LEARNING POSE-ADAPTIVE LIP SYNC WITH CASCADED TEMPORAL CONVOLUTIONAL NETWORK

Ruobing Zheng *Bo Song* *Changjiang Ji*

Deep Innovation R&D Center
Moviebook

ABSTRACT

Speech-driven lip sync has become a promising technique for generating and editing talking-head videos. These studies mainly use 3D morphable models or 2D facial landmarks as the intermediate face representations. However, 2D-based methods have been stagnant recently due to their inability to handle out-of-plane rotations, even though the 2D landmarks have the advantage of fast and accurate extraction. In this paper, we design a cascaded temporal convolutional network to successively generate mouth shapes and corresponding jawlines based on audio signals and template headposes. Instead of explicitly calibrating the rotation between the predicted mouth and the template face, we employ neural networks to learn the pose-adaptive mapping implicitly. We also propose an image-to-image translation-based neural rendering method for producing high-resolution and photo-realistic videos. Experiments show our solution improves both the mapping accuracy and visual performance than baselines. This work could benefit many real-world applications like virtual anchors, telepresence, and conversational agents.

Index Terms— Lip sync, temporal convolutional network, speech process, virtual anchor

1. INTRODUCTION

Speech-driven lip sync works on synthesizing the video of the mouth area and then incorporating it into a head template from other stock footage [1]. The common two-step strategy is first to generate mouth representations from speech content and then synthesize photorealistic appearance [2, 3]. Early studies handle this task using Hidden Markov Models and computer graphics-based rendering techniques [4, 5], which yield compelling results but have the disadvantages of low efficiency. Recent advances in deep learning first boost lip-sync research at the audio-to-mouth stage. RNN-based architectures [3, 6] facilitates the learning of the sequential mapping from audio signals to mouth movements. In the rendering stage, *ObamaNet* [2] is a representative work that demonstrates the power of neural networks [7] in synthesizing the photorealistic appearance. These methods [8, 6] provide a fully-trainable solution for the classical two-stage lip-sync scheme, which significantly improves both the lip-sync accuracy and processing efficiency.

Current neural lip-sync methods mainly use the parameter space, rather than the full pixels, as the target space for learning the audio-to-mouth mapping [9]. The main options are 2D Facial Landmarks (2DFLs) and 3D Morphable Models (3DMMs). 3DMMs have the strength of controllable parameters and free rotation [8, 6]. However, they usually [10, 11] require landmarks for registration, and their parametric nature leads to a weakening of personal talking style. 2DFLs has the advantage of fast and accurate extraction, which also represent the most original facial shapes. But the inability to calibrate out-of-plane rotation is its inherent flaw. This shortcoming significantly affects lip-sync accuracy when facing the target person with rich head motions.

To tackle the above issue, we use the template information to facilitate the mouth generation. In this paper, we design a cascaded temporal convolutional network to learn the pose-adaptive mapping from audio signals to mouth movements. This seq-to-seq model successively generates the mouth shape and corresponding jawline, from the input combining audio signals, target headposes, and target cheek keypoints. Instead of explicitly removing and recovering the rotation, we predict the pose-adaptive landmarks that match the head movement in template videos. The corresponding jawline is generated to assist the rendering module in producing the accurate mouth texture. In cooperation with the mouth generation network, we further propose a neural rendering method to produce photorealistic appearances. We also conduct a comprehensive evaluation of the proposed method and baseline methods.

2. PROPOSED METHOD

The proposed method takes a piece of speech and template footage as input and outputs a synthesized video (Fig. 1). It can be interpreted in the following two stages:

Stage 1: from audio to pose-adaptive landmarks. We train a cascade neural network to learn the seq-to-seq mapping from audio features to 29 talking-related facial landmarks. The headpose from template frame is used as the conditional information in mapping network.

