Vol.2017-MUS-115 No.20
2017/6/17

DNN を用いた話者空間内の変換経路を考慮した声質変換の検討

小谷 岳[1,a)]　齋藤 大輔[1,b)]　峯松 信明[1,c)]

概要：従来の話者変換において，入力-出力話者間のモデル化には様々な手法が検討されてきた．しかし，入力-出力話者の特徴的な組み合わせが変換精度に与える影響の検討や，更には変換精度が低くなる話者対に対する検討は，未だ十分ではない．本研究ではこの問題に対し，入力-出力話者間の話者距離と変換精度の間にある関係を分析することを試み，話者空間内の変換経路を考慮した話者変換を検討する．特に，入力-出力話者間の距離が遠い場合に，第三話者（中間話者）のデータも用いることで，話者変換の精度を向上させることを試みる．具体的には，入力音声から出力音声への変換過程で中間話者の音声を経由するような変換モデルの構築を提案する．実験の結果，メルケプストラム歪みを用いた客観評価において提案手法の有効性が示された．

キーワード：声質変換，Deep Learning，話者空間

1. はじめに

声質変換は，入出力の対応関係を記述するモデルに基づき，任意の文に対して入力音声の声質を所望の声質へ変換する技術である [1]．特に，音声に内在する話者性を操作する変換を話者変換と呼び，テキスト音声合成において出力音声の話者性の変換などに応用されている [2]．従来の話者変換手法では，入力-出力話者間のモデル化に混合ガウス分布モデル (Gaussian mixture model; GMM) やディープニューラルネットワーク (Deep neural networks; DNN) といった様々な手法が検討されてきた [3], [4]．このようなモデル化手法の改善により，話者変換の精度は着実に向上している．しかし，そもそも入力-出力話者間をモデル化する際に，同性間の話者変換よりも異性間の話者変換の方が難しいといった，入力-出力話者の組み合わせにモデル化の難易が存在することが指摘されているが，この問題に対するアプローチは十分ではない [5]．本研究ではこの問題に対し，まず，入力-出力話者間の変換精度と，変換前の入力-出力自然音声間におけるメルケプストラム歪みに基づく話者距離との間にある関係性を明らかにする．次に，この分析的検討に基づき，従来手法では比較的変換精度が劣る入力-出力ペアに対して変換精度を向上させることを試みる．我々は新しい話者変換手法として，DNN を用いた話者空間内の変換経路を考慮した変換を提案する．特に，入力-出力話者間の距離が遠い場合に，入力-出力話者のデータだけではなく，第三話者（中間話者）のデータも用いることで，入力-出力話者間の変換精度を向上させることを検討する．具体的には，入力-中間話者間及び中間-出力話者間の DNN 変換モデルを各々のパラレルデータをもとに学習し，それら変換モデルを連結した入力-中間-出力話者モデルを，入力-出力話者間の事前学習モデルとする．この連結モデルを，入力-出力話者のパラレルデータをもとに再学習を行うことで，入力-出力話者間の変換モデルを構築する．実験の結果，提案手法により変換精度が向上することを客観評価において示す．

[1] 東京大学大学院工学系研究科
Graduate School of Engeneering, The University of Tokyo
[a)] kotani@gavo.t.u-tokyo.ac.jp
[b)] dsk_saito@gavo.t.u-tokyo.ac.jp
[c)] mine@gavo.t.u-tokyo.ac.jp

2. DNN を用いた従来の話者変換

本章では DNN を用いた従来の話者変換手法について述べる．DNN は多層の隠れ層を有するニューラルネットワークである．DNN では層 l の出力特徴量を $h^{(l)}$ とすると，層間を接続する変換関数は，前段の隠れ層からの線形変換と活性化関数 $g(x)$ の組み合わせによって以下のように表される．

