DA323:MMDP-2 Course Assignment

FACE2SPEECH

Presented by Heet Patel (220150010)

THE PAPER

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Face2Speech: Towards Multi-Speaker Text-to-Speech Synthesis Using an Embedding Vector Predicted from a Face Image

Shunsuke Goto[1,2], *Kotaro Onishi*[1,3], *Yuki Saito*[2], *Kentaro Tachibana*[1], *and Koichiro Mori*[1]

[1]DeNA Co., Ltd., Tokyo, Japan
[2]The University of Tokyo, Japan
[3]The University of Electro-Communications, Tokyo, Japan

goto@gavo.t.u-tokyo.ac.jp, koichiro.mori@dena.com

Abstract

We are quite able to imagine voice characteristics of a speaker from his/her appearance, especially a face. In this paper, we propose Face2Speech, which generates speech with its characteristics predicted from a face image. This framework consists of three separately trained modules: a speech encoder, a multi-speaker text-to-speech (TTS), and a face encoder. The speech encoder outputs an embedding vector which is distinguishable from other speakers. The multi-speaker TTS synthesizes speech by using the embedding vector, and then the face encoder outputs the embedding vector of a speaker from the speaker's face image. Experimental results of matching and naturalness tests demonstrate that synthetic speech generated with the face-derived embedding vector is comparable to one with the speech-derived embedding vector.

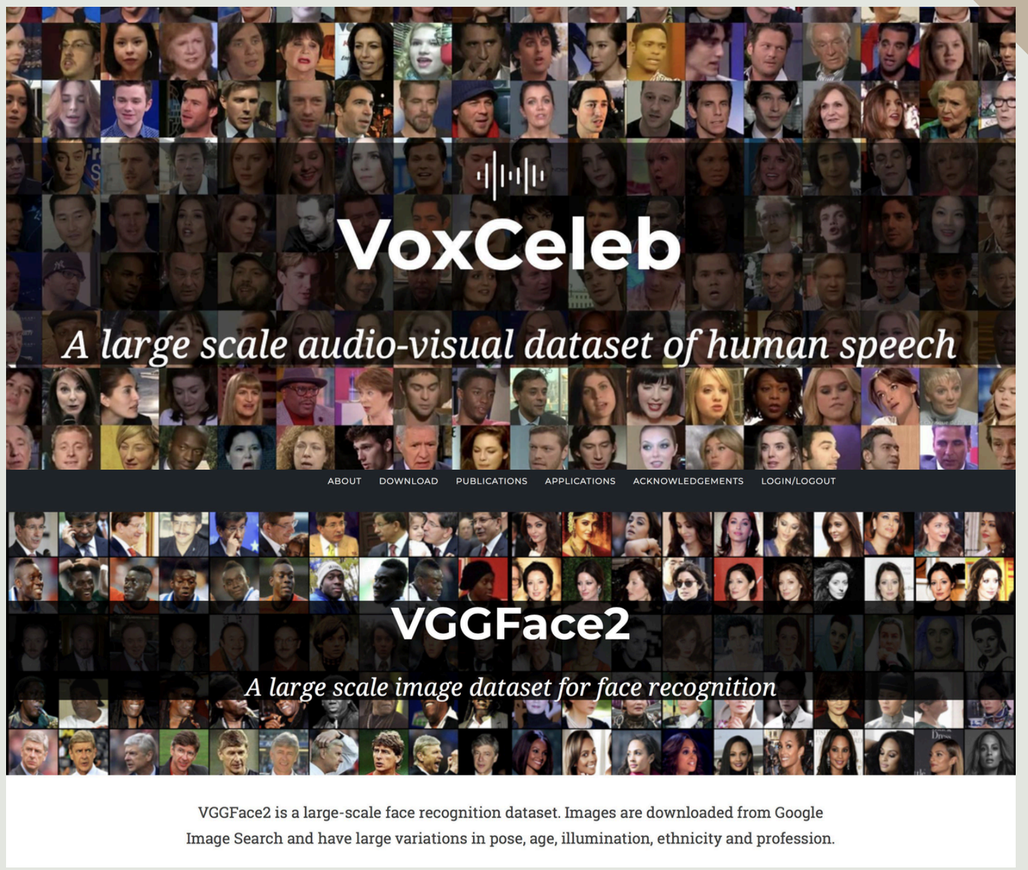
Index Terms: cross-modal face/voice generation, text-to-speech synthesis, multi-speaker modeling, speaker embedding