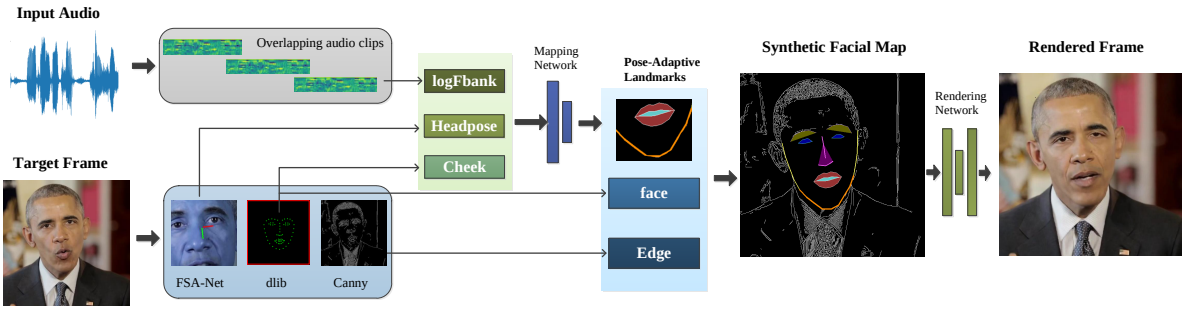

Fig. 1. Flowchart of the proposed lip-sync method. We predict the pose-adaptive landmarks in the first stage and then render the synthetic facial maps to high-resolution photo-realistic videos in the second stage.

stage 2: from landmarks to videos. We render the generated landmarks into final videos via a tailored neural rendering method. We first incorporate the generated landmarks into the face template in each target frame. The synthesized facial maps then go through the rendering network and turn into high-resolution photo-realistic video frames.

2.1. Audio to Pose-Adaptive Landmarks

In this stage, we translate the speech content into the talking-related landmarks via a cascaded Temporal Convolutional Network (TCN), as shown in Fig. 2. The main components are three-fold:

Pose-adaptive landmarks: Recent 2D-based methods [2, 1] employ facial landmarks to represent mouth movements. Their fatal drawback is the inability to completely handle the head rotations. *ObamaNet* [2] removes the in-plane rotation but fails to handle the out-of-plane rotation. The consequence is that the generated mouth is always facing the front even though the target head has big rotations (see Fig. 4(a)). Our solution is to implicitly predict the landmarks that conform to the template head posture. As shown in Fig. 2, along with the audio input, we import the template headposes to the mapping network twice for learning the pose-adaptive mouth shape and jawline.

TCN block: RNN-based models are widely used in lip-sync studies [12, 6]. But a recent report [13] shows that TCN outperforms generic RNN in various tasks. Compared to RNN, TCN has the advantages of large sequential perceptive field, stable gradients, and low memory requirements. To bring such strengths into the lip-sync scenario, we employ TCN blocks as the main component of the mapping network. Within the TCN block, the non-causal structure covers both future and past information, similar to the time-delay LSTM [1] and Bi-directional LSTM [6]. This setting meets the report that the mouth movement is not only determined by the past but also the future sounds [1]. Besides, the dilated convolution [14] provides an exponentially large reception field while preserving stable and fast gradient calculations. Both TCN blocks have the dilation factor of [1,2,4,8], and kernel size equals 3.

Cascaded structure: Jaw movement is closely related to mouth shape [15], but it is inherently difficult to reconstruct the jawline in 2D landmarks. To this end, we design a cascaded structure to successively generate the mouth shape and corresponding jawline, as shown in Fig. 2. The first-phase network G_1 takes as input the audio features ($\mathbf{a}_i \in \mathbb{R}^{256\times 26}$) and the target headpose ($\mathbf{p}_i \in \mathbb{R}^{64\times 3}$). It produces an intermediate output $G_1(\mathbf{a}_i, \mathbf{p}_i)$ which is compared with the PCA mouth features ($\mathbf{m}_i \in \mathbb{R}^{64\times 13}$). The inputs of the second-phase network G_2 consist of the first-phase output $G_1(\mathbf{a}_i, \mathbf{p}_i)$, the target headpose $\mathbf{p}_i$, and the cheek keypoints ($\mathbf{f}_i \in \mathbb{R}^{64\times 16}$) in the target frame. The cheek contains eight facial landmarks, which are important for locating the generated mouth shape and blending the jawline into the template face. The G_2 finally outputs 29 talking-related landmarks ($\mathbf{t}_i \in \mathbb{R}^{64\times 58}$), including 20 mouth keypoints and 9 jaw keypoints. In summary, we first generate the main mouth features with G_1 and then use G_2 to produce final pose-adaptive landmarks of mouth and jawline. The two networks are training together under the following loss.

