Lip Movements Generation at a Glance

Lele Chen, Zhiheng Li, Ross Maddox, Zhiyao Duan and **Chenliang Xu**
{lchen63, zli68, rmaddox}@ur.rochester.edu, {zhiyao.duan, chenliang.xu}@rochester.edu
University of Rochester

Introduction

- **Motivation**: We find that when we talk with someone face to face, we will understand other better than when we talk on the cellphone as we can see their lip movements. So it is crucial to develop a method that, e.g., can enhance speech comprehension while preserving privacy or assistive devices for hearing impaired people.
- **Objective**: In this paper, we consider a task of such: given an arbitrary audio speech and one lip image of arbitrary target identity, generate synthesized lip movements of the target identity saying the speech. To perform well in this task, it inevitably requires a model to not only consider the retention of target identity, photo-realistic of synthesized images, consistency and smoothness of lip images in a sequence, but more importantly, learn the correlations between audio speech and lip movements.

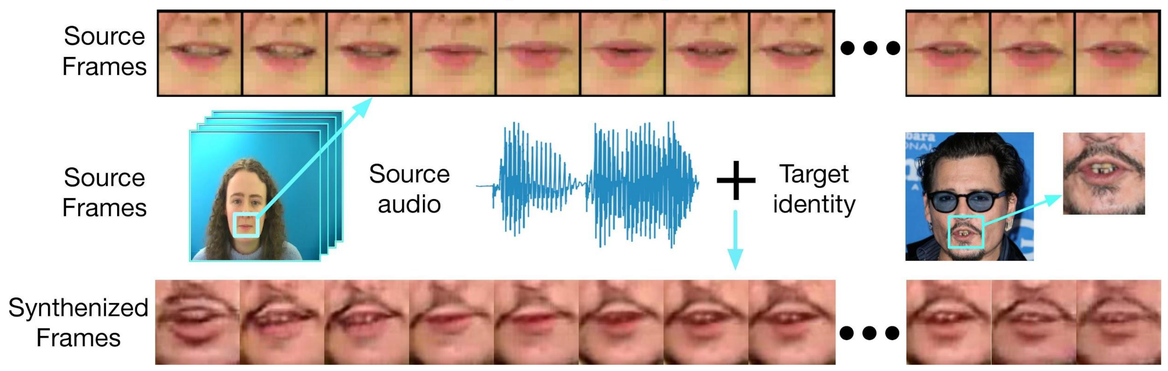

Fig.1: The model takes an audio speech of the women and one lip image of the target identity, a male celebrity in this case, and synthesizes a video of the man's lip saying the same speech.

Lip-Movement Generator Network

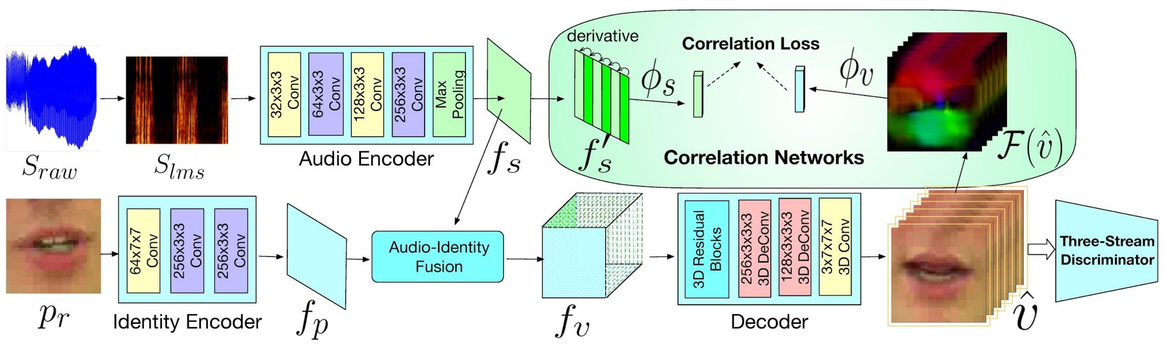

Fig.2: The overall diagram of full model

- **Audio encoder & identity encoder** extracts and fuses audio and visual embeddings.
- **Audio-Identity fusion network** fuses features from two modalities.
- **Decoder** expands fused feature to synthesized video.
- **Correlation Networks** are in charge of strengthening the audio-visual mapping.
- **Three-Stream discriminator** is responsible for distinguishing generated video and real video.

