Imperial College London

ICASSP2020 Barcelona

SAMSUNG

Video-driven Speech Reconstruction using Generative Adversarial Networks

Show & Tell Demo

Rodrigo Mira[1], Pingchuan Ma[1], Konstantinos Vougioukas[1], Stavros Petridis[1,2], Björn Schuller[1,3], Maja Pantic[1,2]

[1]Imperial College London
[2]Samsung AI Centre Cambridge
[3]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

Introduction

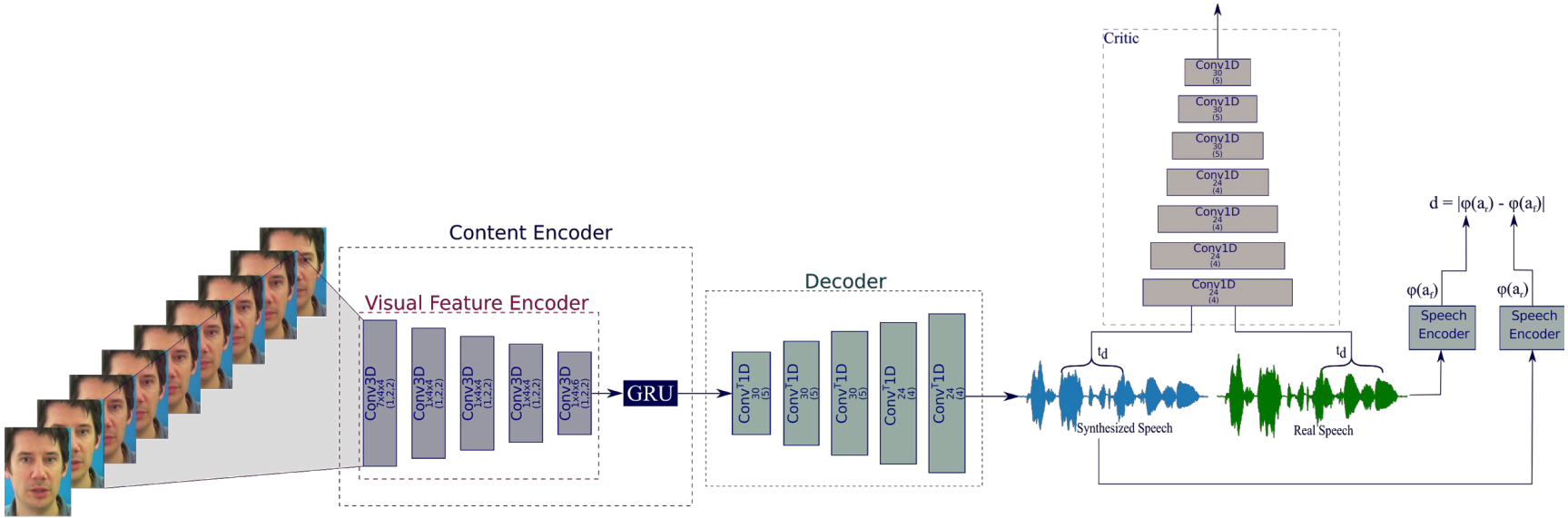
- In this presentation we will demonstrate our **end-to-end speech reconstruction** model on silent videos of unseen live speakers.
- This model is an **extension** of the one presented in ***Vougioukas et al. (2019)*** [1].
- We will be focusing on the **practical details** of the model and subsequently on applying it to live speakers.

[1] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic “Video-Driven Speech Reconstruction using Generative Adversarial Networks“ Interspeech 2019

Motivation

- **Lipreading** is a well developed technique which allows us to transcribe speech from video automatically when the corresponding audio is absent or noisy.
- **Video-to-speech** generates audio directly from video, which has 3 main advantages:
  - Can potentially be applied in **real time** with no delays.
  - Can potentially translate the **emotion and intonation** present in speech.
  - Does not require **transcribed datasets** for training.

Original Model (Interspeech 2019)



Original Model (Interspeech 2019)

- First deep learning model to generate waveform speech from silent video **end-to-end**.
- Features convolutional **encoder-decoder** model which encodes video into compact meaningful features which are then decoded into 16 kHz audio.
- Uses a **waveform critic** which discriminates real from fake samples in order to generate more realistic results.

