4080–4088.

Chen, L.; Li, Z.; Maddox, R. K.; Duan, Z.; and Xu, C. 2018. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 520–535.

Chen, L.; Maddox, R. K.; Duan, Z.; and Xu, C. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7832–7841.

Chung, J. S.; and Zisserman, A. 2016a. Lip reading in the wild. In *Asian conference on computer vision*, 87–103. Springer.

Chung, J. S.; and Zisserman, A. 2016b. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, 251–263. Springer.

Chung, S.-W.; Chung, J. S.; and Kang, H.-G. 2019. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3965–3969. IEEE.

Das, D.; Biswas, S.; Sinha, S.; and Bhowmick, B. 2020. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *European Conference on Computer Vision*, 408–424. Springer.

Gao, R.; and Grauman, K. 2021. Visualvoice: Audio-visual speech separation with cross-modal consistency. *arXiv preprint arXiv:2101.03149*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Jamaludin, A.; Chung, J. S.; and Zisserman, A. 2019. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, 127(11): 1767–1779.

Kaiser, Ł.; Nachum, O.; Roy, A.; and Bengio, S. 2017. Learning to remember rare events. *arXiv preprint arXiv:1703.03129*.

Kim, J. U.; Park, S.; and Ro, Y. M. 2021. Robust Small-Scale Pedestrian Detection With Cued Recall via Memory Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3050–3059.

Kim, M.; Hong, J.; Park, S. J.; and Ro, Y. M. 2021. Multi-modality associative bridging through memory: Speech sound recollected from face video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 296–306.

Kim, S.; and Ro, Y. M. 2021. M-CAM: Visual Explanation of Challenging Conditioned Dataset with Bias-reducing Memory. In *The 32nd British Machine Vision Conference, BMVC 2021*. British Machine Vision Association (BMVA).

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10: 1755–1758.

KR, P.; Mukhopadhyay, R.; Philip, J.; Jha, A.; Namboodiri, V.; and Jawahar, C. 2019. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1428–1436.

Lee, S.; Kim, H. G.; Choi, D. H.; Kim, H.-I.; and Ro, Y. M. 2021. Video Prediction Recalling Long-term Motion Context via Memory Alignment Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3054–3063.

Lee, S.; Sung, J.; Yu, Y.; and Kim, G. 2018. A memory network approach for story-based temporal summarization of 360 videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1410–1419.

Miller, A.; Fisch, A.; Dodge, J.; Karimi, A.-H.; Bordes, A.; and Weston, J. 2016. Key-value memory networks for directly reading documents. *arXiv preprint arXiv:1606.03126*.

Mittal, G.; and Wang, B. 2020. Animating face using disentangled audio representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3290–3298.

Nagrani, A.; Chung, J. S.; Albanie, S.; and Zisserman, A. 2020. Disentangled speech embeddings using cross-modal self-supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6829–6833. IEEE.

Pei, W.; Zhang, J.; Wang, X.; Ke, L.; Shen, X.; and Tai, Y.-W. 2019. Memory-attended recurrent network for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8347–8356.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. 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*, 484–492.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Song, L.; Wu, W.; Qian, C.; He, R.; and Loy, C. C. 2020. Everybody's talkin': Let me talk as you want. *arXiv preprint arXiv:2001.05201*.

Song, Y.; Zhu, J.; Li, D.; Wang, X.; and Qi, H. 2018. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*.

Thies, J.; Elgharib, M.; Tewari, A.; Theobalt, C.; and Nießner, M. 2020. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision*, 716–731. Springer.

Vougioukas, K.; Petridis, S.; and Pantic, M. 2020. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128(5): 1398–1413.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10039–10049.

Weston, J.; Chopra, S.; and Bordes, A. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Yi, R.; Ye, Z.; Zhang, J.; Bao, H.; and Liu, Y.-J. 2020. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*.

Yu, L.; Yu, J.; Li, M.; and Ling, Q. 2020. Multimodal Inputs Driven Talking Face Generation With Spatial–Temporal Dependency. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(1): 203–216.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021. 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*, 3661–3670.

Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; and Wang, X. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9299–9306.

Zhou, H.; Sun, Y.; Wu, W.; Loy, C. C.; Wang, X.; and Liu, Z. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4176–4186.

Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; and Li, D. 2020. MakeltTalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6): 1–15.

Zhu, H.; Huang, H.; Li, Y.; Zheng, A.; and He, R. 2020. Arbitrary Talking Face Generation via Attentional Audio-Visual Coherence Learning. 2334–2340.

Zhu, M.; Pan, P.; Chen, W.; and Yang, Y. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5802–5810.