Audio-Visual Derivative Correlation & Loss

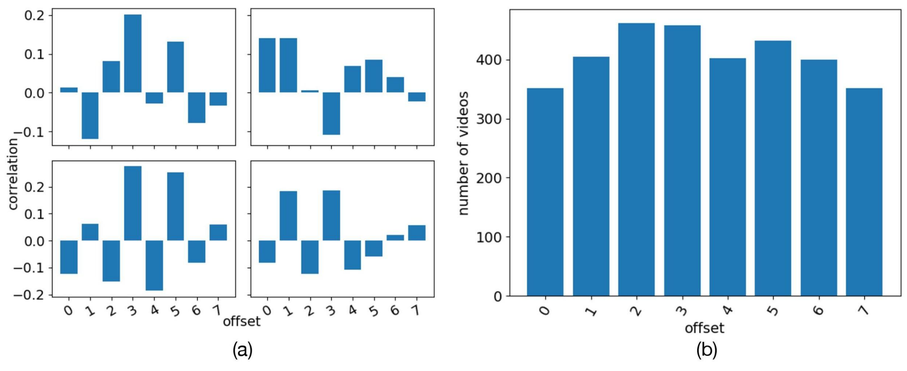

Fig.3: (a): Correlation coefficients with different offsets of four example videos. (b): Number of videos of different offsets with which the video has the maximum correlation coefficient. X-axes of both (a) and (b) stands for time steps of flow field shifted forward.

objective function: $\mathcal{L} = \ell_{corr} + \lambda_1 \ell_{pix} + \lambda_2 \ell_{per} + \lambda_3 \ell_{gen}$

- **Correlation loss:** ensures the correlation between audio and visual information.

$$\ell_{corr} = 1 - \frac{\phi_s(f'_s) \cdot \phi_v(\mathcal{F}(v))}{\|\phi_s(f'_s)\|_2 \cdot \|\phi_v(\mathcal{F}(v))\|_2}$$

- **Pixel-level reconstruction loss:** aims to make the model sensitive to speaker's appearance.

$$\ell_{pix}(\hat{v}, v) = \|v - \hat{v}\|$$

- **Perceptual loss:** utilizes high-level features to compare generated images and ground-truth images, resulting in better sharpness of the synthesized image

$$\ell_{perc}(\hat{v}, v) = \|\varphi(v) - \varphi(\hat{v})\|_2^2$$

- **Adversarial loss:** allows our model to generate overall realistic looking images

$$\ell_{gen} = -\log D([s^j, \hat{v}^j])$$

Qualitative Results

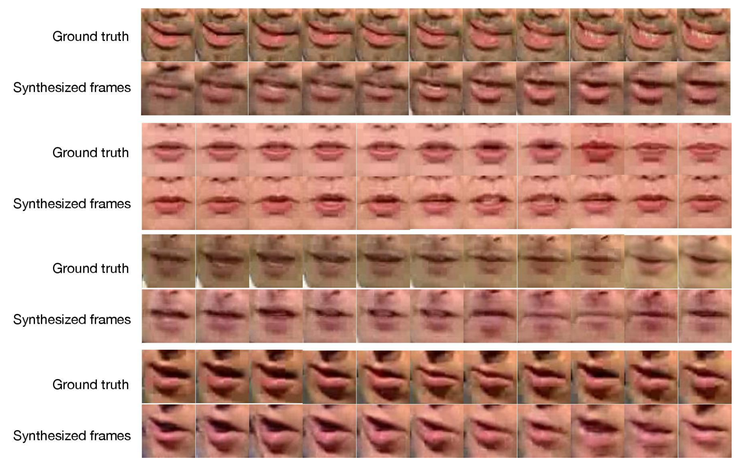

Fig.4: Randomly selected outputs of the full model on the LRW testing set. The lip shape in videos not only synchronize well with the ground truth, but maintain identity information, such as (beard v.s. no beard).

