Vol.2017-SLP-118 No.2
2017/10/13

DNN音声合成のための
話者類似度に基づく教師なし話者適応

高木 信二[1,a)]　西村 祥一[2,b)]　山岸 順一[1,c)]

概要：本論文では，DNN に基づく音声合成において，話者適応にテキストを必要としない，教師なし話者適応について検討する．我々はこれまで，DNN 音声合成において，言語特徴量に加え話者・ジェンダー・年齢コード (入力コードと呼ぶ) を利用した音声合成のための複数話者モデリング，話者適応を提案してきた．本研究では，音声データのみから計算される，学習話者に対する話者類似度を入力コードとして利用する．ここで，話者類似度とは，話者認識において広く用いられているモデル (GMM-UBM や i-vector/PLDA) を利用し計算された，個々の学習話者に対する事後確率を連結したベクトルにより表現されると仮定する．提案教師なし話者適応手法は，目標話者の音声から話者認識モデルにより計算された話者類似度ベクトルを，DNN 音声合成システムの入力コードとして用いることで，実現される．話者認識モデルの構築においては，音声合成に適した話者類似度ベクトルの取得のため，利用する音響特徴量の検討を行った．10 代後半から 80 代までの話者がバランス良く含まれた 135 名からなる高品質巨大コーパスを用い，評価実験を行った．主観評価の結果より，提案法は合成音声の品質を下げることなく，高精度な話者適応が可能であることを確認できた．

キーワード：音声合成，DNN，教師なし話者適応，話者認識

1. はじめに

高い柔軟性を持つ DNN に基づく音声合成システムの研究は広く行われており，DNN に基づく音響モデルにおいて，複数話者モデリングや様々な話者適応手法が提案されている．例えば，話者情報を DNN の入力とすることで複数話者の音声の合成を行う手法 [1], [2] が提案されており，また，話者適応には話者情報の推定 [3]，i-vector[4]，GMM を用いた出力の変換 [4]，Hidden UnitContribution に基づく適応 (LHUC)[5] 等が利用されている．その他，複数話者データを用いることで，隠れ層は全話者で共有されるが話者特有の出力層を持つ音響モデルの構築が行なわれている [6]．この手法では回帰層のみを推定することによる話者適応が検討されている．

我々はこれまで，言語特徴量に加え追加の特徴量として話者・ジェンダー・年齢コード (入力コードと呼ぶ) を利用した DNN 音声合成に基づく，高精度な複数話者モデリング，及び，話者適応を提案した [7]．この手法の話者適応は，音声とその対となるテキストを用い，バックプロパゲーションに基づき目標話者に対する入力コードを逆推定することで行われていた．これに対し，本研究では，入力コードに基づく DNN 音声合成において，テキストを必要としない教師なし話者適応を検討する[*1]．これまで，DNN 音声合成において教師なし話者適応のための d-vector に基づく手法 [9] が提案されているが，その他に教師なし話者適応の研究は報告されておらず，分に研究がなされているとは言い難い．

本研究では，入力コードに基づく DNN 音声合成システムにおいて，音声のみから計算可能な入力コードを話者コードとして利用し，話者適応時にも未知話者の話者コードを音声のみから推定することにより，教師なし話者適応を実現する．この目的のため，話者認識において広く用いられているモデル (GMM-UBM，i-vector/PLDA) を利用し計算された，学習話者に対する事後確率を入力コードと

[1] 国立情報学研究所
National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
[2] 株式会社オルツ
alt Inc., The Canal Gate Akihabara 8F, 3-1-2 Higashikanda, Chiyoda-ku, Tokyo, Japan
a) takaki@nii.ac.jp
b) yoshikazu.nishimura@alt.ai
c) jyamagis@nii.ac.jp

[*1] 厳密には「教師なし話者適応」という用語は正しい使い方ではないが，HMM 音声合成でも同様の適応を教師なし話者適応と呼んでいた [8] ことから、DNN 音声合成でも教師なし話者適応と呼ぶこととする