Original Model (Interspeech 2019)

- The model is trained using **4 separate losses**:
  - **Adversarial loss**, based on ***I. Gulrajani et al. (2017)***[1].
  - **L1 Loss** between the real and synthesized waveforms.
  - **Total Variation Loss** for the synthesized waveform.
  - **Perceptual Loss**, an L1 Loss between the features extracted from the real and synthesized audio. The features are extracted using a pre-trained speech encoder based on ***K. Vougioukas et al. (2018)***[2].

[1]Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville “Improved Training of Wasserstein GANs” NeurIPS 2017
[2]Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic “End-to-End Speech-Driven Facial Animation with Temporal GANs” British Machine Vision Conference 2018.

Seen Speaker Speech Reconstruction (GRID)

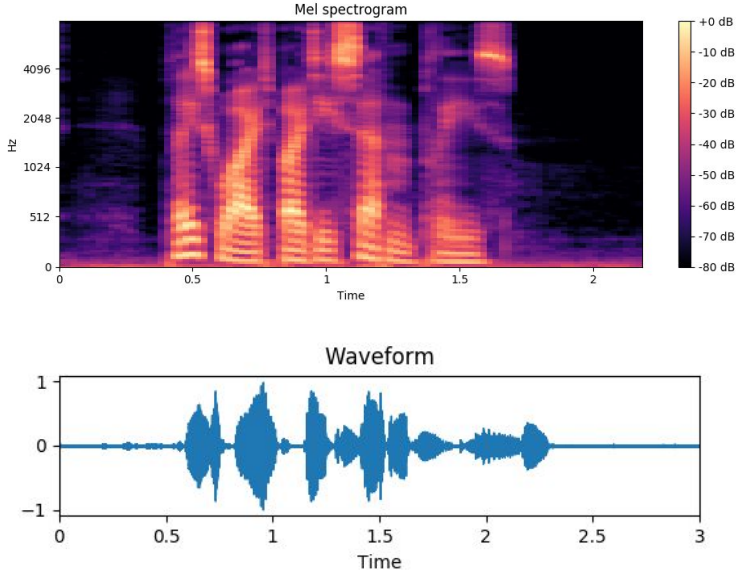
Imperial College London

ICASSP2020 Barcelona

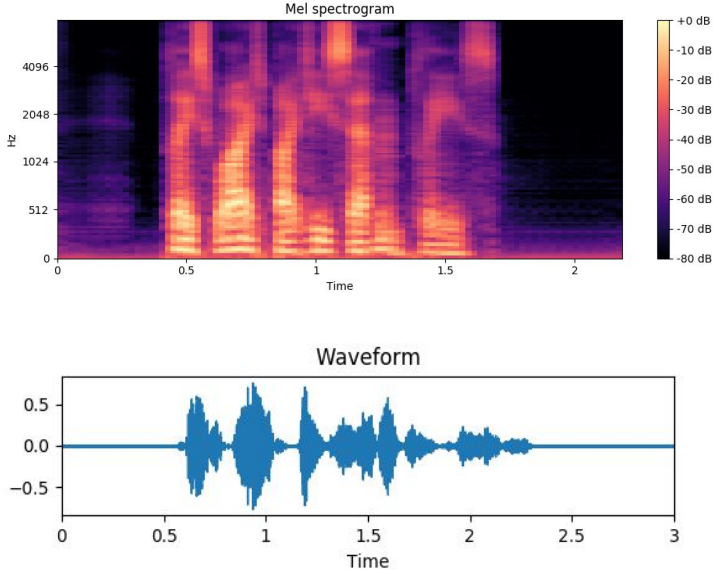
SAMSUNG

Spectrogram/Waveform Comparison (GRID, seen speakers)

Real



Synthesized



Imperial College London

ICASSP2020 Barcelona

SAMSUNG

Unseen Speaker Speech Reconstruction (GRID)

Rodrigo Mira et al.

