Talking Head from Speech Audio using a Pre-trained Image Generator

Mohammed M. Alghamdi[1,2], He Wang[1], Andrew J. Bulpitt[1], David C. Hogg[1]

[1] University of Leeds, [2] Taif University

mmfp1475

Project Page

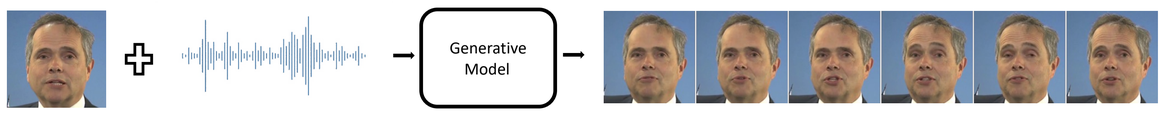
Contact: scmmalg@leeds.ac.uk

ACM Multimedia 2022 Lisbon, Portugal | 10-14 October

Overview

Goal:

- Synthesise a high-resolution talking-head video given an identity image and speech audio.

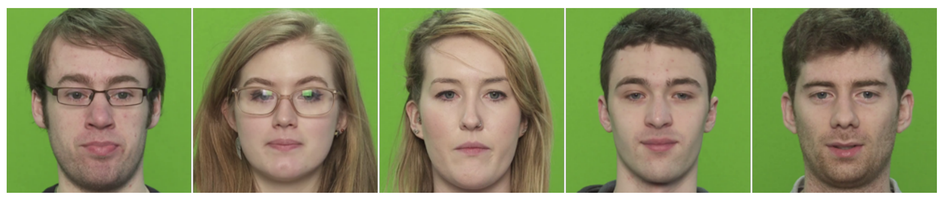


Approach:

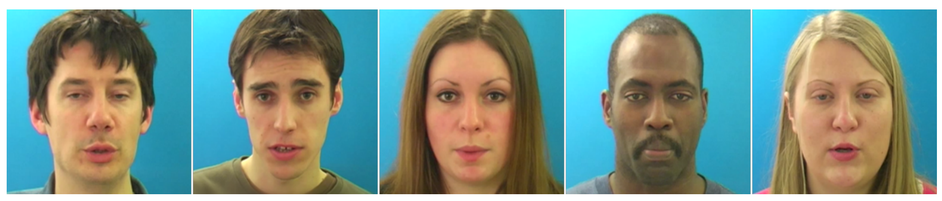
- Build on top of a pre-trained state of the art image generator (StyleGAN).
- Generate talking-head videos by finding motion trajectories in the latent space of StyleGAN conditioned on the speech audio.

