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（ディー・エヌ・エー/東大），大西 弘太郎（ディー・エヌ・エー/電通大），
齋藤 佑樹（ディー・エヌ・エー/東大），橘 健太郎，森 紘一郎（ディー・エヌ・エー）*

1 はじめに

人はある人物の顔から声質を，あるいは声質から顔をある程度想像することが可能であると考えられる．つまり，人の声と顔には何らかの関係性があることは経験的に知っているものであり，その関係性について調査している研究もある [1]．本研究では，機械学習を用いて顔と音声の関係性を学習することを試み，特にテキスト音声合成 (text-to-speech; TTS) に焦点を置く．近年の Deep Neural Network (DNN) に基づく TTS の発展により，朗読型の TTS においては自然音声と同程度の品質の音声が生成できるようになってきた [2]．また，感情や話者性を扱えるような TTS の研究として，Variational Autoencoder (VAE) を用いて感情や話者性の多様な表現を可能にする埋め込みベクトルを学習する VAE-Loop [3] や，同様の考えを end-to-end 音声合成に導入した研究が行われている [4, 5]．また，d-vector [6] や x-vector [7] など，話者認識・話者認証で用いられる埋め込みベクトルを合成音声の話者性の制御に用いる研究も存在する [8]．

上述したような人間の知覚における知覚と聴覚の関係性を考慮すると，音声からだけでなく顔画像の情報からも話者性を表現することが可能であると期待できる．そこで本研究では，顔画像から推定される埋め込みベクトルを用いた DNN 複数話者音声合成モデル (Face2Speech) を提案する．従来の方法では目的話者の合成音声を生成する為にはその話者の音声サンプルを収集する必要があったが，顔画像から本人の声質に近い音声を生成できれば，音声収集の手間を省くことが可能である．また，顔から得られる感情表現を用いた音声合成の応用も期待できる．

提案手法では，（テキスト，音声，顔画像）の 3 つのモダリティのデータが必要になるが，これら全てを揃えた大規模データセットは我々の知る限り存在していない．その為，本研究では（テキスト，音声）と（音声，顔画像）の 2 種類のデータセットを用いて，図 1 に示される 3 つのモジュール，Speech Encoder，Multi-speaker TTS，Face Encoder を学習する．まず，Speech Encoder は話者認識に用いられる誤差関数を最小化することで，音声から他話者と区別可能な埋め込みベクトルを抽出する．次に，Multi-speaker TTS は（テキスト，音声）の組み合わせを用い，与えられたテキストと事前に学習済みの Speech Encoder から得られる埋め込みベクトルを用いて音声を生成する．最後に，Face Encoder は（音声，顔画像）の組み合わせを用い，顔画像からその人物の音声から抽出される埋め込みベクトルを出力するように学習される．推論時には，Face Encoder と Multi-speaer TTS を結合し，顔画像とテキストからそれに対応する音声を生成する．

実験的評価では，4 つのデータセットを用いて Face2Speech を学習する．（音声，顔画像）の組み合わせとして VoxCeleb2 [9] と VGGFace2 [10] を用い，（テキスト，音声）の組み合わせとして VCTK [11] と LibriTTS [12] を用いる．実験的評価の結果，顔由来の埋め込みベクトルから生成された合成音声が，その話者の音声由来の埋め込みベクトルから生成される合成音声と比べて顔画像との親和性，合成音声の自然性共に匹敵するという結果が得られる事を示す．

2 関連研究

近年，映像，画像，音，テキストなどの異なるモダリティの関係性を学習する研究が盛んに行われている．特に，本稿では聴覚と視覚に関わるクロスモーダル学習に焦点を置く．これまで，音声から顔の特徴を推定することで音声から顔の写像を学習する研究 [13] や，主観的な評価に基づいて顔から eigenvoice [14] への変換を行う研究 [15] が行われている．一方から一方のモダリティに変換するだけでなく，顔の特徴と音響的特徴の両方をある人物から別の人物へ変換する研究も行われている [16]．

また，テキスト，音声，顔の 3 つのモダリティを用いて複数話者音声合成を行う手法はあまり提案されていないが，顔の表情から推定される発話方法で音声を生成する研究 [17, 18] や，音声合成に加え，その音声に合わせて口唇が動くような顔映像生成を行う研究 [19] などが存在する．