Hence, in this paper we aim to introduce facial features to a DNN-based multi-speaker TTS framework that can synthesize any arbitrary speakers' voices using the embedding vector of a speaker. The use of facial features would have many advantages in the practical application of TTS. For instance, we can intuitively identify the speaker for a synthesized voice based on visual information about the speaker, rather than using vocal features that are hard to visualize and taking time to listen to the speaker's voice samples. Besides, facial features can offer a natural means to control speaking style within a single speaker (e.g., emotional TTS [13]). Moreover, once the relationships between vocal and facial features are learned, we can introduce them to other media-related applications such as the voice searching systems [14].

In this paper, we propose a new DNN-based multi-speaker TTS framework named *Face2Speech*, which uses a face image to control the voice characteristics of the synthesized speech.

DATASET AND PREPROCESSING

- For Face Encoder:
  - VoxCeleb2: Over 6,000 celebrity videos with speech data. Used for high res face images.
  - VGGFace2: High-resolution face images of 5,993 speakers. Used to extract speech.
- VCTK & LibriTTS: Text-Speech data used for TTS training.
- Preprocessing:
  - Speech: 40-dimensional log Mel-spectrograms
  - Faces: 160x160 pixel images, normalized to [-1, 1]



ARCHITECTURE

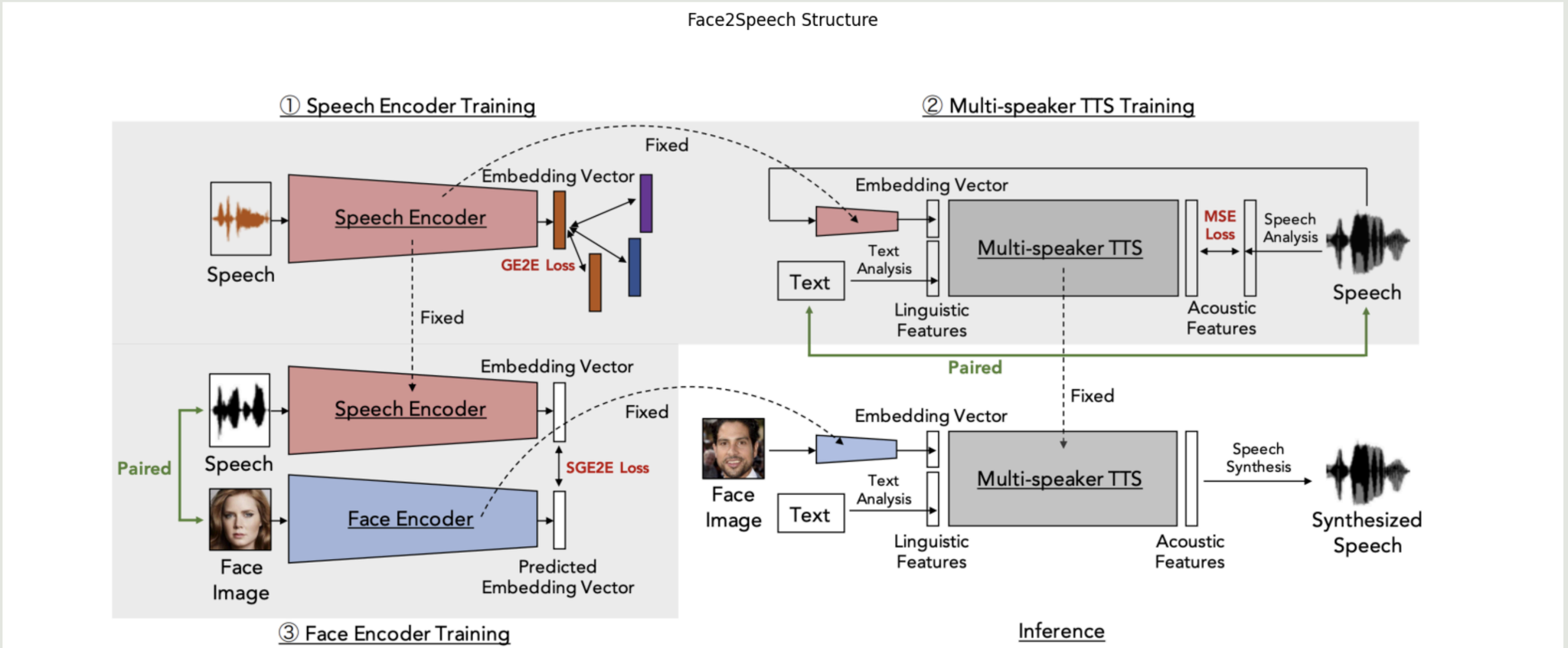
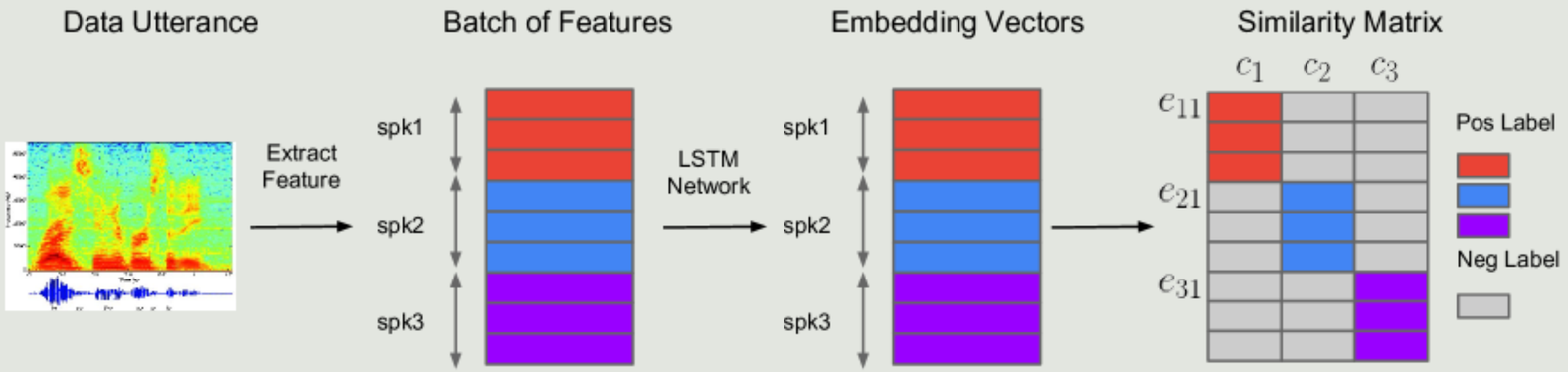


Figure 1: *Overview of Face2Speech. This framework consists of three separately trained modules: 1) speech encoder, 2) multi-speaker TTS, and 3) face encoder. After training, speech can be synthesized from a given text and a face image.*

SPEECH ENCODER

- Input: 40-dimensional log Mel-spectrogram
- Architecture: 3 LSTM layers with 768 hidden units
- Output: 256-dimensional embedding vector
- Loss Function: Generalized End-to-End Loss (GE2E)

