Face2Speech

Introduction to Deep Learning, Carnegie Mellon University
Nikhil Rangarajan, Manini Amin, Shivani Sheth, Sanjana Mahajan
{nrangara, mpamin, sssheth, sanjanam}@andrew.cmu.edu

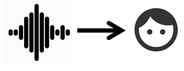
Introduction

Given an audio recording and an image of a face, can we output audio that "sounds" like the face?

Our task is to address the challenge of predicting what a person's voice would sound like if we are only given an image of their face, by utilizing the StarGAN architecture. Potential applications of a working model include speaker-identity modification and pronunciation modification.

Related Work

Speech to Face Models



Crossmodal Speech Conversion: VAE-based model

Face to Speech Models

Predicting faces from voices: GAN based-model

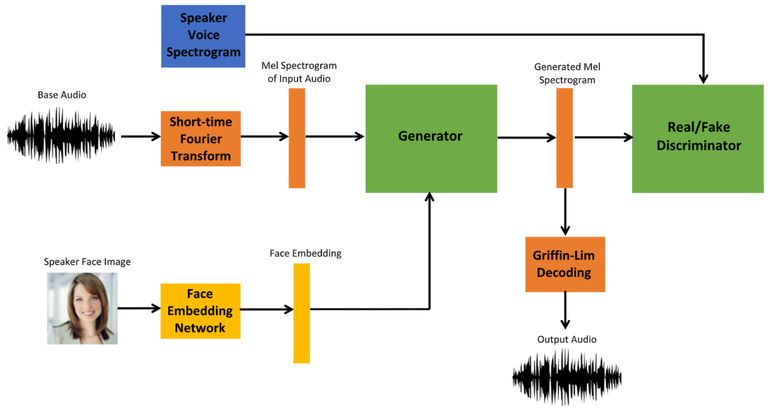
Speech2Face: voice encoder and face decoder

Datasets and Pretrained Models

VGGFace2	VoxCeleb2	Inception v1
9000 identities	5000 identities	Face recognition model Pretrained on VGGFace2 Face Embedding Network
3 million images	1 million utterances	

Utterances: Approx. 7-8s long
Face Images: 299px by 299px

Network Architecture



Overview of our StarGAN Architecture.
The Generator takes in the Mel Spectrogram of a user along with the face embedding of another user to output a newly generated Mel Spectrogram.

Methods

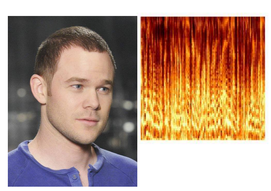
Data Processing

1. Convert VoxCeleb2 audio files to mel spectrograms. (Sampling Rate = 22.05 kHz, 256 mels, Hop Length = 512, FFT Points = 2048)
2. Convert VGGFace2 images to 512 dim face embeddings.

Model Pipeline

1. Utilize base mel spectrogram and target face embedding to input into StarGAN.
2. Generate target mel spectrogram as output from StarGAN.
3. Decode using Griffin-Lim algorithm to generate voice audio.

Base Face Image and Base Spectrogram



Target Face Image and Target Spectrogram

Results and Evaluation

Evaluation Criteria: Compared the outputted voice audio to target user's voice audio from dataset; conditioned on categories such as *gender, age and ethnicity*.

Checked for similarities/differences with respect to *accent, coarseness, pitch and loudness*. Used Mean Opinion Score for comparing the quality of the outputted voice sample with respect to the ground truth voice sample.

Results: Mean Opinion Score/Absolute Category Rating Metric: Achieved 2 (poor perceived quality) on a scale of 1-5.

Voice audio mapped from same category (age, gender, ethnicity) far more accurate and coherent with Mean Opinion Score between 3 and 4.

Further Development

Improve method for decoding spectrograms to voice audio (sharper, clearer sound).

Improve Robustness: Add more noise at training/test time to the voice samples.

Clean Face and Voice Dataset to remove uncorrelated aspects such as hair in images and multiple voices in recording.

Improve encoding features of spectrograms to capture maximal information.

References

[1] Choi, Yunjey, et al. "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation." 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, doi:10.1109/cvpr.2018.00916.
[2] Esler, Tim. "Timesler/Facenet-Pytorch." GitHub, 13 Nov. 2019, github.com/timesler/facenet-pytorch.
[3] Hirokazu Kameoka, Kou Tanaka, Aaron Valero Puche, Yasunori Ohishi, Takuhiro Kaneko, "Crossmodal Voice Conversion," arXiv:1904.04540 [cs.SD], Apr. 2019
[4] J. S. Chung*, A. Nagrani*, A. Zisserman. VoxCeleb2: Deep Speaker Recognition. INTERSPEECH, 2018.
[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman. VGGFace2: A dataset for recognising face across pose and age. International Conference on Automatic Face and Gesture Recognition, 2018.
[6] Wen Yandong, Singh Rita, Raj Biksha. "Reconstructing Faces from Voices." 25 May 2019, doi:1905.10604