3 Face2Speech モデル

提案手法は，図 1 に示すように Speech Encoder，Multi-Speaker TTS，Face Encoder の 3 つのモジュールから構成される．

3.1 Speech Encoder

Speech Encoder はメルスペクトログラムを入力とし，他話者と区別可能な埋め込みベクトルを出力する．話者の特徴を表現する埋め込みベクトルは話者認識において広く用いられてきた [20, 21]．提案手法では，Wan らの手法 [20] に従った枠組みを用いる．学習時の各バッチは N 人の話者の M 個の発話で構成され，j 番目の話者による i 番目の発話における埋め込みベクトルを L2 正規化したものを $\mathbf{e}_{ji} (1 \leq j \leq N, 1 \leq i \leq M)$ とする．話者 j の重心を $\mathbf{c}_j = \frac{1}{M} \sum_{m=1}^{M} \mathbf{e}_{jm}$ とすると，バッチ内の各発話と各話者の重心 $\mathbf{c}_k$ から求められる類似度行列 $\mathbf{S}$ の要素 $S_{ji,k}$ は式 (1) のように cos 類似度で定義される．

$$S_{ji,k} = w \cdot \cos(\mathbf{e}_{ji}, \mathbf{c}_k) + b \tag{1}$$

*Multi-speaker text-to-speech synthesis using an embedding vector based on a face image by GOTO, Shunsuke (DeNA/The University of Tokyo), ONISHI, Kotaro (DeNA/The University of Electro-Communications), SAITO, Yuki (DeNA/The University of Tokyo), TACHIBANA, Kentaro (DeNA), and MORI, Koichiro (DeNA)

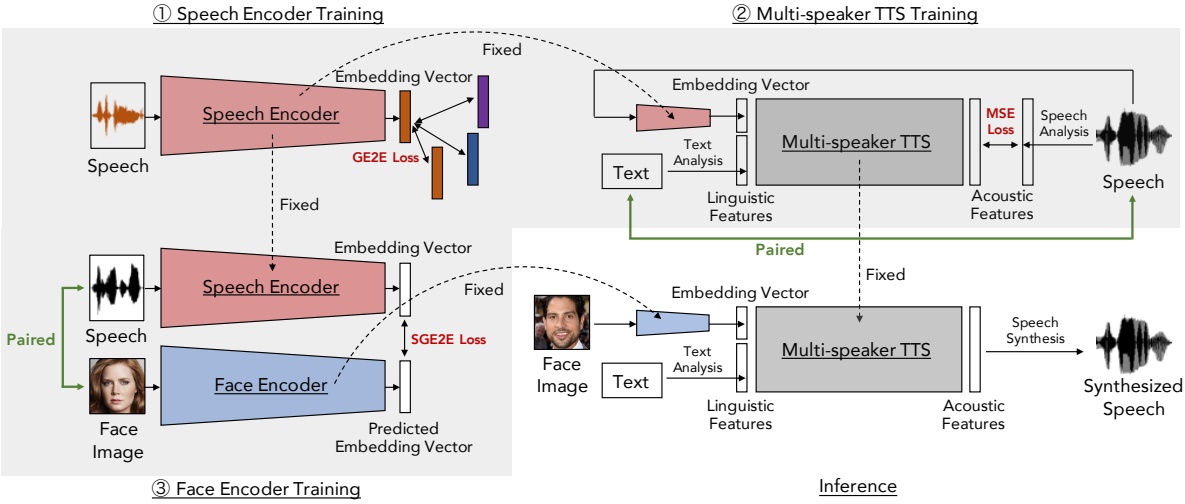

Fig. 1　提案手法の概略図

ここで，w と b は学習可能なスカラー値である．j 番目の話者による i 番目の発話における正規化埋め込みベクトル $\mathbf{e}_{ji}$ の誤差関数（generalized end-to-end loss; GE2E Loss）は，類似度行列 $\mathbf{S} = (S_{ji,k})_{(N\cdot M)\times N}$ を用いて式 (2) を用いて以下のように表される．

$$L(\mathbf{S}) = -\log \frac{\exp(S_{ji,j})}{\sum_{k=1}^{N} \exp(S_{ji,k})}. \tag{2}$$