して利用することを提案する．概念的には，話者認識モデルの各学習話者に対する事後確率が，学習話者の違いの類似度を表現していると仮定し，その情報を One-hot ベクトルの代わりとして利用することで，DNN 音声合成のための複数話者モデルの構築を行う．これにより，学習された DNN 音声合成モデルは，学習話者の違いの類似度を反映し，さらに，入力される話者類似度ベクトルが変化した場合には，出力音声も変化すると期待できる．言い換えれば，提案 DNN 音声合成モデルに対して，未知の目標話者の音声から計算された話者類似度を入力コードとして利用すれば，出力音声も変化し，教師なし話者適応が実現できると期待される．本手法では，話者認識モデルにより計算された個々の学習話者に対する事後確率を，DNN 音声合成システムの入力コードとしており，One-hot ベクトルの拡張ともみなせる．i-vector などを直接入力として利用した従来の話者適応手法と比べ，汎用的な枠組みであることもメリットの一つである．

2. 入力コードを用いた DNN 複数話者音声合成モデル

ここでは，文献 [7] の手法を簡潔にレビューする．DNN に基づく複数話者音声合成システムは，複数話者コーパスを用い特定話者モデル構築と同様の手順で構築を行う．しかし，各話者の音響特性を保持することに加え話者適応を可能とするため，言語特徴量に加え追加の特徴量 (入力コードと呼ぶ) を DNN 音声合成システムの入力として用いる．これら入力コードを用いることで，話者，ジェンダー，年齢等を区別した DNN に基づく音響モデルの学習，また，音声合成を行うことが期待できる．

2.1 入力コード

入力コードとして話者・ジェンダー・年齢コードを用いる．例えば，話者コードに One-hot ベクトル，ジェンダーコードにバイナリ値 (0 が女性，1 が男性)，年齢コードに年齢そのものを表現する 1 次元の数値が利用される．

2.2 入力コードを用いた DNN 音響モデルにおける話者適応

少量の音声データと対となるテキストを用い，目標話者の音声を合成するのに適した話者コードの推定を行う．文献 [10] と同様に，目標話者の適応データに対して予測誤差が少なくなるように，バックプロパゲーションに基づき話者コードを逆推定する．適応時に平均声モデルが初期値として有効であると知られていることから [11]，本手法では DNN に基づく複数話者モデルにおいて，話者コードの初期値に平均値を利用し，適応を行う．本研究では，この教師あり話者適応手法を提案法との比較に用いた．

3. 話者類似度に基づく教師なし話者適応

提案法は，セクション 2 で記した入力コードに基づく DNN 音声合成において，テキストを必要としない教師なし話者適応を目的とする．提案法における，音声合成のための複数話者モデルの構築，話者適応の手順は以下の通りである．

(1) 個々の学習話者に対する話者認識モデルを構築する．利用する音声データは DNN に基づく複数話者音声合成システム構築に用いる音声データと同様である．本研究では，話者認識モデルに GMM-UBM[12]，または，i-vector/PLDA[13] を利用した．
(2) Step.1 で構築した話者認識モデルと学習話者の音声データを用い，個々の学習話者に対する事後確率を計算する．得られた事後確率を連結し，話者類似度ベクトルとする*2．本研究では，学習データには学習話者が 112 人含まれるため，112 次元のベクトルにより話者類似度が表現される．
(3) Step.2 で求めた話者類似度ベクトルを話者コードとし，ジェンダー・年齢コードを加えた入力コードを用い，DNN に基づく複数話者音声合成システムの構築を行う．
(4) Step.1 で構築した話者認識モデルを用い，目標話者の適応データから話者コードを表現する話者類似度ベクトルを推定する*3．

これにより，話者類似度ベクトルに基づく DNN に基づく複数話者音声合成システムの構築，目標話者に対する話者類似度ベクトルの推定が行われ，教師なし話者適応が可能となる．以下，本研究で用いた話者認識モデル (GMM-UBM, i-vector/PLDA) について記す．

3.1 GMM-UBM

GMM-UBM では，不特定話者データを用い Universal Background Model (UBM) を学習し，話者モデルの構築には当該話者の音声データを用いた適応を行う．通常，モデルには GMM が利用される．

3.2 i-vector/PLDA

i-vector/PLDA は，i-vector と呼ばれる各話者を表現する特徴量を，Probabilistic Linear Discriminant Analysis (PLDA) によりモデル化する手法である．i-vector は GMM スーパーベクトル空間において UBM の平均からの次元圧縮された差とされ，発話 u における GMM スーパベクトル

