ICML 40 Years International Conference On Machine Learning

ICML Workshop Artificial Intelligence & Human-Computer Interaction — AI & HCI

ZHEJIANG UNIVERSITY

ByteDance

Ada-TTA: Towards Adaptive High-Quality Text-to-Talking Avatar Synthesis

Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Chen Zhang, Xiang Yin, Zejun Ma Zhou Zhao

Zhejiang University | ByteDance

Motivation

- Recent progress in **Zero-shot Text-to-Speech (TTS)** make it possible to synthesize identity preserving speech given the input text and a short speech prompt.
- Advanced 3D Talking Face Generation (TFG) such as **GeneFace++** could synthesize high-quality audio-driven talking portrait video given only a few-minute long video.
- **Ada-TTA** combine the greatness of zero-shot TTS model and GeneFace++, so that the joint system could allow users to create a talking video with **only text input**.

- Ada-TTA achieves:
  - ➢ Identity-preserving speech synthesize.
  - ➢ Real-time talking video generation with only text input.

Methodology

- ➢ Overall Design: Zero-shot TTS + GeneFace++ (See Fig.1).
- ➢ VQGAN-based Zero-shot TTS (See Fig.2).

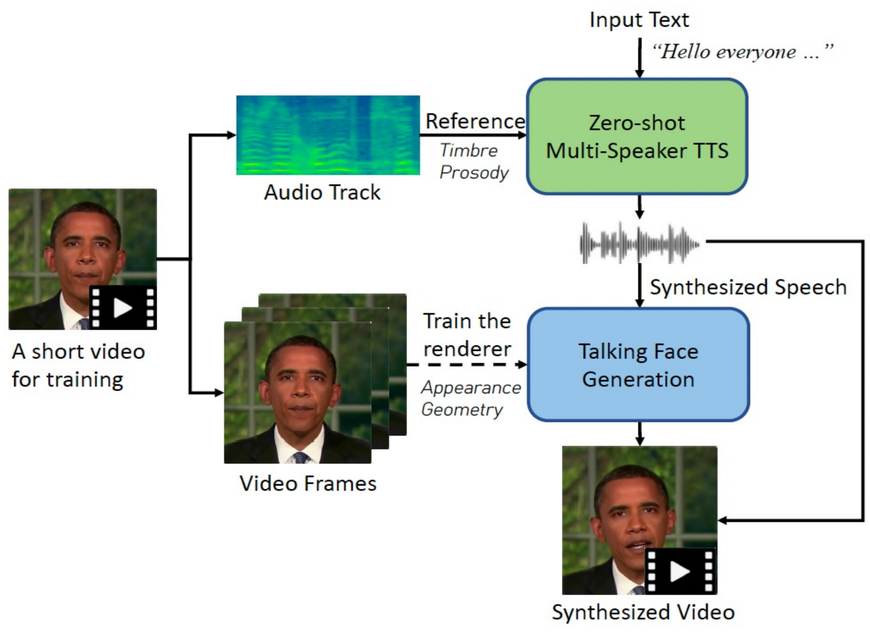

Fig.1. The overall pipeline of Ada-TTA

Experiments

- Visualized Results (See Fig.3 and the QR code.)
- Ada-TTA achieves better speaker similarity and image quality than the baseline.
- The users prefer our method in terms of audio and video quality.

Method	Spk-Sim↑	FID↓
YourTTS + Wav2Lip	0.9392	55.43
Ada-TTA (Ours)	**0.9854**	**28.36**

Fig.4. Objective evaluation of the TTA systems

Method	CMOS-A	CMOS-V	CMOS
Y+W	0.00	0.00	0.00
Ada-TTA	$+0.84 \pm 0.50$	$+0.76 \pm 0.42$	$+0.74 \pm 0.31$

Fig.4. Subjective evaluation of the TTA systems

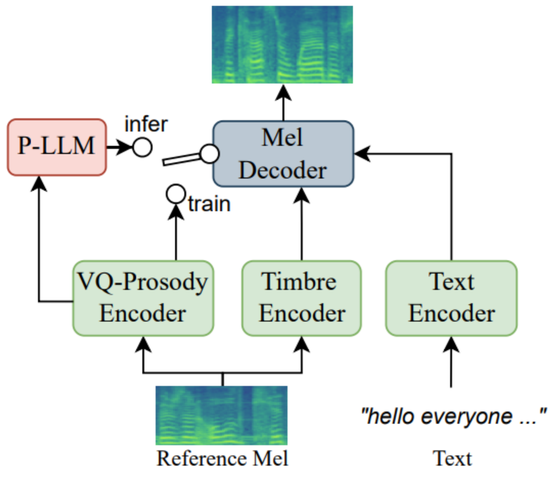

Figure 2: The overall structure of our internal zero-shot multi-speaker TTS model.

Fig.2. Design of Zero-shot TTS system

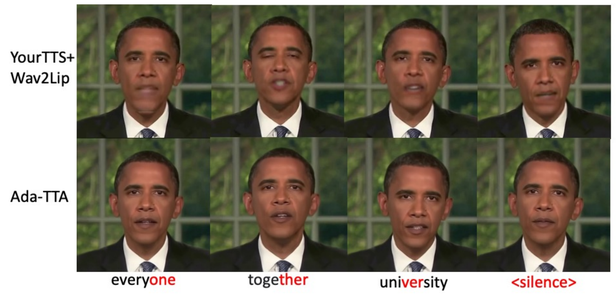

Fig.3. Visualized Results

Demo Video

Code for GeneFace