$$h^{(l)} = g(W^{(l)}h^{(l-1)} + b^{(l)}) \tag{1}$$

本研究では活性化関数 $g(x)$ としてランプ関数 (Rectified Linear Unit; ReLU) を用いる．最終層の活性化関数につい

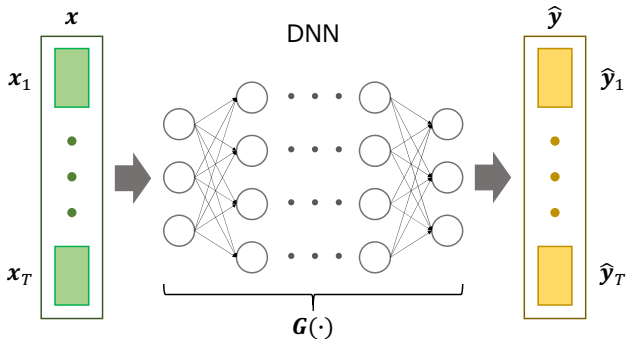

図 1　DNN を用いた話者変換

ては，声質変換のような連続値に対する回帰問題では線形写像が広く持ちいられており，本研究においても同様である．また一般的な DNN の学習は，微分可能な誤差基準のもとで誤差逆伝搬法を用い，ミニバッチ単位での確率的勾配降下法によってパラメータの最適化が行われる．本研究では声道スペクトルの静的特徴量を用い，バッチサイズを 1 文として二乗誤差基準のもとに学習を行う．図 1 に示すように，入力音声特徴量系列 $x = [x_1^\top, \cdots, x_t^\top, \cdots, x_T^\top]^\top$ から，DNN により合成音声の静的特徴量系列 $\hat{y} = G(x)$ を推定する．

3. 話者距離と音声変換精度の関係

3.1 メルケプストラム歪みに基づいた話者距離を利用した場合

3.1.1 メルケプストラム歪みに基づいた話者距離

入力-出力話者のパラレルデータに対するメルケプストラム歪みを話者距離とすることを考える．つまり，話者変換前の入力-出力音声の声道スペクトル特徴量間距離が，変換音声と出力音声の声道スペクトル特徴量間距離に与える影響を検討する．メルケプストラム歪みは，時間構造が一致した 2 音声の特徴量ベクトル系列 ($X = [x_1, x_2, ..., x_T]^\top$, $Y = [y_1, y_2, ..., y_T]^\top$) に対し，式 (2) で定義される．

$$\text{Mel-CD [dB]} = \frac{1}{T}\sum_{t=1}^{T}\frac{10}{\ln 10}\sqrt{2||x_t - y_t||^2} \qquad (2)$$

3.1.2 実験条件

実験データには JNAS を用いた [6]．JNAS データ中の音素バランス文 503 文中の 50 文からなるサブセット A を読み上げている話者 32 人 (男女各 16 人ずつ) を対象に，男性話者 m001 を入力話者，残り 31 名の話者を出力話者として，31 通りの入力-出力話者ペアに対して実験を行った．学習データのサンプリング周波数は 16 kHz，フレームシフトは 5 ms とした．スペクトル特徴量として STRAIGHT 分析 による 0 次から 24 次のメルケプストラム係数を用いた [7]．各変換モデルの学習には，各話者の 50 文のデータのうち 1 ～ 40 文目を用いた．41 ～ 50 文目を評価データとし，話者距離の計算と変換精度の評価に用いた．変換モデルとしては 2 章の DNN に基づいた変換手法を用いた [4]．隠れ層数と素子数は全話者ペアに対して固定としてそれぞれ 5，128 とした．最適化手法として，学習率 0.001 の Adam を用いた [8]．学習データ 40 文のうち，33 ～ 40 文目の 8 文をバリデーションデータとし，モデルの学習はバリデーションデータに対する誤差が減少しなくなるまで反復を繰り返すことで行った．また学習データの前処理として，DTW (Dynamic Time Warping) と 大局的なアフィン変換による大まかな話者変換を繰り返し交互に 10 回適用することでパラレルデータの時間構造を一致させた．

