2-Q-49: 顔画像から予測される埋め込みベクトルを用いた複数話者音声合成

☆後藤 駿介[1,2]，大西 弘太郎[1,3]，齋藤 佑樹[1,2]，橘 健太郎[1]，森 紘一郎[1]
[1]ディー・エヌ・エー　[2]東大　[3]電通大

1. Introduction

背景

人の声と顔には何らかの関係性[1]

従来の複数話者音声合成

未知の話者の音声を埋め込みベクトルから推測

提案手法

未知の話者の音声を顔画像から推測

2. Method

SpeechEncoder

メルスペクトログラムから埋め込みベクトルを出力
Generalized End-to-End(GE2E) Lossで学習[2]
同じ話者の埋め込みベクトルは近づけ、異なる話者の埋め込みベクトルは遠ざける

N人の各M個の発話から構成されるバッチに対して
・c_k：話者kの埋め込みベクトルの重心
・e_{ji}：話者jのi番目の埋め込みベクトル
として類似度行列Sの要素を計算
類似度行列SからGE2E Lossを計算

$$S_{ji,k} = w \cdot \cos\left(\mathbf{e}_{ji}, \mathbf{c}_k\right) + b$$

$$L\left(\mathbf{S}\right) = -\log \frac{\exp\left(S_{ji,j}\right)}{\sum_{k=1}^{N} \exp\left(S_{ji,k}\right)}$$

FaceEncoder

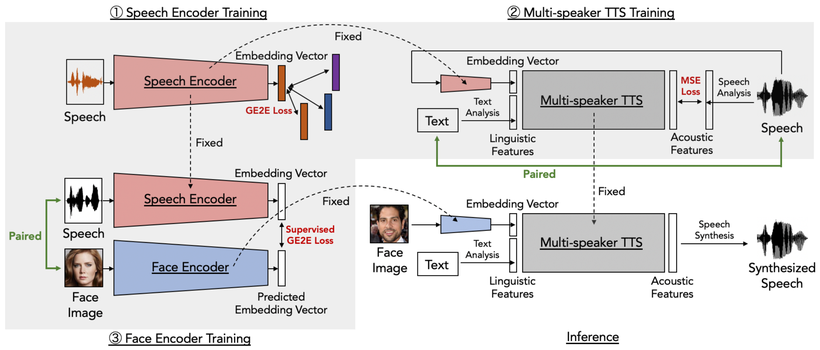
顔画像から顔画像の人物の埋め込みベクトルを出力

教師ありGE2E Lossで学習

N人の各M個の画像から構成されるバッチに対して
・$\tilde{c_k}$：SpeechEncoderの話者kの出力
・e_{ij}：話者jのi番目の埋め込みベクトル

$$\tilde{S}_{ji,k} = w \cdot \cos\left(\mathbf{e}_{ji}, \tilde{\mathbf{c}}_k\right) + b$$

$$L\left(\tilde{\mathbf{S}}\right) = -\log \frac{\exp\left(\tilde{S}_{ji,j}\right)}{\sum_{k=1}^{N} \exp\left(S_{ji,k}\right)}$$



として類似度行列Sの要素を計算
類似度行列SからSpeechEncoderと同様にGE2E Lossを計算

Multi-speaker TTS

継続長予測と音響特徴量推定のパイプライン型

3. Experiment

SYNTH-FACE(FaceEncoder + MS-TTS)と
従来法:SYNTH-SPEECH(SpeechEncoder + MS-TTS)を比較

実験条件

SpeechEncoder

入力	40次元メルスペクトログラム
出力	256次元
構造	768次元3層LSTM
バッチに含まれる話者数(N)	32
バッチに含まれる話者あたりの発話数(M)	4
データセット	VoxCeleb2
学習用話者数，テスト話者数	5593,118

FaceEncoder

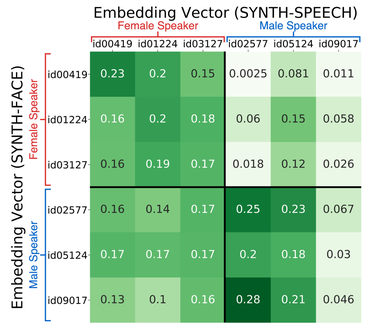
入力	160x160
出力	256次元
構造	VGG19
データセット	VGGFace2
学習用話者数，テスト話者数	5993,118

Multispeaker-TTS

入力	継続長:420次元言語特徴量+256次元埋め込みベクトル 音響: 425次元言語特徴量+256次元埋め込みベクトル
出力	継続長: 1次元継続長 音響:127次元音響特徴量
構造	512次元双方向LSTM
データセット	VCTK, LibriTTS
学習用話者数，テスト話者数	5993,118
Vocoder	WORLD

埋め込みベクトルの比較

SYNTH–SPEECHとSYNTH-FACEの埋め込みベクトルのコサイン類似度を計算



親和性評価

被験者30名は20サンプルの音声がその話者の顔画像に
1）ふさわしい、2）ややふさわしい、3）あまりふさわしくない、4）ふさわしくない
の４段階で評価

システム	スコア
SYNTH-FACE	2.01 ± 0.07
SYNTH-SPEECH	1.91 ± 0.06

自然性評価

被験者30名によるプリファレンスABテスト
被験者は10ペア（20サンプル）の音声を聞き、どちらが自然であるかを選択

システム	スコア
SYNTH-FACE	0.548 ± 0.049
SYNTH-SPEECH	0.452 ± 0.049

顔画像から従来手法と遜色ない音声を合成可能

[References]
[1] H. M. J. Smith, et al. Vol. 14, No. 1, 2016. [2] L. Wan, et al. InProc. ICASSP, pp. 4879-4883, Apr.2018.