We designed a composite loss function to cooperate with the above structure. The whole cascaded network is defined as G , and the training data pair is $\{(\mathbf{a}_i, \mathbf{p}_i, \mathbf{m}_i, \mathbf{f}_i, \mathbf{t}_i)\}_{i=1}^{F}$. We first employ the L_2 regression loss to measures the first-phase mapping accuracy of the main mouth features:

$$\mathcal{L}_{pca} = \|\mathbf{m}_i - G_1(\mathbf{a}_i, \mathbf{p}_i)\|_F^2 \tag{1}$$

Then the same L_2 loss is used on the final output 29 talking-related landmarks:

$$\mathcal{L}_{L2} = \|\mathbf{t}_i - G_2(G_1(\mathbf{a}_i, \mathbf{p}_i), \mathbf{p}_i, \mathbf{f}_i)\|_F^2 \tag{2}$$

We also use a pairwise inter-frame loss improve the temporal stability in final generated landmarks. The the L_2 distance between the differences of consecutive frames is calculated as:

$$\mathcal{L}_{int} = \left\|(\mathbf{t}_i - \mathbf{t}_{i-1}) - \left(G_{2_i} - G_{2_{i-1}}\right)\right\|_F^2 \tag{3}$$

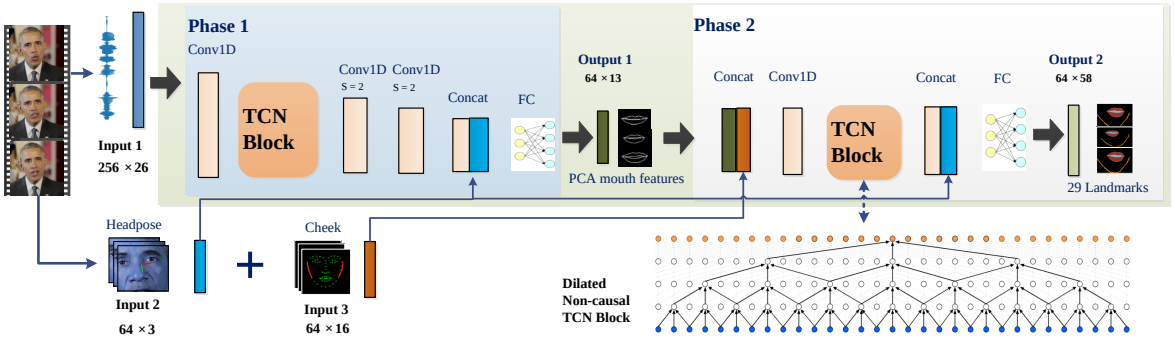

Fig. 2. Architecture of the cascaded temporal convolutional network for learning the audio-to-landmarks mapping. We import the headpose in both Phase 1 and Phase 2 to predict the pose-adaptive talking-related landmarks.

The final loss function ($\lambda_1 = 0.05, \lambda_2 = 0.6$) can be formed as:

$$\mathcal{L}_{all} = \mathcal{L}_{L2} + \lambda_1 \mathcal{L}_{pca} + \lambda_2 \mathcal{L}_{int} \tag{4}$$

2.2. Neural Rendering

We further propose a neural rendering strategy for producing a high-resolution and photoreal appearance. To avoid the mismatch between the generated mouth and the template face (see in Fig. 3(a)), we design a synthetic facial map for generating facial texture at the full-face level. We first recover the generated talking-related landmarks in the template face. Then we augment the composite landmarks with Canny edges extracted from the template frame. The edge information facilitates the rendering model to produce accurate and continuous details. Figure 3(c) shows the same mouth with different jawlines result in different rendered results, which confirm the necessity of accurate jawlines. We further build the rendering model by augmenting hierarchical image-to-image translation model *pix2pixHD* [16] with self-attention blocks. In subsequent experiments, we successfully rendered a 720 $\times$ 720 resolution head images based on the facial maps.

3. EXPERIMENTS

3.1. Dataset and pre-processing

The experiments were conducted on the weekly videos of ex-President Barack Obama [2]. We process 2-hour videos for training both the mapping and rendering networks. The resolution of original videos is 1080 $\times$ 720 and we crop the face area in 720 $\times$ 720. We extract the 68 2D-landmarks from each frame. We calculate the relative coordinates based on the center of the nose and do not explicitly remove any rotations. The 29 normalized landmarks are predicted and then recovered in given nose centers and face scale. As for audio, we extract 26-D *logfBank* features.

