Vol.2016-SLP-112 No.3
2016/7/28

Deep Learningを利用した任意話者の声質変換

関井 祐介[1,a)]　折原 良平[1]　小島 圭介[2]　清 雄一[1]　田原 康之[1]　大須賀 昭彦[1]

概要：声質変換手法として Gaussian Mixture Model(GMM) を用いた手法や Deep Neural Network(DNN) を用いた手法が研究されている．これらの多くは一対一の声質変換手法を提案しており，複数話者の入力に対応した研究は多くない．また，従来の DNN を用いた声質変換手法では，一対一変換および多対一変換において複雑なネットワークを用いるため，変換に要する時間が長くなるという問題がある．本研究では，複数話者の声質変換に対応するにあたり，オートエンコーダを用いた声質変換手法を提案する．提案手法では，オートエンコーダで次元圧縮した高次特徴量を目的話者の高次特徴量へ DNN で変換し，目的話者のオートエンコーダを用いて音響特徴量に復元する．評価実験では，従来の DNN を用いた声質変換手法より声質変換精度が向上し，変換に要する時間を短縮できたことを確認した．

1. はじめに

1.1 背景

近年，入力話者の声質のみを目的話者の声質に変換する声質変換技術が盛んに研究されている．声質変換技術は，アニメーション作品において声優の変更による違和感を緩和させることや [1]，海外映画の吹替音声を役者本人の声質で作成すること [2] などに適用できると考えられている．

従来の代表的な声質変換手法として，GMM(Gaussian Mixture Model) を用いた声質変換手法が研究されている [3], [4]．しかし，近年 DNN(deep neural network) を用いた声質変換手法が GMM を用いた手法より高い変換精度をもたらすことが報告されている [5]．これは，人間の声道形状が非線形的であるのに対し，GMM を用いた手法は線形変換をベースにしており，一方，DNN は非線形ベースの変換を行っているためであると考えられる [6]．非線形ベースの声質変換手法として，RBM(restricted Boltzmann machine) を用いた変換手法 [7] や RBM を拡張した DBN(deep belief network) を用いた変換手法 [9]，CRBM(conditional restricted Boltzmann machine) を用いた変換手法 [10] 等が提案されている．また，DNN を用いた声質変換手法では，事前学習に RBM やオートエンコーダを用いることにより変換精度が向上することが報告されている [11], [15]．

声質変換の研究の多くは，音響特徴量として MFCC(Mel-Frequency Cepstrum Coefficients) を用いているが，スペクトル包絡の変換を行った方が類似性が高くなるという結果が報告されている [7], [12]．MFCC 変換では，MFCC からスペクトル包絡に復元する際，高周波数域の情報が損失してしまい，高周波数域の類似性がスペクトル包絡の変換に比べ劣ってしまうことが原因と考えられる [8]．このため，声質変換に用いる音響特徴量としてスペクトル包絡を選択すべきだと言える．しかし，スペクトル包絡は MFCC に比べ次元数が大きく，声質変換器の作成や声質変換に時間を要するという問題がある．特に，声質変換に時間を要するという問題は，アプリケーションの形態に制約を課してしまうため，声質変換技術の応用先を広げるためにも，この問題を解決する必要がある．

以上のことから，高精度な声質変換を行うためには，音響特徴量としてスペクトル包絡を用い，非線形ベースの変換を行うことが好ましいと考えられる．DNN を用いてスペクトル包絡の変換を行う手法 [13] が提案されているが，DNN の入力に次元数の大きい対数スペクトル包絡を用いているため，DNN の構造が複雑になり，変換に要する時間が長くなっている．

今まで述べた手法は特定の入力話者音声から特定の目的話者音声への一対一変換であったが，任意の入力話者音声を任意の目的話者音声へ変換する多対多変換を実現する手法 [14], [15] が提案されている．任意話者の声質変換を行うために，複数話者の音声データを用いて訓練させるため，一対一変換の声質変換器よりも多くの音声データが必要となる．しかし，いずれの研究も十分な精度を得るための学習に要する最小の訓練話者数は明記されておらず，任意話