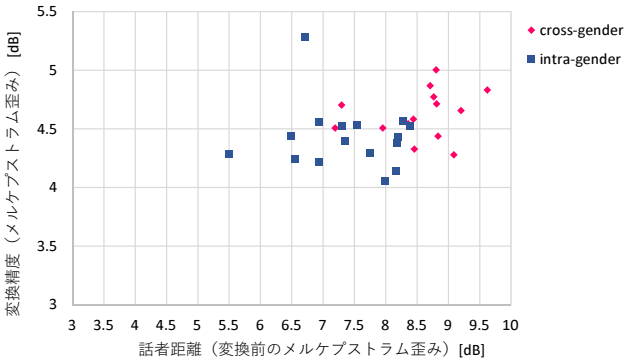

図 2　自然音声間のメルケプストラム歪みに基づく話者距離と変換精度の関係

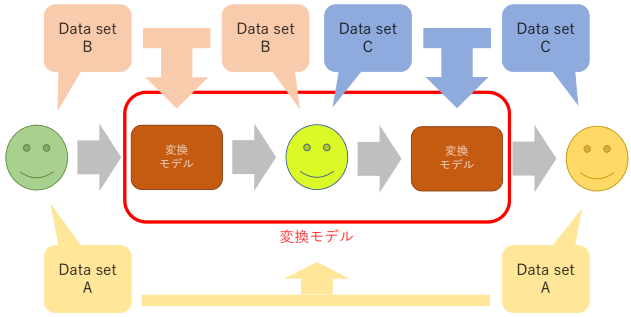

図 3　提案手法

3.1.3 結果

実験の結果得られた話者距離と変換精度の関係を図 2 に示す．相関係数は 0.23 であり， 話者距離と変換精度の間には係数は小さいが確実に正の相関があることが分かる．つまり，入力-出力音声間で声道スペクトルの距離が遠い程，変換精度が低くなる．また，話者変換の精度は異性間の方が同性間よりも低いことが図 2 より分かる．

4. 提案手法

4.1 話者空間内の変換経路を考慮した話者変換

3 章で分析した入力-出力話者間距離と変換精度の関係に基づき，話者空間内の変換経路を考慮した話者変換を提案する．3.1.3 節の結果は，異性間などの話者性が大きく異なる入力-出力話者間の変換を行う際に，入力-出力特徴量が複雑な対応関係を持つため，モデルが合理的に学習され