*2 ここで，話者認識モデルの構築と学習話者に対する事後確率の計算には同じ音声データが用いられるが，話者認識が目的ではなく，話者類似度の計算が目的であることに注意する．
*3 提案法は，年齢コード，ジェンダーコードの推定にも利用可能であるが，本論文では [7] と同様に話者コード以外のコード推定については議論は行っていない．

$\boldsymbol{M}_u$ の生成過程を，

$$\boldsymbol{M}_u = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}_u + \boldsymbol{\epsilon} \tag{1}$$

とすると，i-vector は $\boldsymbol{w}_u$ と表現される．ここで，$\boldsymbol{m}$ は GMM スーパベクトル，$\boldsymbol{T}$ は基底行列であり，$\boldsymbol{w}$ は標準ガウス分布，，$\boldsymbol{\epsilon}$ は残差成分でありガウス分布 $\mathcal{N}(\boldsymbol{w} \mid \boldsymbol{0}, \boldsymbol{\Sigma})$ に従う．

i-vector を用いた話者間変動や話者内変動をモデル化する手法として，PLDA が提案されている [13]．本研究では，Gaussian PLDA と呼ばれる手法を用いる．本手法では，i-vector を観測とし，以下のようにモデル化を行う．

$$\boldsymbol{w}_u = \bar{\boldsymbol{w}} + \boldsymbol{\Phi}\boldsymbol{\beta} + \boldsymbol{\Gamma}\boldsymbol{\alpha}_u + \boldsymbol{\epsilon}_u \tag{2}$$

ここで，$\bar{\boldsymbol{w}}$ は i-vector 空間におけるオフセット，$\boldsymbol{\Phi}$，$\boldsymbol{\Gamma}$ はそれぞれ話者部分空間，チャネル部分空間における基底行列，$\boldsymbol{\beta}$，$\boldsymbol{\alpha}_u$ は話者因子，チャネル因子を表現し, 標準ガウス分布に従う．$\boldsymbol{\epsilon}_u$ は残差成分でありガウス分布 $\mathcal{N}(\boldsymbol{w} \mid \boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$ に従う．本研究では，式 (2) における第 3 項は利用しなかった．

3.3 提案話者コード

話者識別には，GMM-UBM では UBM を用いた対数尤度比が，i-vector/PLDA では二つの発話が同じ話者によるものか否かに関するスコア [13] が用いられることが多いが，本研究では，話者認識モデルは個々の学習話者に対する事後確率を求めるために用いられる．また，得られた個々の学習話者に対する事後確率を連結したベクトルを話者コードとし，DNN 音声合成の入力として用いる．これまで，i-vector を DNN 音声合成システムの入力として直接利用する手法 [4] が報告されているが，提案法では i-vector は PLDA モデルの構築，事後確率の計算に使用され，DNN 音声合成システムの入力としては用いない．

また，文献 [9] では教師なし話者適応が検討されており，話者認識 DNN の中間層に PCA を適用，事後確率で重み付けし，音声合成システムに利用している．そのため，話者認識のための DNN やその音響特徴量に音声合成システムが直接依存している．提案法は個々の学習話者に対する事後確率のベクトルのみを音声合成に利用するため，音声合成システムが話者認識用モデルパラメータや音響特徴量に直接的に依存しておらず，シンプルかつ汎用性の高い方法と言える．

4. 実験

4.1 実験条件

実験には 10 代後半から 80 代までの男性 65 名，女性 70 名からなる高品質日本語コーパスを用いた．このコーパスから男性 56 名，女性 56 名の音声データを複数話者モデルの学習に用い，残りの男性 9 名，女性 14 名の音声データを適応実験に用いた．学習セットに含まれる話者は年齢，ジェンダーがバランス良く含まれるように設定し，各年齢帯とジェンダーそれぞれで 8 名である．発話数も話者間で大きな偏りがないように設定し各話者それぞれ約 100 文を用い，合計で 11,154 発話である．

適応実験では，学習データに含まれない 23 話者から 10，50，100 文のいずれかを適応データとして用いた．適応実験ではデータ量の関係から学習データ選択のようにバランスの良い話者選択は行えなかったが，可能な限り年齢，ジェンダーをバランス良く含むように選択を行っている．