[1] 電気通信大学大学院情報システム学研究科
Graduate School of Information Systems, The University of Electro-Communications
[2] ソリッドスフィア株式会社
Solid Sphere, inc.
[a)] sekii.yusuke@ohsuga.is.uec.ac.jp

者の声質変換を行うために何人分の音声データを用いるべきか明らかでないという問題もある．

1.2 研究目的

本研究では，オートエンコーダと簡易な DNN を用いることにより，一対一の声質変換および多対一の声質変換に要する時間の短縮を行うことを目的とする．

本稿の構成は以下の通りである．第 2 章ではオートエンコーダを利用した声質変換手法や任意話者の声質変換手法を提案している研究を紹介する．第 3 章では，本研究で提案する手法の全体像を説明し，オートエンコーダの仕組みとメリットについて述べる．第 4 章では評価実験として，一対一変換および多対一変換について既存手法との比較実験を行う．第 5 章で本稿をまとめ，今後の課題について議論する．

2. 関連研究

声質変換に関する研究は数多く存在するが，ここではオートエンコーダを利用した声質変換手法，DNN を用いた任意話者の声質変換手法に関する研究について示す．

Nguyen ら [12] は，スペクトル包絡，F0（基本周波数）と発話の長さを総合的に変換する話者変換手法を提案した．声質変換にあたるスペクトル包絡の変換では，重みに L1 正則化を用いたオートエンコーダで事前学習する手法を提案しており，重みをランダムに初期化するものよりも高い精度でスペクトル包絡の変換を行った．しかし，高精度に変換を行える一方，DNN の入力に 512 次元の対数スペクトル包絡を用いており，DNN の隠れ層が 3 層で，隠れ層の素子数が 3000 と大規模な NN となっているため，変換に長時間を要するという問題がある．

Mohammadi ら [11] は，DNN を用いた声質変換の事前学習にディープオートエンコーダを利用した声質変換手法を提案した．入力話者，目的話者それぞれのディープオートエンコーダを用いて入力特徴量を圧縮し，入力話者の圧縮された特徴量（以下，高次特徴量）を目的話者の高次特徴量に変換する ANN（Artificial Neural Network）を作成した．作成したディープオートエンコーダと ANN を結合した DNN を作成し，最後に fine-tuning（再学習）を行った．これは，小規模なコーパスでの訓練の時，GMM 等の既存手法よりも優位であった．DNN の入力は 24 次元の MCEP(melcepstrum) であるため，DNN の隠れ層の層数が増えても，変換に要する時間は比較的短い．

Liu ら [15] は，DNN を用いた話者非依存の声質変換手法を提案した．変換したいフレームの特徴量とその前後のフレームの特徴量を合わせて入力とし，かつ複数話者の音声データを訓練に利用することで話者非依存の声質変換手法を実現した．また，話者非依存の DNN を初期値とし，一対一の話者依存 DNN を作成する手法は，事前学習に DBN を用いる手法よりも優れていた．評価実験では，一対一の声質変換手法である GMM および既存の DNN を用いた手法と大差ない精度で任意話者の声質変換を実現した．Liu らも Mohammadi らと同様，音響特徴量に 24 次元 MCEP を用いている（ただし，前後 1 フレーム分を合わせて入力しているので，入力は 72 次元となる）ため，変換に要する時間は比較的短いと考えられる．