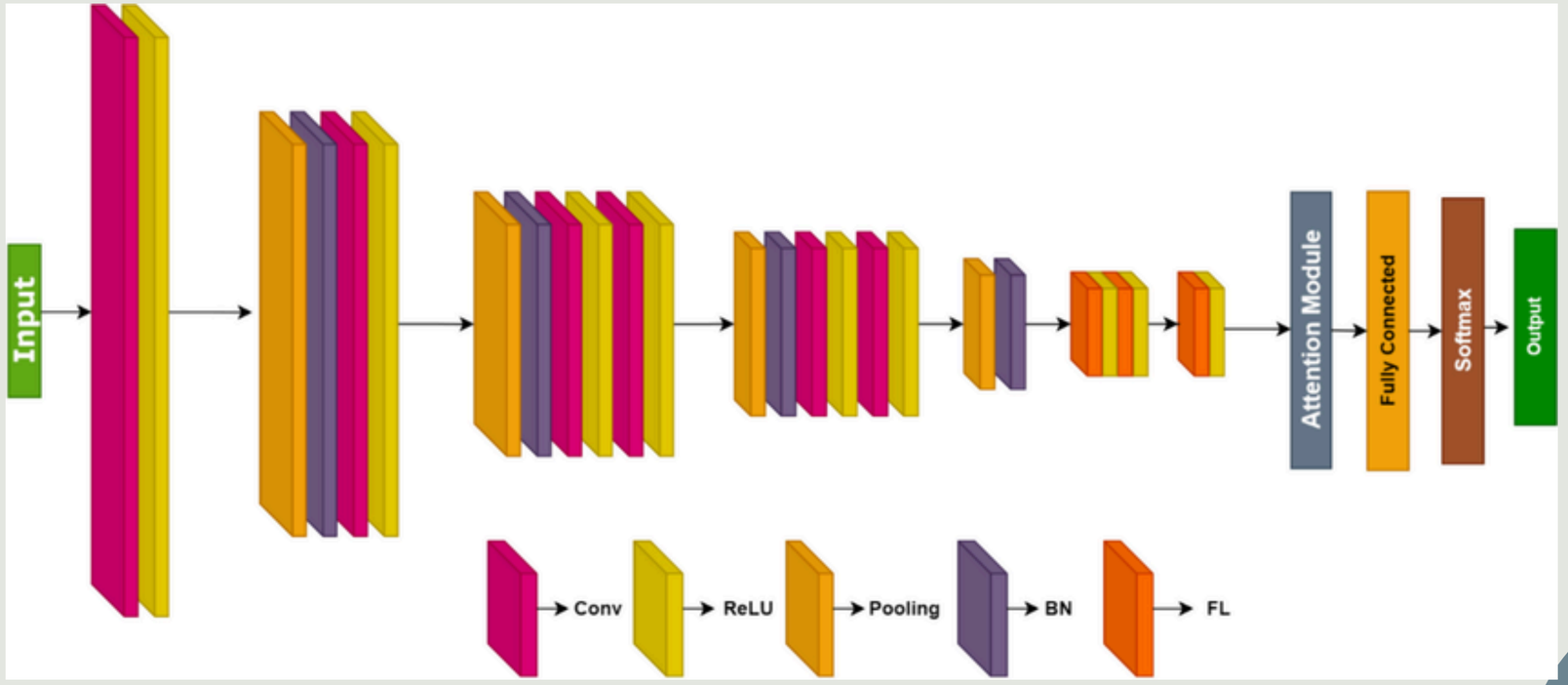


MULTI SPEAKER TTS

- Inputs: Text and embedding vectors from the Speech Encoder
- Models:
  - Duration Model: Predicts frame counts
  - Acoustic Model: Generates Mel-cepstral coefficients (MCEPs), log F0, and aperiodic measures
- Loss: Mean Squared Error (MSE)

FACE ENCODER

- Input: Face images (160x160 pixels)
- Architecture: VGG19
- Output: 256-dimensional embedding vector
- Loss Function: Supervised GE2E Loss (SGE2E)



EVALUATION

Evaluation Metrics

Table 1: *Matching scores on a four-point scale and preference scores of naturalness with 95% confidence intervals. Note that* ***the lower matching score is the better****, while the higher preference score is the better.*

System	Matching Score	Preference score
SYNTH-FACE	2.01 ± 0.07	0.548 ± 0.049
SYNTH-SPEECH	1.91 ± 0.06	0.452 ± 0.049

FUTURE SCOPE

- Instead of using GE2E and SGE2E losses, which capture joint embeddings of voice and facial images, we can train another network to predict the voice embeddings from facial embeddings. This way, we can retain original characteristics of the voice embeddings.
- Using better and more generalised multi speaker TTS models like TortoiseTTS or YourTTS.

Thank You