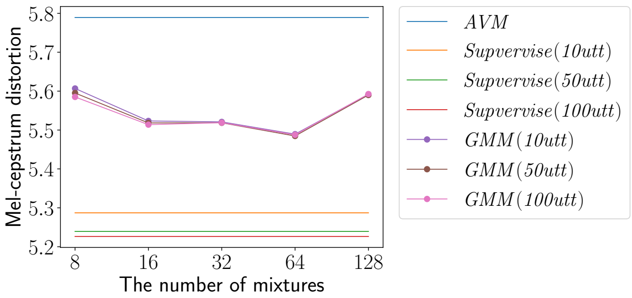
複数話者音声合成システムを用いた実験，適応実験の評価ではテスト文として各話者異なる 10 文を用いた．

話者認識モデルの学習には，SIDEKIT[14] を用いた．特徴量には 0 次を含む 19 次 MFCC とその Δ，Δ^2 も用いた．本研究では音声合成に適した話者類似度ベクトルを求めるため，WORLD[15] を用いて抽出した 0 次を含む 19 次メルケプストラム (MGC) とその Δ，Δ^2，また，F0 に関する特徴量の利用も検討した．F0 に関する特徴量としては，F0 をそのまま用いるとスペクトル特徴量と次元数が大きく異なることから，当該フレームと前後 32 フレームの無声部を補完した log F0 に対して DCT を行い 20 次元にした特徴量とその Δ，Δ^2(DCTF0) を利用した．これら，特徴量を用い GMM-UBM の学習，i-vector 抽出のための UBM の学習を行なった．i-vector 抽出に用いた GMM のミクスチャ数は 64，i-vector の次元数は 400，G-PLDA の話者部分空間に対する基底数は 20 である．

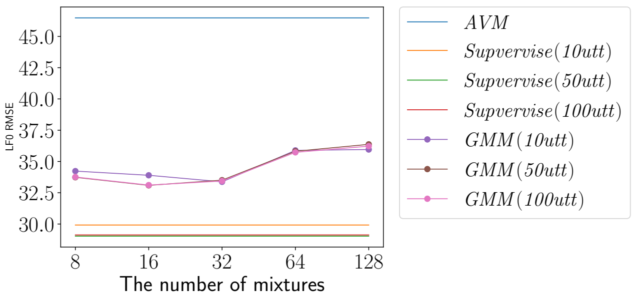
DNN 音声合成システムの構築には，0 次を含む 59 次メルケプストラム係数，無声部を補完した log F0，無声/有声パラメータ，25 次非周期成分を用いた．これら音響特徴量は F0 適応窓を用いて WORLD により各フレーム毎に抽出を行った．フレームシフトは 5ms である．また，無声/有声パラメータ以外のパラメータについてはその Δ，Δ^2 も用いた．DNN に基づく音響モデル構築に利用した音響特徴量は合計 259 次元となる．

言語特徴量は 389 次元であり，DNN に基づく音響モデルの入力として用いられる．言語特徴量に含まれる音素継続長は HMM により学習データ，テストデータともに Forced-alignment により得た．つまり，音声合成時にテストデータから得られた音素継続長を利用している．また，実験では言語特徴量に加え話者，ジェンダー，年齢を表現する入力コードを DNN に基づく音響モデルの入力として利用する．

音響モデルは隠れ層数 5 で各隠れ層が 1024 ユニットからなるフィードフォワード DNN を用いた．シグモイド関数を隠れ層，出力層の活性化関数として用いた．音響モデルのパラメータはランダム値により初期化を行い，学習率 (0.05)，学習回数 (10 epochs)，ミニバッチサイズ (256 フレーム) を固定し確率的勾配降下法により学習を行った．