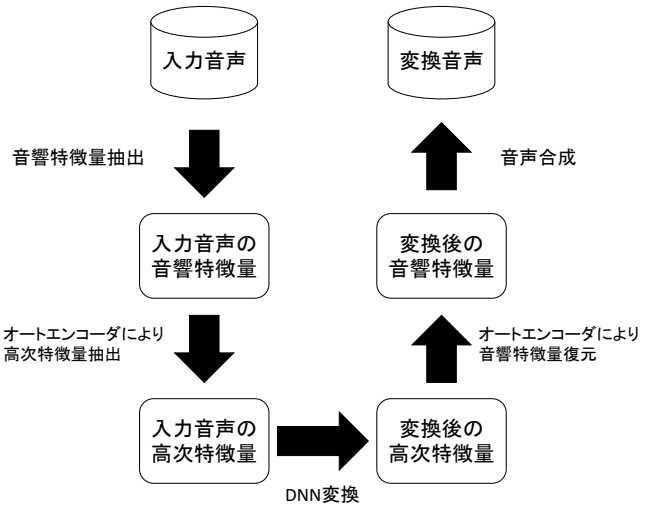

図 1 声質変換の流れ

3. 提案手法

3.1 提案手法の全体像

本研究では**図 1** の流れで声質変換を行う．

まず，入力音声から音響特徴量（本研究ではスペクトル包絡）を抽出する．次に入力話者音声で訓練されたオートエンコーダを用いて高次特徴量の抽出を行う．入力音声の高次特徴量を DNN を用いて目的話者音声の高次特徴量へ変換する．変換された高次特徴量を目的話者音声で訓練されたオートエンコーダを用いて音響特徴量へ復元する．得られた音響特徴量を元に，音声合成によって変換音声を求める．

3.2 オートエンコーダ

一般的な Neural Network(NN) は教師あり学習の手法であり，入力値と出力値の組が必要となる．オートエンコーダ（Autoencoder）[16] は教師なし学習の一手法であり，出力値が入力値をそのまま再現するようなな Neural Network(NN) であるため，入力値のみを必要とする．

図 2 のように入力層（x）と隠れ層（h），出力層（y）の 3 層からなる NN を考えた時，オートエンコーダを以下のように表す．

$$\mathrm{h} = f(\mathrm{W}_1\mathrm{x} + \mathrm{b}_1) \tag{1}$$

$$\mathrm{y} = g(\mathrm{W}_2\mathrm{h} + \mathrm{b}_2) \tag{2}$$

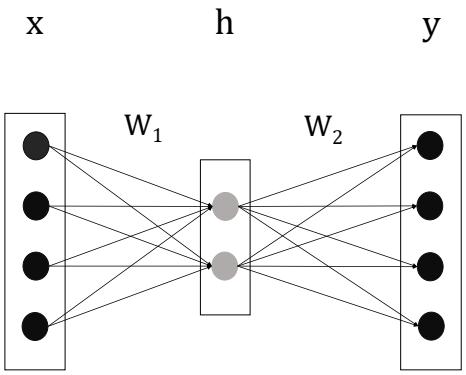

図 2　オートエンコーダ

W_1，b_1 はそれぞれ x を h へ変換する際の重みとバイアス，W_2，b_2 はそれぞれ h を x へ変換する際の重みとバイアスで，f と g は活性化関数である．式（1），（2）より，入力 x を変換し出力 y を求める式は以下のようになる．

$$y = g(W_2 f(W_1 x + b_1) + b_2) \tag{3}$$

オートエンコーダでは y が x に近くなるようにパラメータである重み W_1, W_2 とバイアス b_1, b_2 を決定する．つまり，y と x の近さを測るための誤差関数の値を最小化するようにパラメータを決定する．またここで，$W_1 = W_2^{\mathrm{T}}$ と制約することもある．誤差関数は一般的に平均二乗誤差が用いられることが多い．

$$E = ||x - y||^2 \tag{4}$$

オートエンコーダは RBM と同じく事前学習の手法として用いられることが多く，オートエンコーダを利用して NN に初期値を与え，fine-tuning することでより良い結果を得ることができる [11]．オートエンコーダの隠れ層を入力層の次元よりも小さくすることで，次元圧縮された特徴量（高次特徴量）を抽出することもできる．これにより，次元の大きい特徴量を比較的次元の小さい特徴量として表すことが可能となる．

3.3　特徴量変換

本研究では，オートエンコーダから抽出した高次特徴量を利用する．次元の小さい高次特徴量を利用することにより，従来手法よりも変換に要する時間を短縮することが目的である．