この誤差関数は同じ話者の埋め込みベクトルの平均との cos 類似度を大きくし，異なる話者の埋め込みベクトルの平均との cos 類似度を小さくする効果がある．

3.2　Multi-speaker TTS

Multi-speaker TTS としては，継続長推定モデルと音響特徴量推定モデルから構成される統計的パラメトリック音声合成 [22] を用いる．継続長推定における入力は音素毎の言語特徴量と発話毎の埋め込みベクトルを連結した値であり，出力は音素継続長に当たるフレーム数である．一方，音響特徴量推定は，入力はフレーム毎の言語特徴量と発話毎の埋め込みベクトルを連結した値を用い，出力はフレーム毎の音響特徴量 (メルケプストラム，F_0，非周期性指標) である．継続長推定モデルと音響特徴量推定モデルの誤差関数はどちらも平均二乗誤差（mean squared error; MSE Loss）を使用する．

3.3　Face Encoder

Face Encoder の学習には顔画像と音声を用いる．入力は顔画像であり，出力はその話者の発話から得られる埋め込みベクトルの重心である．Face Encoder は Speech Encoder と同じく埋め込みベクトルを出力するため，Face Encoder の誤差関数は式 (2) と同様のものになる．ここで，M をバッチ内の各話者の発話数，$\tilde{M}$ を各話者の全発話の数とする．Speech Encoder では各バッチにおいて j 番目の話者の重心は $\mathbf{c}_j = \frac{1}{M}\sum_{m=1}^{M} \mathbf{e}_{jm}$ と計算されるが，Face Encoder では重心は Speech Encoder の出力を用いて $\tilde{\mathbf{c}}_j = \frac{1}{\tilde{M}}\sum_{m=1}^{\tilde{M}} \mathbf{e}_{jm}$ と計算され，この重心は全てのミニバッチにおいて一定である．その為，Face Encoder の類似度行列 $\tilde{S}_{ji,k}$ は，式 (3) のように定義される．

Fig. 2　VoxCeleb2 と VGGFace2 の解像度の違い

$$\tilde{S}_{ji,k} = w \cdot \cos(\mathbf{e}_{ji}, \tilde{\mathbf{c}}_k) + b. \tag{3}$$

類似度行列が与えられると，誤差関数（supervised generalized end-to-end loss; SGE2E Loss）は式 (2) によって計算される．

4　実験的評価

4.1　データセット

顔画像と音声のペアのデータセットとしては，VoxCeleb2 [9] と VGGFace2 [10] を用いた．VoxCeleb2 は 6000 人以上の有名人の発話を YouTube から抽出したデータセットであり，VGGFace2 は VoxCeleb2 と同じ人物が存在する大規模画像データセットである．顔と音声の対応を取ることができれば Face Encoder は学習できるが，図 2 に示すように動画から顔画像の部分を切り出すと解像度が低くなり顔情報の特徴抽出に向かないと考えられるため，音声は VoxCeleb2 から，顔画像は VGGFace2 から用意した．本研究では学習用に 5993 人，テスト用に 118 人の顔と音声のデータを用いた．

テキストと音声のペアのデータセットとしては，VCTK [11] と LibriTTS [12] を用いた．VCTK は 108 話者，90361 発話，LibriTTS は 805 話者，18744 発話を含む．両方のデータセット共にバックグラウンドノイズは少なく，VoxCeleb2 よりもクリーンな音声である．また，今回使用する音声は全てサンプリング周波数 16 kHz にダウンサンプリングした．

4.2 実験条件

4.2.1 Speech Encoder

Speech Encoder に用いる音声は窓長，ホップ長，FFT 長をそれぞれ，400 サンプル (25 ms)，160 サンプル (10 ms)，512 サンプル (64 ms) として分析した．窓関数はハン窓を使用した．Speech Encoder の入力は長さ 160 フレームの 40 次元 log-Mel スペクトログラムとし，出力の埋め込みベクトルは 256 次元のベクトルとした．DNN の隠れ層は 768 次元の 3 層の Long-Short Term Memory (LSTM) とし，出力層は 256 次元とした．隠れ層と出力層の活性化関数は，それぞれ tanh 関数，線形関数とした．誤差関数の計算の為に得られた出力は L2 正規化した．学習時の各バッチに含まれる話者数 N は 32，発話数 M は 4 とした．学習エポック数を 500，学習率を 10^{-5} とし，最適化手法は Adam [23] を用いた．