3.2. Results

In cascaded TCN, we import the headpose in both phases and the cheek information only in the second phase. We first explore the contribution of these conditional variables via ablation study. Then, we compare our model with two representative RNN-based lip-sync baselines (LSTM [2] and BiLSTM [6]) with only mouth output. We record the MSE of predicted mouth and jawline respectively, as well as the training time and inference time (5min audio).

Table 1 shows that headposes in both phases are helpful to improve the mapping accuracy. Adding headpose in Phase 1 brings a greater lift than adding it alone in Phase 2, showing that learning a pose-adaptive PCA mouth is the basis for accurately generating the final talking-related landmarks. The result also indicates that inputting the headpose again in Phase 2 is necessary to calibrate the jawline and fine-tune the mouth shape. Besides, we find that the lack of cheek information dramatically affects the mapping accuracy for both mouth and jawline. We infer that the large MSE comes from the wrong location. Therefore, cheek information is essential for both locating the predicted mouth and connecting the jawline to the template face. Moreover, we find a positive correlation between training time and the increase in conditional inputs. Compared to RNN-based baselines, all TCN-based models have a significant advantage of time consumption, especially for inference time.

Fig. 4 shows the impact of headpose and cheek information on generating landmarks and final rendered results. In Figure 4(a), the middle row generates the pose-independent mouth shape. We find that the generated mouth is obviously wrong when encountering such a big out-of-plane head rotation, and the rendering result is also affected. Fig. 4(b) shows the problem caused by generating jawlines without cheek information. Although the rendered model has some corrective power, the jaw position is still clearly wrong in the rendering result. Fig. 5 shows the final rendered results of our lip-sync methods. The generated mouths movements are consistent

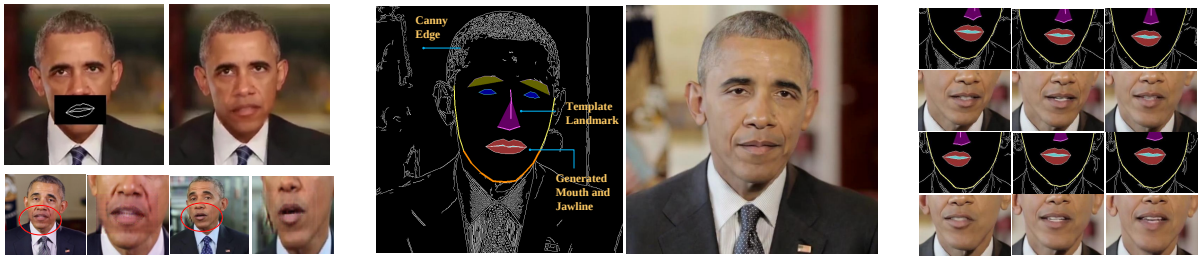

Fig. 3. (a) The inpainting strategy used in *ObamaNet*. (b) The full-face generation used in our method. (c) The different rendered results from the same mouth with different jawlines.

Table 1. Ablation study on the headpose, cheek used in the mapping network, and the comparison to baseline models.

	No pose	G1 pose	G2 pose	Ours	No cheek	LSTM	BiLSTM
Mouth (10^{-6})	4.095 ± 0.086	3.882 ± 0.047	4.004 ± 0.027	**3.628** ± **0.058**	69.71 ± 0.426	12.94 ± 0.278	12.93 ± 0.343
Jawline (10^{-6})	2.955 ± 0.051	2.778 ± 0.031	2.939± 0.017	**2.723** ± **0.037**	116.5 ± 0.283	–	–
Train (m)	58.86 ± 7.26	73.54 ± 4.08	64.92 ± 8.86	82.57 ± 10.87	**21.12** ± **0.73**	131.6 ± 23.8	169.3 ± 26.5
Infer (s)	0.043 ± 0.004	0.045 ± 0.003	0.044 ± 0.004	**0.043** ± **0.004**	0.043 ± 0.002	7.043 ± 0.338	12.44 ± 0.214