本研究では，**図 3** のような特徴量変換構造を提案する．まず，入力話者の音響特徴量 x，目的話者の音響特徴量 x′ を入力とし，入力話者と目的話者各々のオートエンコーダを作成する．次に，入力話者のオートエンコーダと目的話者のオートエンコーダからそれぞれ高次特徴量（隠れ層の出力値）h，h′ を抽出する．入力話者オートエンコーダから抽出した高次特徴量 h を入力データ，目的話者オートエンコーダから抽出した高次特徴量 h′ を正解データとする DNN を作成する．最後に，高次特徴量変換を行う DNN で変換された高次特徴量 h″ を目的話者オートエンコーダのデコーダ重み W_2' を利用することで音響特徴量を復元し，変換された音響特徴量 y″ を得る．

任意話者変換（多対一変換）に対応するためには，訓練データとして複数の入力話者を用意し学習させる．オートエンコーダの高次特徴量は複数話者から構成される訓練データを用いることで，より一般化された高次特徴量となると考えられる．一般化された高次特徴量を目的話者の高次特徴量に変換する DNN を用いることで，訓練データとして用いていない任意の入力話者の音声でも高精度に変換できることが期待される．

4.　評価実験

4.1　評価方法

一対一変換と任意話者変換（多対一変換）において提案手法を含む複数手法の比較実験を行った．実験には，ソリッドスフィア社が作成した音声データセット [*1] を用いた．一対一変換では，男性話者 2 人（KJM，YMGT）と女性話者 2 人（HM，TK）からなる変換の組合せを 4 組（YMGT → KJM，KJM → HM，TK → YMGT，HM → TK）作成し実験を行った．任意話者変換の声質変換器作成において，目的話者 2 人と訓練に利用する話者数を 2，4，6，8 と変化させた 4 パターンの直積である計 8 組の変換器を作成した．なお，目的話者には男性話者（KRT）と女性話者（HM）を用い，一方を目的話者とし，もう一方を評価用の入力話者として実験を行った．また，一対一変換と任意話者変換ともに，300 発話を訓練データ，50 発話をテストデータとして評価を行った．なお，訓練データには入力話者と目的話者の同一内容発話から動的計画法でアラインメントを取ることにより作成されたパラレルデータを用いた．

実験では提案手法であるオートエンコーダの高次特徴量を用いた手法と 3 つの従来手法を用いて声質変換精度比較を行った．提案手法には，50 次元の高次特徴量を用いた手法（our1）と 100 次元の高次特徴量を用いた手法（our2）の 2 つの手法を用いた．従来手法には，GMM を用いた手法（JDGMM）[4]，MFCC を DNN を用いて変換した手法（mfcc）[5]，対数スペクトル包絡を DNN を用いて変換した手法（spec）[12] の 3 つの手法を用いた．入出力音響特徴量として，TANDEM-STRAIGHT[17] により求めた 513 次元の対数スペクトル包絡を our1，our2，spec で用い，スペクトル包絡より計算された 25 次元 MFCC を JDGMM，mfcc で用いた．JDGMM における混合数は 64 とし，mfcc，spec，our1，our2 の隠れ層及び隠れ層素子数はそれぞれ，2 層 50 素子，3 層 3000 素子，2 層 200 素子，