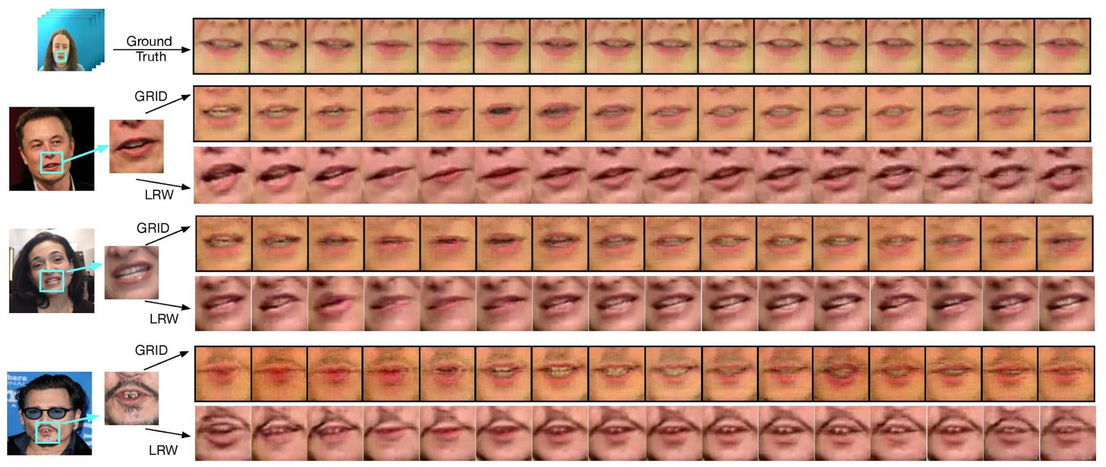

Fig.5: Generated images based on three identity images outside of dataset, which is also not paired with the input audio from GRID dataset. Two full models trained on GRID and LRW datasets are used here for a comparison.

Quantitative Evaluation

Methods	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
ℓ_{corr}					✓		✓		✓
ℓ_{corr}(Non-Derivative Corr.)						✓			
ℓ_{gen}	✓		✓	✓	✓	✓	✓	✓	✓
ℓ_{pix}		✓	✓	✓	✓	✓	✓	✓	✓
ℓ_{perc}				✓	✓	✓	✓	✓	✓
Two-Stream D.							✓	✓	
Three-Stream D.	✓		✓	✓	✓	✓			
Three-Stream D.(Frame-Diff.)									✓
Metrics									
LMD	1.37	1.28	1.33	1.31	**1.18**	1.96	1.39	1.42	1.40
SSIM	0.67	**0.79**	0.66	0.70	0.73	0.52	0.68	0.59	0.66
PSNR	29.46	29.81	29.66	29.40	**29.98**	28.6	29.59	29.46	29.51
CPBD	0.176	0.006	0.182	0.209	0.175	**0.218**	0.187	0.176	0.210

Tab.1: Ablation results on GRID dataset. The full model (method (e)) uses all four losses as described in Sec.4. For LMD, the lower the better. SSIM, PSNR and CPBD, the higher the better. We bold each leading score.

Method	GRID				LDC				LRW			
	LMD	SSIM	PSNR	CPBD	LMD	SSIM	PSNR	CPBD	LMD	SSIM	PSNR	CPBD
G. T.	0	N/A	N/A	0.141	0	N/A	N/A	0.211	0	N/A	N/A	0.068
Vondrick [17]	2.38	0.60	28.45	0.129	2.34	**0.75**	27.96	0.160	3.28	0.34	28.03	0.082
Chung [10]	1.35	**0.74**	29.36	0.016	2.13	0.50	28.22	0.010	2.25	0.46	28.06	**0.083**
Full model	**1.18**	0.73	**29.89**	**0.175**	**1.82**	0.57	**28.87**	**0.172**	**1.92**	**0.53**	**28.65**	0.075

Tab.2: Results on three datasets compared with State-of-the-Art methods. Models mentioned in this table are trained from scratch (no pre-training included) and be tested on each dataset a time. We bold each leading score.

Acknowledgement: This work was supported in part by NSF BIGDATA 1741472 and the University of Rochester AR/VR Pilot Award. We gratefully acknowledge the gift donations of Markable, Inc., Tencent and the support of NVIDIA with the donation of GPUs used for this research. This article solely reflects the opinions and conclusions of its authors and not the funding agents.

Paper details: https://arxiv.org/abs/1803.10404.

Demo video: https://www.youtube.com/watch?v=7IX_sIL5v0c.

Reference

10. Chung, J.S., Jamaludin, A., Zisserman, A.: You said that? In BMVC. (2017)
17. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: NIPS. (2016)