SUPPLEMENTARY MATERIAL

In this supplementary material, we provide more **qualitative analysis** about our audio and video results. In the enclosed folder, we also include a 8-minute **overall video** about our work, a **PyTorch implementation** of our method, as well as **examples** of voice conversion and audiovisual synthesis. We humbly urge the reviewers to watch the videos, to see and hear the synthesis results for themselves!

1 AUDIO CONVERSION

In the enclosed folder *AudioConversion*, we provide examples of audio conversion. We choose several recording sentences, and convert each of them to 10 target speakers (Alan Kay, Alexei Efros, Carl Sagan, Claude Shannon, John Oliver, Oprah Winfrey, Richard Hamming, Robert Iger, Stephen Hawking, and Takeo Kanade) in our CelebAudio dataset.

We also provide the multi-lingual results that drive John Oliver to speak Mandarin Chinese and Hindi. This indicates our exemplar autoencoders work on phonemes that are shared by different languages.

2 VIDEO SYNTHESIS

In the enclosed folder *VideoSynthesis*, we show several audiovisual samples in same-speaker generation and cross-speaker generation.

For same-speaker generation, we perform audio-to-video generation for one specific speaker. From the comparisons with Speech2vid (Chung et al., 2017) and LipGAN (KR et al., 2019), we can see clear artifacts in the face region. This is because those methods only generate the face region and paste it onto a still image (Speech2vid) or a footage video (LipGAN). When the morphed mouth shapes are very different from the footage, or there are a few dynamic facial movements, the pasting around face becomes "goofs". On the contrary, our approach generates full face and does not have those artifacts.

We also provide several failure cases for Speech2vid. This is due to the failure of facial registration, which is the first step of Speech2vid. This indicates Speech2vid is based on techniques of facial registration and struggles when such technique fails. However, our approach does not rely on any facial keypoints, and thus can generate even difficult samples (profile views, old archive footages etc.) successfully.

For cross-speaker generation, we input anyone's audio and output the audiovisual stream of a specific style, which is not possible by any other existing method.

For ablation analysis of video synthesis (sec D.4 in appendix), we conduct 3 comparative experiments: (a) Train only the audio-to-video translator. (b) Jointly train audio decoder and video decoder from scratch. (c) First train an audio autoencoder, then train video decoder. Here we provide concrete samples for qualitative analysis: (a) *onlyvideo.mp4* (b) *scratch.mp4* (c) *ours.mp4*. From the results, we note that if we only train the audio-to-video translator, the model cannot build the relationship between lip movements and speech (result of (a)). If we jointly train audio decoder and video decoder from scratch, we can see lip motion in the result, but still not perfectly consistent with the speech (result of (b)). If we use the whole model, the lip motion corresponds to the speech very well (result of (c)).

REFERENCES

Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017.

Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1428–1436, 2019.