[*1]　男性話者 4 名，女性話者 6 名，各話者 500 発話収録

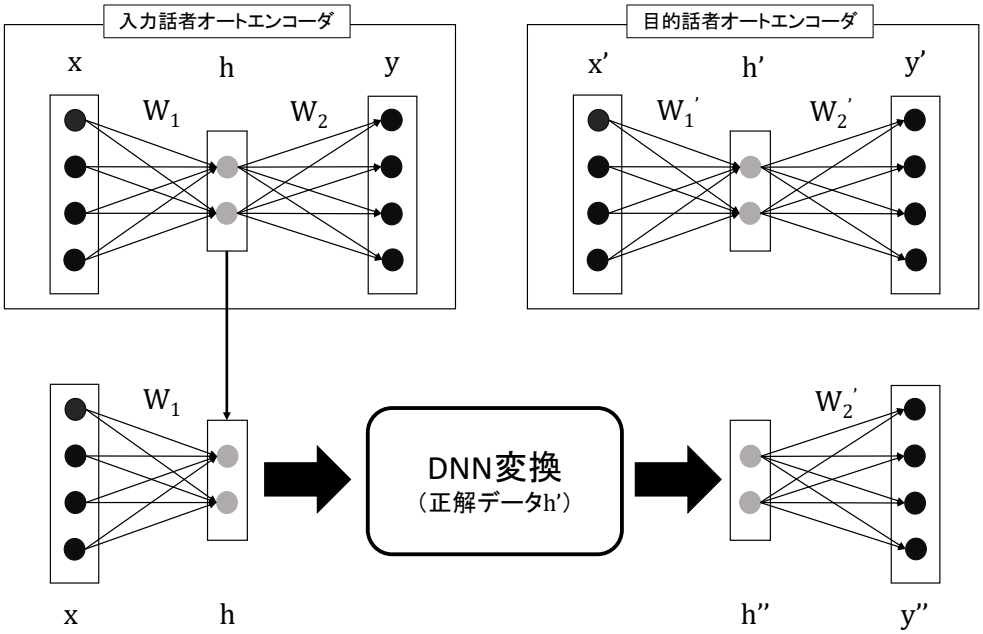

図 3 特徴量変換の全体構造

3 層 500 素子とした．オートエンコーダおよび各 DNN の活性化関数，学習最適化アルゴリズムはそれぞれ ReLU 関数 [18]，ADAM（α=0.0001）[19] を用いた．学習回数については，mfcc，spec はそれぞれ 200，20 とし，our1 と our2 ではともに，オートエンコーダの学習回数は 100，DNN の学習回数は 30 とした．これらの値について，mfcc は Desai らの手法 [5] を参考にし，他は予備実験により決定した．

客観評価基準として，変換音声スペクトルが目的話者音声スペクトルにどのくらい近いかを表す尺度である LSD（log spectral distortion）と音響特徴量の変換所要時間の 2 つを用いた．また，LSD の評価式は以下の通りである．

$$\mathrm{LSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(10\log_{10}\frac{x_i}{y_i}\right)^2} \tag{5}$$

x_i は変換音声の i 番目のスペクトル，y_i は目的話者音声の i 番目のスペクトルである．MOS（mean opinion score）に基づく主観評価では，被験者 9 人に目的話者音声と変換音声を視聴させ，類似性（変換音声の声質が目的話者音声の声質に似ているか）と自然性（発話がはっきりしているか）の 2 項目について 1 から 5 の 5 段階で評価させた．また，一対一変換における主観評価には 2 つの変換組（TK → YMGT と HM → TK）の声質変換器を用い，それぞれ訓練データに用いてないランダムに選択した 1 発話を評価に用いた．任意話者変換における主観評価には，HM を目的話者とした声質変換器を用い，訓練データに用いてないランダムに選択した 1 発話を評価に用いた．

4.2 実験結果と考察

4.2.1 一対一変換の結果

一対一変換において，各手法によりスペクトル変換した 4 組の LSD 評価結果を**表 1** に示した．LSD 値を比較すると，提案手法である our1，our2 および先行研究手法である spec は JDGMM および mfcc より高いスペクトル変換精度が得られた．これは，JDGMM および mfcc は音響特徴量に MFCC を利用しているため，スペクトル包絡に復元した際，高周波数成分が欠落してしまい，スペクトル包絡の類似度が低くなってしまうためと考えられる．また，our1，our2，spec の 3 手法のスペクトル変換精度に大きな差はなかったが，our1 より our2 の精度が高かったことから，次元の大きい高次特徴量を用いた方が高い精度を得られることがわかる．

図 4 では，変換した音声の類似性と自然性を人間の聴覚に基づき評価した結果を示した．また，各手法の MOS 値は 2 つの変換組でそれぞれ得られた値の平均値である．類似性では手法間に大きな差はなかったが，提案手法である our1 が最も MOS 値が高かった．自然性では提案手法である our1 と our2 が他手法よりも値が高く，高品質な声質変換を行えていることがわかる．主観評価において，手法間の差が LSD におけるものよりも小さくなったのは，MFCC が人間の音声知覚を考慮した特徴量であるためと考えられる．