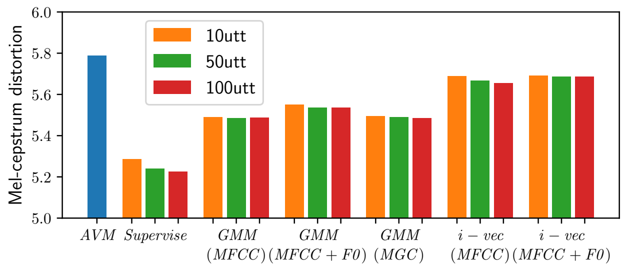


(a) メルケプストラム歪み

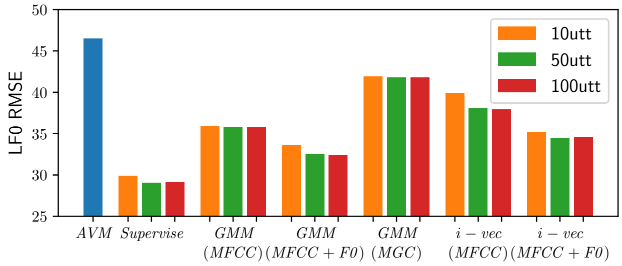


(b) LF0 RMSE

図 1　平均声 (*AVM*)，教師あり話者適応 (*Supervise*)，GMM-UBM を利用した教師なし話者適応 (*GMM*) の客観評価結果を示す．GMM のミクスチャ数が 8，16，32，64，128，適応文章数は 10，50，100 の結果である．



(a) メルケプストラム歪み



(b) LF0 RMSE

図 2　*AVM*，教師あり話者適応 (*Supervise*)，GMM-UBM，または，i-vector/PLDA 利用した教師なし話者適応 (*GMM*, *i-vec*) の結果を示す．システム名のカッコ内は話者認識モデルの学習に利用した特徴量を示す．

提案法との比較のため，話者コードとして One-hot ベクトルを用いた DNN 複数話者音声合成システムによる，教師あり話者適応システムを構築した [7]．また，全ての手法において，目標話者に対するジェンダー，年齢は既知とし，適応は行っていない．

4.2　実験結果 (話者適応)

GMM-UBM，i-vector/PLDA に基づく話者認識モデルを用いた，提案法による教師なし話者適応 (*GMM*，*i-vec*) の結果を示す．教師あり話者適応 (*Supervise*)，*Supervise* において話者コードに平均値を入力した平均声モデル (*AVM*)，との比較を行なった．

4.2.1　客観評価 (*AVM*，*Supervise*，*GMM-UBM* の比較)

図 1 に *AVM*，*Supervise*，*GMM* のメルケプストラム歪み，LF0 RMSE の結果を示す．GMM のミクスチャ数を 8，16，32，64，128，適応文章数は 10，50，100 とそれぞれ変更し，実験を行った．まず，*AVM* と *GMM* を比較すると，全ての適応文章数でメルケプストラム歪み，LF0 RMSE が減少しており，提案法により話者適応が行われていることがわかる．しかし，教師なし話者適応 *GMM* と教師あり話者適応である *Supervise* を比較すると，予想されるように *Supervise* においてメルケプストラム歪み，LF0 RMSE が小さい．

次に，図 1 において，提案法 *GMM* のミクスチャ数に注目すると，ミクスチャ数が 32，64 でそれぞれ LF0 RMSE，メルケプストラム歪みが最小となっている．話者認識において通常用いられるミクスチャ数より少数となっているが，提案法では GMM-UBM が話者類似度ベクトルを求めるために用いられることから，適切なミクスチャ数が異なったと考えられる．ミクスチャ数が少ない場合 (例えば，8) では，モデルがシンプルであり各学習話者を十分に表現できず，適切な話者類似度が求められていないと考えられる．また，提案法において，ミクスチャ数が多い複雑なモデルを用いると，学習話者に対する話者コードは One-hot ベクトルに近づいていく．これは，話者認識モデルの学習と話者類似度の計算に同じ音声データが用いられているためである．そのため，ミクスチャ数が多い場合には，話者類似度ベクトルを用いる話者適応に適した複数話者音声合成システムの構築が行われておらず，性能が低下したのだと考えられる．

以降の実験では，GMM のミクスチャ数を 64 に固定した．また，i-vector の計算のための UBM の GMM のミクスチャ数も 64 とした．

4.2.2　客観評価 (話者認識モデルの比較)