4.2.2 Multi-speaker TTS

音声分析・合成には WORLD ボコーダ [24][25] を用いた．継続長モデルは，音素ごとに 420 次元の言語特徴量と，発話ごとに Speech Encoder から得られる 256 次元の埋め込みベクトルを連結した 676 次元のベクトルを入力とし，1 次元の継続長を出力とした．音響モデルは，フレーム毎に 425 次元の言語特徴量と埋め込みベクトルを連結した 681 次元を入力とし，127 次元の音響特徴量（メルケプストラム 40 次元，対数 F_0 1 次元，非周期性指標 1 次元，これら全ての動的特徴量と，有声/無声フラグ 1 次元）を出力とした．ここで入力に与える埋め込みベクトルは発話内平均とした．継続長モデル，音響モデル共に入力は $[0.01, 0.99]$ に値を取るように正規化をし，出力は平均 0 分散 1 になるように正規化をした．どちらのモデルも隠れ層は 512 次元の双方向 LSTM3 層とし，隠れ層の活性化関数には tanh 関数を用いた．学習エポック数を 40，学習率を 10^{-4} とし，最適化手法は Adam を用いた．

4.2.3 Face Encoder

入力となる顔画像のサイズは 160×160 にスケールし，各ピクセル内の値は $[-1, 1]$ に値を取るように正規化をした．学習データ拡張の為にランダムに水平反転を行った．また，画像から顔の部分を切り出すために，[26] に基づいた事前学習モデルを用い，顔が検出されなかった画像については学習データから除外した．出力は 256 次元の埋め込みベクトルであり，顔画像と対応した話者の話者平均埋め込みベクトルである．DNN には画像認識に用いられる VGG19 [27] を用いた．学習エポック数を 124，最適化手法を Adam とし，学習率は 2.0×10^{-3} に設定した．

4.3 データセットの違いが Speech Encoder の学習に与える影響

Speech Encoder はテキストなしで音声から話者固有の特徴を抽出する．[28] では，Speech Encoder の学習に複数のデータセットを用いた場合でも，データセット間の環境の違いは合成音声と元音声との類似性に対して悪い影響を与えなかったことが示されている．しかし，提案手法ではテキスト音声合成に用いるデータセットと，顔から話者の埋め込みベクトルを推定する Face Encoder のデータセットが異なるため，データセットの混合には注意が必要である．本稿では，Speech Encoder を VoxCeleb2，VCTK，LibriTTS の全てを用いて学習した場合と VoxCeleb2 のみを用いた場合についての予備実験を行った．図 (3) に Speech Encoder での埋め込みの比較を示す．両手法共に話者の性別の違いが埋め込みベクトルに表れていることがわかるが，前者の場合クリーンである VCTK，LibriTTS とノイジーである VoxCeleb2 の違いも埋め込みベクトル上に表れている．一方で，後者は埋め込みベクトルには明確にデータセット間の違いは表れていない．つまり Speech Encoder は話者の特徴だけでなくデータセットの特徴も捉えてしまう為，異なるデータセットで Multi-speaker TTS と Face Encoder の学習を行う提案手法では，データセット間の違いを捉えない Speech Encoder が理想的である．よって，提案手法では VoxCeleb2 の音声のみで Speech Encoder の学習を行った．

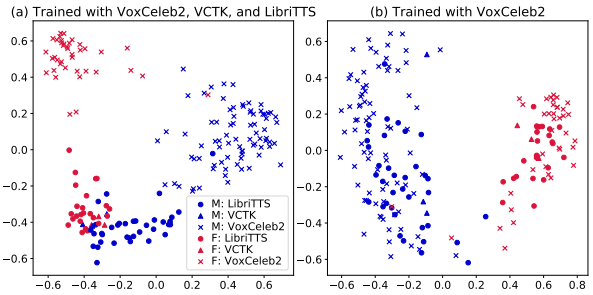

Fig. 3　Speech Encoder から出力される埋め込みベクトルの主成分分析結果．(a) VoxCeleb2，VCTK，LibriTTS で学習，(b) VoxCeleb2 のみで学習．