Datasets

- **TCD-TIMIT:** 59 speakers each uttering 100 sentences

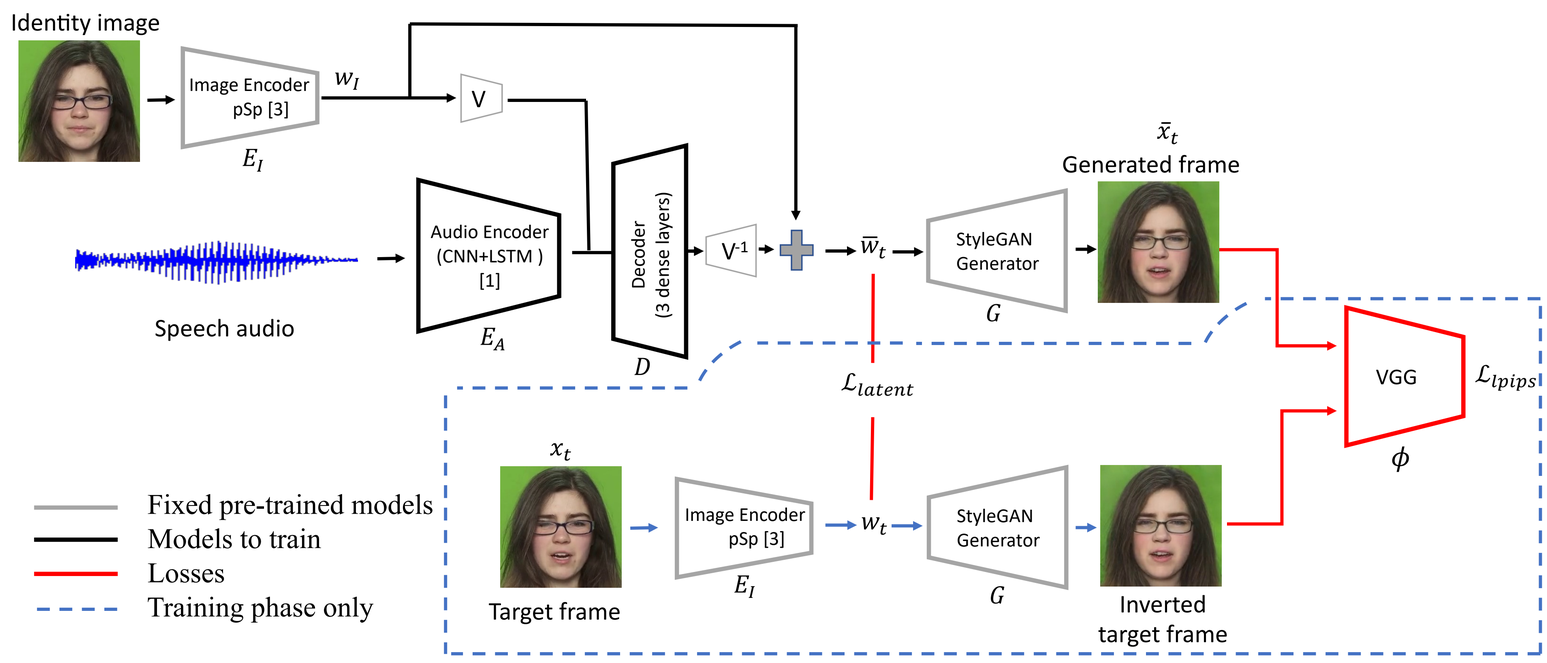


- **GRID:** 33 speakers each uttering 1000 sentences



Method

Stage 1:

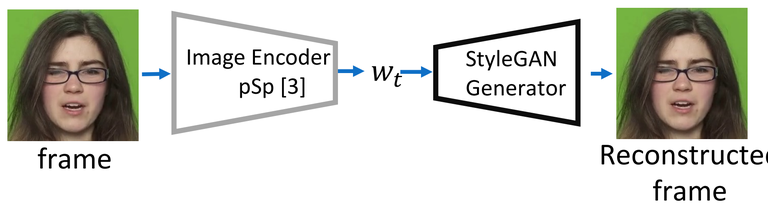


$$\mathcal{L}_{\text{latent}} = \sum_{t=1}^{T} ||w_t - \bar{w}_t||_2$$

$$\mathcal{L}_{\text{lpips}} = \sum_{t=1}^{T} ||\phi(\bar{x}_t) - \phi(G(E_I(x_t)))||_2$$

Stage 2:

Improve the visual quality of the generated videos further by tuning the generator only on a single image or short video of a target subject using the PTI [4] method.



Evaluation

Quantitative comparisons

Method	TCD-TIMIT				GRID			
	PSNR↑	SSIM↑	FID↓	LMD↓	PSNR↑	SSIM↑	FID↓	LMD↓
Vougioukas, et al. [5]	17.24	0.60	16.05	3.42	16.72	0.62	13.58	3.08
Chen, et al. [1]	15.31	0.58	11.79	3.66	16.80	**0.69**	13.27	3.74
Zhou, et al. [6]	18.10	0.58	18.02	2.59	18.53	0.61	11.87	2.64
Prajwal, et al. [2]	18.26	0.64	15.24	2.19	17.83	**0.69**	11.11	**2.05**
Ours	**20.55**	**0.65**	**8.11**	**2.18**	**20.33**	0.65	**5.30**	2.18

Qualitative results

Qualitative comparisons

Ablation analysis

Method	TCD-TIMIT		
	PSNR↑	SSIM↑	LMD↓
w/o $\mathcal{L}_{\text{latent}}$	20.57	0.65	2.30
w/o $\mathcal{L}_{\text{lpips}}$	**20.78**	**0.66**	2.75
Stage 1	17.55	0.49	2.37
Proposed model	20.55	0.65	**2.18**

References

[1] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, 2019.

[2] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM Multimedia*, 2020.

[3] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, 2021.

[4] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021.

[5] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans. In *BMVC*, 2018.

[6] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: Speaker-aware talking-head animation. *ACM Trans. Graph.*, 2020.