図 2 に *AVM*，*Supervise*，*GMM*，*i-vec* のメルケプストラム歪み，LF0 RMSE の結果を示す．*GMM(MFCC)*，*i-vec(MFCC)* では話者認識で通常用いられる 19 次 MFCC+Δ, Δ^2，*GMM(MFCC+F0)*，*i-vec(MFCC+F0)* では 19 次 MFCC とその Δ, Δ^2 に加え，実験条件に示した DCTF0 特徴量，*GMM(MGC)* では音声合成で利用されるスペクトル特徴量に近い，WORLD を用いて抽出された 19 次メルケプストラムとその Δ，Δ^2 を用い，話者認識モデルの構築を行なった．適応文章数は 10，50，100 である．

表 1　学習話者自身の話者類似度の平均

GMM(MFCC)	0.15
GMM(MFCC+F0)	0.087
i-vec(MFCC)	0.99
i-vec(MFCC+F0)	0.98

表 2　適応時に用いられる話者コードにおける話者類似度の上位数人の累積値

	上位 1 人	上位 2 人	上位 3 人
GMM(MFCC)	0.039	0.072	0.10
GMM(MFCC+F0)	0.041	0.075	0.11
i-vec(MFCC)	0.83	0.96	0.99
i-vec(MFCC+F0)	0.58	0.75	0.84

まず，*GMM(MFCC)* と *GMM(MFCC+F0)* を比較すると，*GMM(MFCC+F0)* はメルケプストラム歪みは若干増加するものの，LF0 RMSE の値は減少している．F0 に関する特徴量を利用することで，音声合成で重要となる F0 を考慮した話者類似度ベクトルが求められたと考えられる．*i-vec(MFCC)* と *i-vec(MFCC+F0)* においても同様の傾向が見られる．次に，*GMM(MFCC)* と *GMM(MGC)* を比較すると，メルケプストラム歪みはほぼ同程度の結果となっているが，*GMM(MGC)* は LF0 RMSE の値は非常に大きくなっている．これは音声合成で用いられるメルケプストラムと異なり，MFCC には F0 の情報も多く含まれていたためだと考えられる．

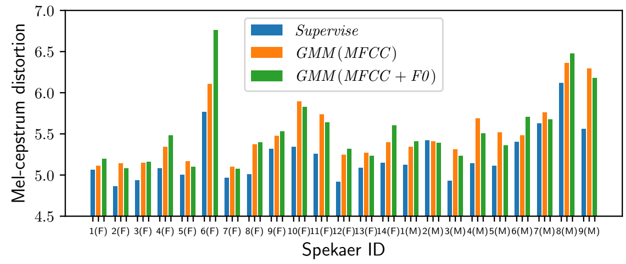
また，GMM-UBM を利用する *GMM* と i-vector/PLDA を利用する *i-vec* について比較すると，i-vector/PLDA に基づく手法はメルケプストラム歪み，LF0 RMSE の値が大きい．原因を調べるため，各手法で推定された話者コードについて調査した．表 1 に，各手法の学習話者の話者類似度ベクトルにおける学習話者自身の話者類似度の平均，表 2 に，各手法の適応時に用いられる話者類似度ベクトルにおける話者類似度の上位数人の累積値の平均を示す．表より，*GMM(MFCC)*，*GMM(MFCC+F0)* と比較して，

- *i-vec(MFCC)*，*i-vec(MFCC+F0)* では，複数話者音声合成システム構築に用いられる話者類似度ベクトルは，One-hot ベクトルに近い表現となっている．
- *i-vec(MFCC)*，*i-vec(MFCC+F0)* では，推定された目標話者に対する話者類似度ベクトルは，少数の学習話者に対する話者類似度が非常に高い．

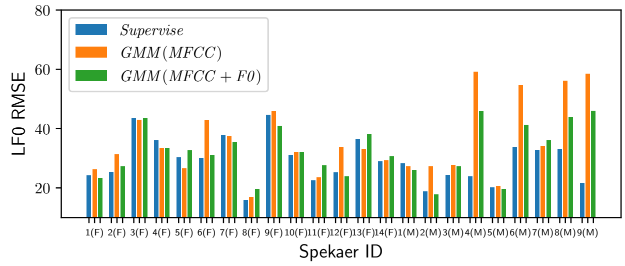
つまり，現在のパラメータ設定において，i-vector/PLDA を利用した話者認識モデルでは，学習話者に対しては話者類似度が適切に推定されておらず，適応時には学習話者から少数の話者の選択が行われていると考えられる．