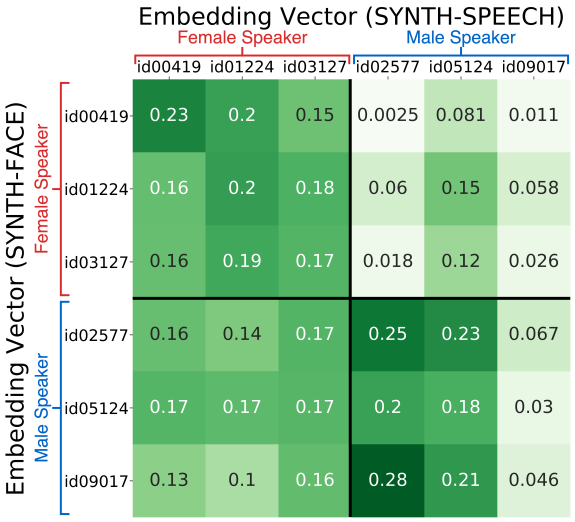

Fig. 4　2 つの埋め込みベクトル（**SYNTH-FACE** と **SYNTH-SPEECH**）のコサイン類似度を示すヒートマップ

4.4 客観評価

本実験では 2 つの方法（**SYNTH-FACE** と **SYNTH-SPEECH**）で音声を生成する．**SYNTH-SPEECH** では，ある話者の発話を Speech Encoder に入力することで埋め込みベクトルを得るが、

System	Matching Score
SYNTH-FACE	2.01 ± 0.07
SYNTH-SPEECH	1.91 ± 0.06

Table 1　親和性に関する 4 段階評価と 95%信頼区間

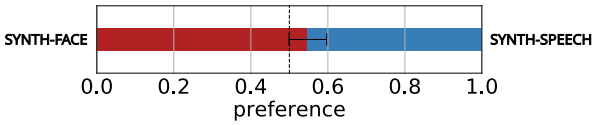

Fig. 5　自然性に関するプリファレンススコア．エラーバーは 95%信頼区間を示す．

SYNTH-FACE は，顔画像から対応する話者の平均埋め込みベクトル，つまり **SYNTH-SPEECH** の埋め込みベクトルを Face Encoder から予測する．Face Encoder は，Speech Encoder から出力される埋め込みベクトルとの SGE2E Loss を最小化するように学習される為，**SYNTH-SPEECH** は本実験における上限であると言える．本節では 2 つの方法で得られるテスト話者の埋め込みベクトルのコサイン類似度を計算し，図 4 にそのヒートマップを示す．

各話者の **SYNTH-FACE** 埋め込みベクトル (図 4 の各行) は，同じ人物，あるいは同性の話者の **SYNTH-SPEECH** 埋め込みベクトル (図 4 の各列) と最も高い類似度を示している．この結果から，特に性別に関して言えば 2 つの埋め込みベクトルが比較的高い相関があると考えられる．すなわち，Face Encoder は，顔画像から精緻に本人の声質を推定することはできないが，大まかな特徴を捉えることができると考えられる．

4.5　主観評価

4.5.1　親和性評価

親和性評価では，顔画像から推定される埋め込みベクトルを用いて生成した音声とがその人物の顔画像とどれだけ適合しているかを評価した．音声サンプルは以下のリンクから確認できる[1]．4.4 節と同様に **SYNTH-FACE** と **SYNTH-SPEECH** の 2 種類の方法で音声を生成した．それぞれの音声が顔画像に対してふさわしいかどうかの主観実験を音声ごとに別に実験を行った．被験者は計 30 名とし，各被験者は 20 サンプルの音声を聞き，それぞれ 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 段階から評価した．

親和性評価の結果を表 1 に示す．**SYNTH-FACE** は **SYNTH-SPEECH** に対しわずかに親和性が劣るものの，著しい違いはないことが確認できる．人間が判断できる限りでは，当該話者の音声から得られる埋め込みベクトルを用いて生成した音声と同程度に，顔画像から推定される埋め込みベクトルを用いて生成した音声はその顔画像と合っているということを示している．