表 2 では，各手法によるスペクトル変換に要する時間の比較を行った．対数スペクトル包絡を変換した手法は入力対数スペクトル包絡から変換されたスペクトル包絡を求

表 1　一対一変換の LSD（dB）

target	our1	our2	JDGMM	mfcc	spec
YMGT → KJM	4.08	4.04	5.06	5.19	4.06
KJM → HM	4.29	4.20	4.72	4.96	4.21
TK → YMGT	4.02	3.96	5.04	5.10	3.96
HM → TK	3.88	3.82	4.55	4.50	3.88
average	4.07	**4.01**	4.84	4.94	4.03

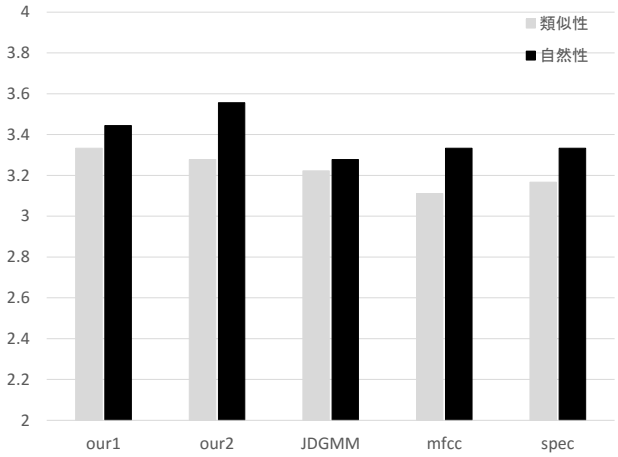

図 4　一対一変換の主観評価

表 2　変換時間（s）

our1	our2	JDGMM	mfcc	spec
14.6	17.7	41.7	6.1	73.0

めるまでの時間，MFCC を変換した手法は，入力 MFCC から変換された MFCC を求めるまでの時間を計測した．また，ここでの変換時間は 50 発話分のスペクトルを全て変換するのに要した時間である．表 1 より，our1，our2，spec に大きな変換精度差はなかったが，変換時間の比較では，提案手法である our1 と our2 が変換に要した時間が 20 秒以内であるのに対し，先行研究手法である spec は変換に要する時間が 73 秒と 4 倍以上の時間を要するという結果となった．また，従来手法である JDGMM は mfcc より変換精度は高かったが，mfcc の変換時間が約 6 秒に対し，JDGMM は変換に 40 秒以上要するということから，音響特徴量として MFCC を用いる場合は変換精度と変換時間はトレードオフの関係になると考えられる．

以上の結果より，一対一変換において提案手法である our2 が変換精度，変換時間共に優れていたことがわかる．

4.2.2　任意話者変換の結果

各手法によりスペクトル変換の対象とする目的話者 2 人と訓練話者数の組合せ 4 組の直積である 8 通りに対し，各手法の LSD 評価を行った結果を**表 3** に示した．例えば，KRT_mix2 というのは，目的話者を KRT とする任意話者声質変換器の作成に用いた訓練データの話者が 2 人であることを意味する．また，JDGMM については，各変換組（TK → KRT，KRT → TK）の一対一変換を行った結果を

表 3　任意話者変換の LSD（dB）

target	our1	our2	JDGMM	mfcc	spec
KRT_mix2	4.15	4.11	(4.72)	5.44	4.16
KRT_mix4	4.16	4.08	—	5.13	4.12
KRT_mix6	4.11	4.12	—	5.16	4.19
KRT_mix8	4.14	4.07	—	5.31	4.16
TK_mix2	4.60	4.60	(5.01)	5.09	4.64
TK_mix4	4.48	4.49	—	5.09	4.45
TK_mix6	4.48	4.40	—	4.99	4.47
TK_mix8	4.49	4.46	—	4.93	4.56
average	4.33	**4.29**	4.87	5.14	4.34

記載した．表 1 と同様，JDGMM と mfcc は他手法に比べ精度が落ちるが，理由は前項で述べた通りである．任意話者変換では，our1，our2 共に spec よりも高い精度が得られた．このことから，オートエンコーダを用いることにより，複数話者の訓練データから一般的な高次特徴量が得られ，直接特徴量を変換するより話者に依存しない変換が行えるようになったと考えられる．