Video-driven Speech Reconstruction using Generative Adversarial Networks
Show & Tell Demo

Imperial College London

ICASSP2020 Barcelona

SAMSUNG

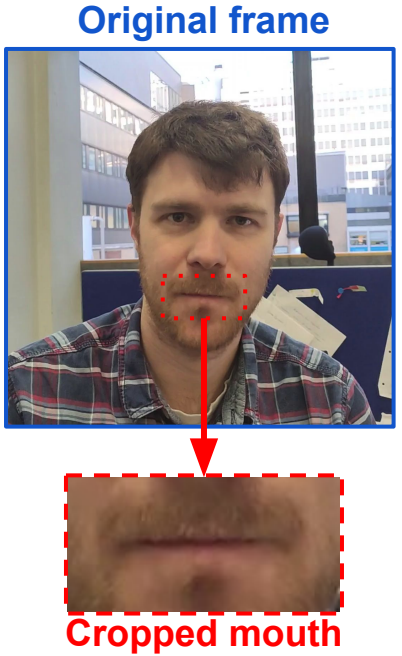
Unseen Speaker Speech Reconstruction in the Wild (LRW)

Rodrigo Mira et al.

Video-driven Speech Reconstruction using Generative Adversarial Networks
Show & Tell Demo

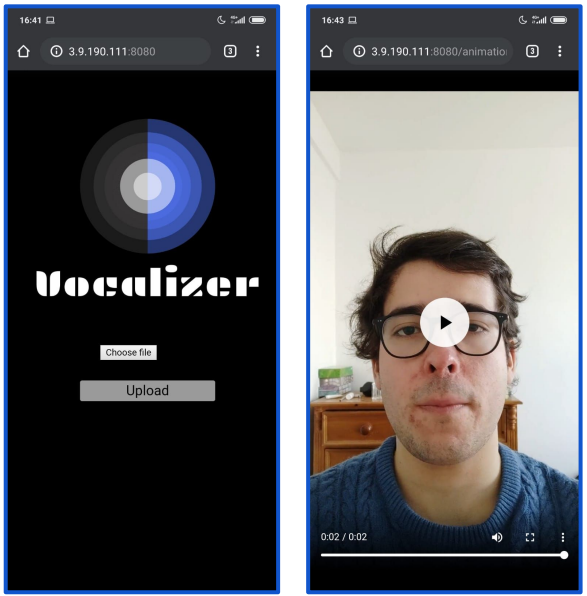
Demo (Step 1)

- Record video and convert it to **25 frames per second**.
- Perform face detection and alignment on each frame using ***Dlib***'s 68-landmark model.
- Align each frame to a reference **mean face shape**.
- Crop **mouth ROI** (Region of Interest) on each frame using a fixed 74x150 bounding box.
- Compile frames into cropped video.

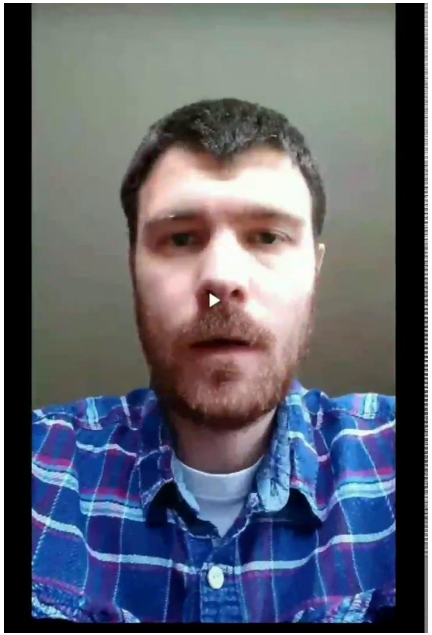


Demo (Step 2)

- Feed the video into our **model** (excluding the critic).
- Save a video featuring the old uncropped video and the **new reconstructed audio**, and display it.
- On an average CPU, the entire process takes around **40 seconds** for a 3 second video.
- Excluding pre-processing and post-processing, on a high end machine with an RTX 2080 TI, generating the waveform takes around 1 second.



Simulated Live Demo



Imperial College London

ICASSP2020 Barcelona

SAMSUNG

More Live Samples

Rodrigo Mira et al.

Video-driven Speech Reconstruction using Generative Adversarial Networks
Show & Tell Demo

Conclusion

- Thank you for watching our demo.
- We have shown that **intelligible speech reconstruction** is possible for live unseen speakers.
- In the future, we hope to find a way to capture the voice of new speakers efficiently, to create realistic voiced speech for live unseen speakers.
- The samples shown here can be found under **https://sites.google.com/view/speech-synthesis/home/extension**