4.2.3　客観評価 (各目標話者の結果)

図 3 に，*Supervise*，*GMM(MFCC)*，*GMM(MFCC+F0)* について目標話者 23 名それぞれの客観評価結果を示す．



(a) メルケプストラム歪み



(b) LF0 RMSE

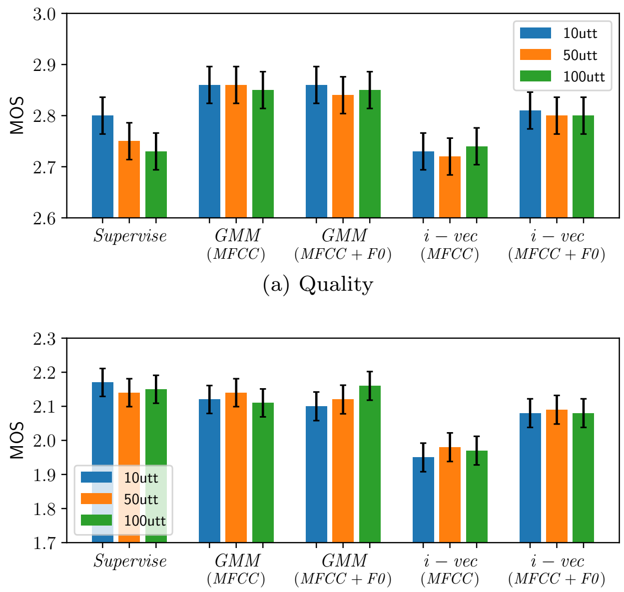
図 3　目標話者 23 名それぞれの客観評価結果を示す．適応文章数は 100 である．ここで，話者 ID のカッコ内は F は女性，M は男性を表す．

図 3 より，*Supervise* と比較し提案法は，多くの目標話者で同等，もしくは，若干誤差が大きい結果となっている．しかし，*GMM(MFCC)* における目標話者 4(M)，6(M) の LF0 RMSE ように，*Supervise* と比較しメルケプストラム歪み，LF0 RMSE が大きい話者が存在する．F0 に関する特徴量を用いることで，LF0 RMSE の改善が見られる話者もいるが，目標話者 6(F) のように F0 特徴量を用いたことにより，メルケプストラム歪みが非常に大きくなる話者も存在した．このような話者適応における外れ話者の対処は，今後の課題と言える．

4.2.4　主観評価

主観評価実験を行った．被験者はクラウドソーシングを用いて集め，日本語話者 180 人である．*Supervise*, *GMM(MFCC)*, *GMM(MFCC+F0)*, *i-vec(MFCC)*, *i-vec(MFCC+F0)* において，適応データとして各目標話者の 10, 50, 100 発話を用いて適応した，計 15 システムを評価した．音声サンプルは計 3,450 個 (15 システム ×23 目標話者 ×10 テスト文) である．合成音声の品質，リファレンス音声と比較した話者類似性をそれぞれ 5 段階 MOS により評価した．各音声サンプルは品質，話者類似性に関してそれぞれ 10 回ずつ評価した．

図 4 に主観評価実験の結果を示す．まず，品質において，*Supervise* の適応データの発話数の違いに注目すると，発話数が多いほど品質が低下していることがわかる．これは，複数話者音声合成システムの学習時に用いられる話者コード (One-hot ベクトル) と推定された話者コード (連続値) の表現が大きく異なることや，バックプロパゲーションによる推定に適切な停止基準が設定できていないことが原因とし



(a) Quality

(b) Speaker Similarity

図 4　主観評価実験結果

て考えられる．*Supervise* と提案法である *GMM(MFCC)*，*GMM(MFCC+F0)* を比較すると，提案法が高品質な音声の合成が行えていることがわかる．また，i-vector/PLDA をもちいた手法 *i-vec(MFCC)*，*i-vec(MFCC+F0)* は，GMM-UBM を用いた手法 *GMM(MFCC)*，*GMM(MFCC+F0)* と比較し，低評価となった．これらの結果から，GMM-UBM を用いることで，音声合成のための複数話者モデルの学習，話者適応に適切な話者類似度ベクトルを推定でき，高品質な音声合成が行われたと考えらえれる．