表 4 では各手法の訓練話者数別の LSD を示した．訓練に用いた話者を変化させた時，訓練に用いた話者が 2 人の時の精度が低かったことは各手法に共通しているが，訓練用話者が 4 人以上になると大きな差はなかった．今回は訓練用話者を 10 人以上として実験を行わなかったためこれ以降の変化は不明だが，この結果を考慮すると必ずしも訓練に多くの話者を用いることが声質変換の精度向上に結びつくとは言えない．10 人以上に増やした場合も精度に大きな影響がない可能性も十分あり得ると考えられる．

図 5 では，任意話者変換における主観評価の結果を示した．類似性と自然性共に，同じ手法であれば訓練話者数が 2 人の声質変換器より訓練話者数が 8 人の声質変換器の方が MOS 値が高かった．これは表 3 の結果からもわかるように，訓練話者数は 2 人よりも 4 人以上用いた方がより目的話者に似ている高品質な音声を作成できるからと考えられる．手法間の差については，表 3 の結果に反し，our1 や our2 よりも spec の MOS 値が高かった．また，一対一変換である JDGMM より任意話者変換を行う spec が類似性と自然性共に優れていたことから，任意話者変換においても，スペクトル包絡を用いることで従来の一対一変換手法である GMM を用いた変換手法並かそれ以上の精度で声質変換を行うことができることがわかった．spec の主観評価結果が優れていた理由として，spec で用いた DNN の構造が複雑なため，様々な入力に対応しやすかったことが考えられる．ただし，主観評価に用いた評価用の音声が 1 発話だったため，偶然上手く変換できた発話を選んでいた可能性も考えられる．

5.　まとめ

本研究では，複数話者の声質変換に対応させ，かつ特徴量

表 4　訓練話者数に対する LSD（dB）

訓練話者数（人）	our1	our2	mfcc	spec
2	4.38	4.35	5.26	4.40
4	4.32	4.28	5.11	4.29
6	4.30	4.26	5.08	4.33
8	4.32	4.27	5.12	4.36

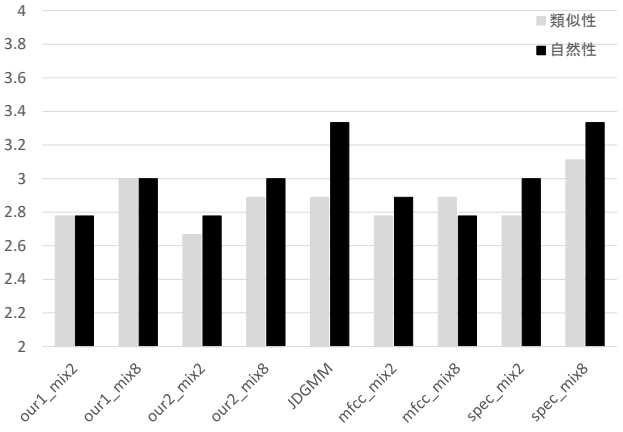

図 5　任意話者変換の主観評価

の変換時間を短縮することを目的として，オートエンコーダを利用した声質変換手法を提案した．評価実験では，一対一変換，任意話者変換ともに，従来手法よりも変換精度，変換時間で優れていた．今後は，高次特徴量変換の DNN の初期値を事前学習を用いて決定することや，オートエンコーダと高次特徴量変換 DNN を結合した DNN を構成し，fine-tuning を行うなど，更なる精度向上を目指す予定である．また，任意話者変換実験で訓練に用いる話者数を増やすことで，精度がどのように変化するか観察し，適切な訓練人数の特定を行いたい．

謝辞　本研究は JSPS 科研費 26330081，26870201，16K12411 の助成を受けたものです．本研究を遂行するにあたり，研究の機会と議論・研鑽の場を提供して頂き，御指導頂いた国立情報学研究所／東京大学 本位田 真一 教授をはじめ，活発な議論と貴重な御意見を頂いた研究グループの皆様に感謝致します．