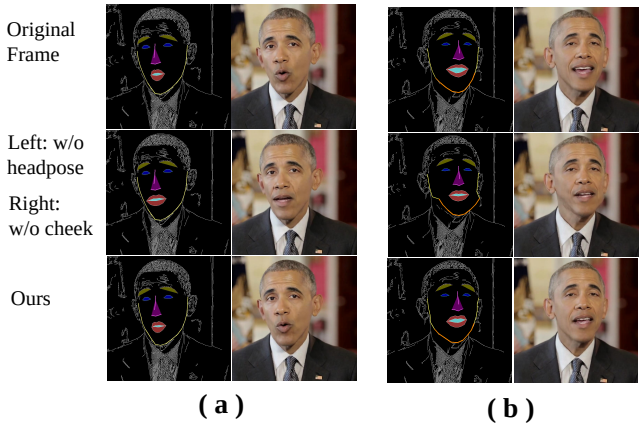

Fig. 4. Typical results of missing headpose (a), cheek information (b) at the audio-to-landmarks mapping stage.

with the audio source, while fitting well with the headpose and skin texture of template frames.

4. CONCLUSION

We present a neural-based framework for learning high-fidelity and pose-adaptive lip sync from speech. We hope our solution will make it possible to bring such methods into real-world applications.

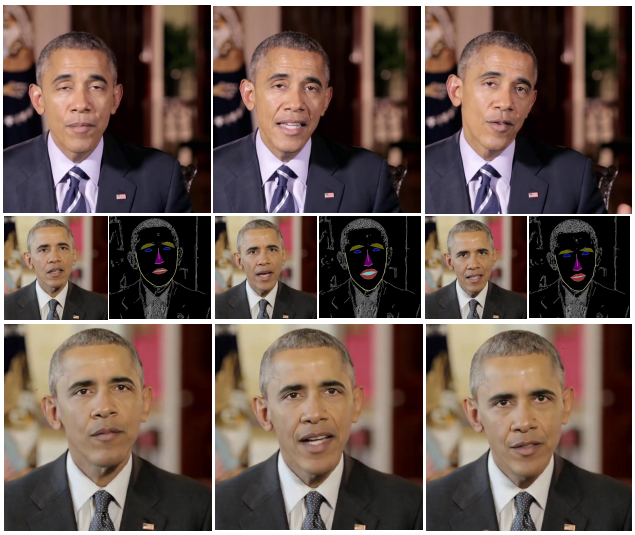

Fig. 5. Final results of the proposed lip-sync method. The first row shows the reference frames from the audio-source video. The second row shows the template frames and corresponding synthetic facial maps. The last row presents the final video frames rendered at 720 resolution.

5. REFERENCES

[1] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman, "Synthesizing obama: learn-

ing lip sync from audio," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 95, 2017.

[2] Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brébisson, and Yoshua Bengio, "Obamanet: Photo-realistic lip-sync from text," *arXiv preprint arXiv:1801.01442*, 2017.

[3] Bo Fan, Lijuan Wang, Frank K Soong, and Lei Xie, "Photo-real talking head with deep bidirectional lstm," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4884–4888.

[4] Carlos Busso, Zhigang Deng, Michael Grimm, Ulrich Neumann, and Shrikanth Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1075–1086, 2007.

[5] Lijuan Wang, Wei Han, Frank K Soong, and Qiang Huo, "Text driven 3d photo-realistic talking head," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[6] Guanzhong Tian, Yi Yuan, and Yong Liu, "Audio2face: Generating speech/face animation from single audio with attention-based bidirectional lstm networks," in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2019, pp. 366–371.

[7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2016.

[8] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner, "Neural voice puppetry: Audio-driven facial reenactment," *arXiv preprint arXiv:1912.05566*, 2019.

[9] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy, "Everybody's talkin': Let me talk as you want," *arXiv preprint arXiv:2001.05201*, 2020.

[10] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li, "Towards fast, accurate and stable 3d dense face alignment," *arXiv preprint arXiv:2009.09960*, 2020.

[11] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero, "Learning a model of facial shape and expression from 4d scans.," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.

[12] Lei Xiao and Zengfu Wang, "Dense convolutional recurrent neural network for generalized speech animation," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 633–638.

[13] Shaojie Bai, J Zico Kolter, and Vladlen Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[14] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[15] Jintao Jiang, Abeer Alwan, Patricia A Keating, Edward T Auer, and Lynne E Bernstein, "On the relationship between face movements, tongue movements, and speech acoustics," *EURASIP Journal on Advances in Signal Processing*, vol. 2002, no. 11, pp. 506945, 2002.

[16] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.