4.5.2　自然性評価

親和性評価に加えて，顔画像から推定される埋め込みベクトルを用いて生成した音声がどれだけ自然であるかを測る自然性評価も行った．評価方法は，30 名の被験者によるプリファレンス AB テストとした．各被験者は 10 ペア (計 20 サンプル) の音声をランダムな順序で聞き，どちらの音声が自然であるかを選択した．

自然性評価の評価結果を図 5 に示す．**SYNTH-FACE** は **SYNTH-SPEECH** と同等のスコアであることが示された．すなわち，未知話者の音声の生成の際に顔画像 1 枚から音声を生成しても，その話者の音声を用いた場合に比べて遜色のない品質の音声が生成できる事が示された．

[1] https://dena.github.io/Face2Speech/

5　おわりに

本稿では，顔画像から推定される声質の音声を合成するテキスト音声合成を提案し，実験的評価によりその有効性を示した．今後は話者の変動に対する提案手法の頑健性を調査する．

参考文献

[1] H. M. J. Smith, et al. Vol. 14, No. 1, 2016.

[2] J. Shen, et al. In *Proc. ICASSP*, pp. 4779–4783, Apr. 2018.

[3] K. Akuzawa, et al. In *Proc. Interspeech*, pp. 3067–3071, Sep. 2018.

[4] Y. Wang, et al. *arXiv*, Vol. abs/1803.09017, , 2018.

[5] W.-N. Hsu, et al. In *Proc. ICLR*, May 2019.

[6] E. Variani, et al. In *Proc. ICASSP*, pp. 4080–4084, May 2014.

[7] D. Snyder, et al. In *Proc. ICASSP*, pp. 5329–5333, Apr. 2018.

[8] F. Fang, et al. In *Proc. SSW*, pp. 155–160, Sep. 2019.

[9] J.-S. Chung, et al. In *Proc. Interspeech*, pp. 1086–1090, Sep. 2018.

[10] Q. Cao, et al. In *Proc. FG*, pp. 67–74, 2018.

[11] C. Veaux, et al. 2017.

[12] H. Zen, et al. In *Proc. Interspeech*, pp. 1526–1530, Sep. 2019.

[13] T.-H. Oh, et al. In *Proc. CVPR*, pp. 7539–7548, June 2019.

[14] R. Kuhn, et al. *IEEE Trans. on Speech and Audio Processing*, Vol. 8, No. 6, pp. 695–707, Nov. 2000.

[15] Y. Ohsugi, et al. In *Proc. Interspeech*, pp. 1001–1005, Sep. 2018.

[16] F. Fang, et al. In *Proc. ICASSP*, pp. 6795–6799, May 2019.

[17] É. Székely, et al. *Speech Communication*, Vol. 57, pp. 63–75, Feb. 2014.

[18] É. Székely, et al. In *Proc. SLAPT*, pp. 5–8, June 2012.

[19] J. Schroeter, et al. In *Proc. ICME*, pp. 571–574, July 2000.

[20] L. Wan, et al. In *Proc. ICASSP*, pp. 4879–4883, Apr. 2018.

[21] L. Chao, et al. *arXiv*, Vol. abs/1705.02304, , 2017.

[22] H. Zen, et al. *Speech Communication*, Vol. 51, No. 11, pp. 1039–1064, Nov. 2009.

[23] D.-P. Kingma and J. Ba. In *Proc. ICLR*, May 2015.

[24] M. Morise, et al. *IEICE Trans. on Information and Systems*, Vol. 99, No. 7, pp. 1877–1884, July 2016.

[25] M. Morise. *Speech Communication*, Vol. 84, pp. 57–65, Nov. 2016.

[26] K. Zhang, et al. *IEEE Signal Processing Letters*, Vol. 23, No. 10, pp. 1499–1503, Oct. 2016.

[27] K. Simonyan and A. Zisserman. *arXiv*, Vol. abs/1409.1556, , 2014.

[28] Y. Jia, et al. In *Proc. NIPS*, pp. 4480–4490, Dec. 2018.