次に，図 4 に示される話者類似性の結果に注目すると，*Supervise*，*GMM(MFCC)*，*GMM(MFCC+F0)* は適応データの発話数に関わらず，ほぼ同等の性能となっている．F0 特徴量を利用した *i-vec(MFCC+F0)* は *i-vec(MFCC)* からの改善は見られるが，どちらの手法も他の手法と比較して評価は低い．i-vector/PLDA を用いた手法では，4.2.2 での客観評価結果と同様に，適切な話者類似度ベクトルが推定されなかったため，話者適応が適切に行われなかったと考えられる．GMM-UBM を用いた提案法では，高精度な話者適応が行えることが示された．

5. おわりに

本研究では，入力コードに基づく DNN 音声合成において，テキストを必要としない，教師なし話者適応について検討した．話者認識において広く用いられているモデル (GMM-UBM，i-vector/PLDA) を利用し計算された，個々の学習話者に対する事後確率を連結したベクトルを，入力コードとして利用した．各学習話者に対する事後確率が，話者類似度を表現すると仮定し，DNN に基づく複数話者音声合成システムの構築を行い，目標話者の音声データから計算された話者類似度ベクトルを入力コードとして利用することで，話者適応を行った．主観評価実験の結果，GMM-UBM に基づく話者認識モデルを用いた提案法において，合成音声の品質を下げることなく，高精度な話者適応が可能であることを確認できた．

今後の課題として，ニューラルネットワークによる話者類似度の計算，ノイズ等を含む低品質音声データを用いた，提案法による教師なし話者適応が挙げられる．

6. 謝辞

本研究は株式会社オルツの助成を受けた．

参考文献

[1] N. Hojo, Y. Ijima, and H. Mizuno, “An investigation of DNN-based speech synthesis using speaker codes,” in *Proc. Interspeech*, 2016.

[2] S. Ö. Arik, G. F. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, “Deep voice 2: Multi-speaker neural text-to-speech,” *CoRR*, vol. abs/1705.08947, 2017. [Online]. Available: http://arxiv.org/abs/1705.08947

[3] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, “Voice synthesis for in-the-wild speakers via a phonological loop,” *CoRR*, vol. abs/1707.06588, 2017. [Online]. Available: http://arxiv.org/abs/1707.06588

[4] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, “A study of speaker adaptation for DNN-based speech synthesis,” in *Proc. Interspeech*, 2015.

[5] P. Swietojanski and S. Renals, “Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models,” in *Proc. SLT*, 2014, pp. 171–176.

[6] Y. Fan, Y. Qian, F. K. Soong, and L. He, “Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis,” in *Proc. ICASSP*, 2015, pp. 4475–4479.

[7] H.-T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, “Adapting and controlling DNN-based speech synthesis using input codes,” *Proceedings of ICASSP*, pp. 4905–4909, 2017.

[8] S. King, K. Tokuda, H. Zen, and J. Yamagishi, “Unsupervised adaptation for HMM-based speech synthesis,” *Proc. Interspeech*, pp. 1869–1872.

[9] R. Doddipatla, N. Braunschweiler, and R. Maia, “Speaker adaptation in DNN-based speech synthesis using d-vectors,” *Proc. Interspeech*, pp. 3404–3408, 2017.

[10] J. S. Bridle and S. Cox, “RecNorm: Simultaneous normalisation and classification applied to speech recognition.” in *Proc. NIPS*, 1990, pp. 234–240.

[11] J. Yamagishi and T. Kobayashi, “Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training,” *IEICE T. Inf. Syst.*, vol. 90, no. 2, pp. 533–543, 2007.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing*, vol. 10, pp. 19–41, 2017.

[13] P. Kenny, “Bayesian speaker verification with heavy-tailed priors,” *Odyssey 2010*, 2010.

[14] A. Larchera, K. A. Lee, and S. Meignier, “An extensible speaker identification sidekit in python,” *Proceedings of ICASSP*, 2016.

[15] M. Morise, “Cheaptrick, a spectral envelope estimator for high-quality speech synthesis,,” *Speech Communication*, vol. 67, pp. 1–7, 2015.