参考文献

[1] 塩出萌子，小泉悠馬，伊藤克亘: 中間話者コーパスを用いたアニメーション演技音声のための話者変換，第 76 回全国大会講演論文集, pp. 495-496, 2014.
[2] 見原隆介，齋藤大輔，峯松信明，広瀬啓吉: 音声の構造的表象に基づく異言語間・異話者間の音声変換手法，電子情報通信学会技術研究報告. SP, 音声 109(308), pp. 55-60, 2009.
[3] Y. Stylianou, O. Cappe and E. Moulines: Continuous probabilistic transform for voice conversion, IEEE Transactions on Speech and Audio Processing, vol.6, no.2, pp. 131-142, 1998.
[4] T. Toda, A. W. Black and K. Tokuda: Voice conversion based on maximum likelihood estimation of spectral parameter trajectory, IEEE Transactions on Speech and Audio Processing, vol.15, no.8, pp. 2222-2235, 2007.
[5] S. Desai, E.V. Raghavendra, B. Yegnanarayana, A. W. Black and K. Prahallad: Voice conversion using articial neural networks, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3893-3896, 2009.
[6] 中鹿亘，滝口哲也，有木康雄: 話者依存型 Recurrent Temporal Restricted Boltzmann Machine を用いた声質変換，日本音響学会研究発表会講演論文集，pp. 219-222, 2014.
[7] L.H. Chen, Z.H. Ling, Y. Song, and L.R. Dai: Joint spectral distribution modeling using restricted Boltzmann machines for voice conversion, Proceeding of INTERSPEECH, pp. 3052-3056, 2013.
[8] 中鹿亘，滝口哲也，有木康雄: 話者適応型 Restricted Boltzmann Machine を用いた声質変換の検討，電子情報通信学会技術研究報告. SP, 音声 114(365), pp. 165-170, 2014.
[9] T. Nakashika, R. Takashima, T. Takiguchi and Y. Ariki: Voice Conversion in High-order Eigen Space Using Deep Belief Nets, Proceeding of INTERSPEECH, pp. 369-372, 2013.
[10] Zhizheng Wu, Eng Siong Chng and Haizhou Li: Conditional restricted boltzmann machine for voice conversion, IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP), pp. 104-108, 2013.
[11] S.H. Mohammadi and A. Kain: Voice conversion using deep neural networks with speaker-independent pretraining, IEEE Spoken Language Technology Workshop (SLT), pp. 19-23, 2014.
[12] Hy Quy Nguyen, Siu Wa Lee, Xiaohai Tian, Minghui Dong and Eng Siong Chng: High quality voice conversion using prosodic and high-resolution spectral features, Multimedia Tools and Applications, Volume 75, Issue 9, pp. 5265-5285, 2016.
[13] Feng-Long Xie, Yao Qian, Yuchen Fan, Frank K. Soong and Haifeng Li: Sequence error (SE) minimization training of neural network for voice conversion, Proceeding of INTERSPEECH, pp. 2283-2287, 2014.
[14] T. Toda, Y. Ohtani and K. Shikano: Eigenvoice conversion based on Gaussian mixture model, Ninth International Conference on Spoken Language Processing (ICSLP), pp. 2446-2249, 2006.
[15] Li-Juan Liu, Ling-Hui Chen, Zhen-Hua Ling and Li-Rong Dai: Spectral conversion using deep neural networks trained with multi-source speakers, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4849-4853, 2015.
[16] G. E. Hinton and R. R. Salakhutdinov: Reducing the dimensionality of data with neural networks, Science, 313(5786), pp. 504-507, 2006.
[17] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino and H. Banno: Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3933-3936, 2008.
[18] Vinod Nair and Geoffrey E. Hinton: Rectified Linear Units Improve Restricted Boltzmann Machines, Proceedings of the 27th International Conference on Machine Learning (ICML), pp. 807-814, 2010.
[19] Diederik Kingma and Jimmy Ba: Adam: A Method for Stochastic Optimization, International Conference for Learning Representations (ICLR), 2015.