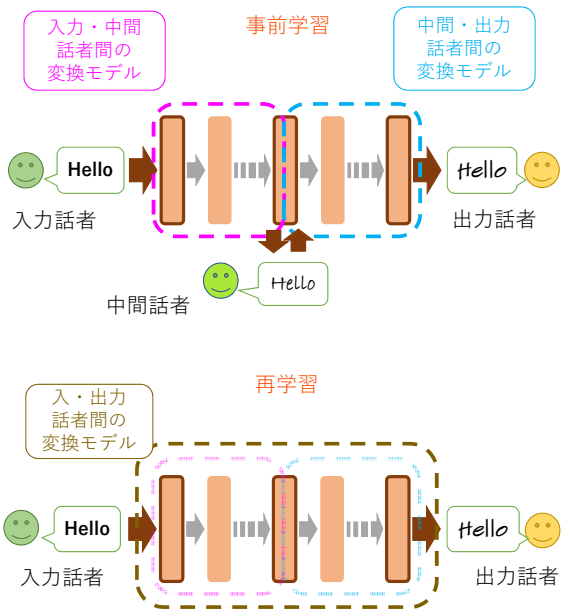

図 4　提案手法におけるモデル学習の流れ

ないことが一因と考えられる．我々はこの問題に対し，入力-出力話者と比較的話者性の近い第三話者（中間話者）を用意し，入力音声から出力音声への変換過程で中間話者の音声を経由するような変換モデルを構築することで，入力-出力話者間の変換精度を向上させることを検討する．ここで，一度入力音声から中間音声に変換しその後中間音声から出力音声へと単純に変換すると，入力-中間話者変換と中間-出力話者変換の誤差が累積することは容易に想像できる．そこで本研究では，DNN の事前学習として中間話者のデータを用いる手法を提案する．つまり，入力-出力話者間の距離が遠い場合に，入力-出力話者と比較的話者性の近い中間話者を用意し (図 3)，中間話者の特徴量をガイドとしながら入力特徴量を出力特徴量に変換するような変換モデルを構築することで，入力-出力話者間の変換精度を向上させることを試みる．具体的な手順は，図 4 に示すように，入力-中間話者間及び中間-出力話者間の DNN 変換モデルを各々のパラレルデータをもとに学習し，それら変換モデルを連結した入力-中間-出力話者モデルを，入力-出力話者間の事前学習モデルとする．この連結モデルを，入力-出力パラレルデータをもとに再学習を行うことで，入力-出力話者間の変換モデルを構築する．このようにモデルを構築することで，話者性が大きく異なる入力-出力話者間に対し入力音声から出力音声への変換に方向性を与え，より合理的な変換モデルの学習が期待できる．

4.2　実験条件

実験データは 3.1.2 節と同様に JNAS を用い，話者ペアの選び方のみが異なる．入力話者は男性話者 m001 とした．出力話者は，3 章の実験結果において m001 を入力話者とした際の変換精度が低い話者の中から，女性話者 f032 を出力話者とした．中間話者は各々 JNAS 内の subset A を読み上げているその他の話者 30 名とした．

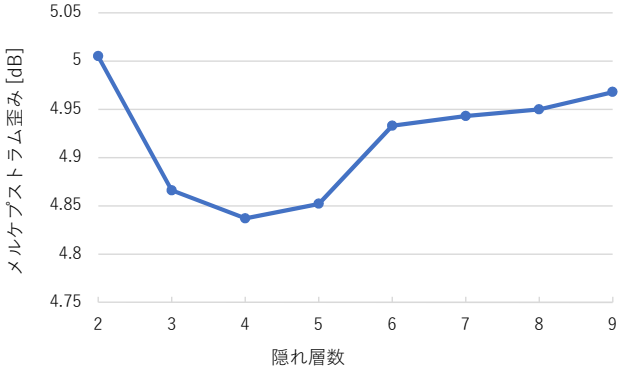

図 5　従来手法による話者変換精度

まず，この入力-出力話者ペア (m001-f032) に対し，従来の DNN 入力-出力話者変換手法による実験を行った．これは，従来手法で過学習を避けて変換モデルを学習することが可能なパラメータ数を調べる意図がある．本実験では，DNN の隠れ層数を 2 ～ 9 と変化させた場合の変換精度について実験を行った．その他の実験条件は 3.1.2 節と同様とした．

中間話者を経由する DNN 変換モデルのパラメータを以下に示す．入力-中間話者間の変換モデルについては，隠れ層数と素子数はそれぞれ 4, 128 とした．中間-出力話者間の変換モデルについても同様である．つまり，連結した入力-出力話者間の変換モデルの隠れ層数は 9 となる．その他の条件は 3.1.2 節と凡そ同様であるが，最適化手法として学習率 0.0001 の Adam を用いた点のみ異なる．

4.3　実験結果・考察

従来の DNN 入力-出力話者変換手法による実験の結果を図 5 に示す．図 5 より隠れ層数が 4 の時，変換精度が 4.837 [dB] となり最も良く，その後層数を増やすにしたがって変換精度が悪くなっている，つまり過学習が起きていることが分かる．これに対し，中間話者を経由した変換手法の実験結果 (図 6) において，従来手法の変換精度 (4.837 [dB]) を上回る結果が得られた．提案手法の変換モデルの隠れ層数 (9 層) は，従来手法 (図 5) では過学習が起きている層数である．つまり，提案手法は中間話者のデータを利用した事前学習により過学習を回避したと言え，提案手法の有効性が示された．

また，図 6 より，変化精度が中間話者の選択に依存していることが分かる．この結果に対し変換精度と入力-中間話者距離，中間-出力話者距離，及びその和を比較したが，十分に解釈可能な結果は得られなかった．また，他の入力-出力話者ペアに対し提案手法を適用した場合，選択した中間話者によっては必ずしも変換精度が向上しない入力-出力話者ペアが存在した．これらの結果から，本研究では 2

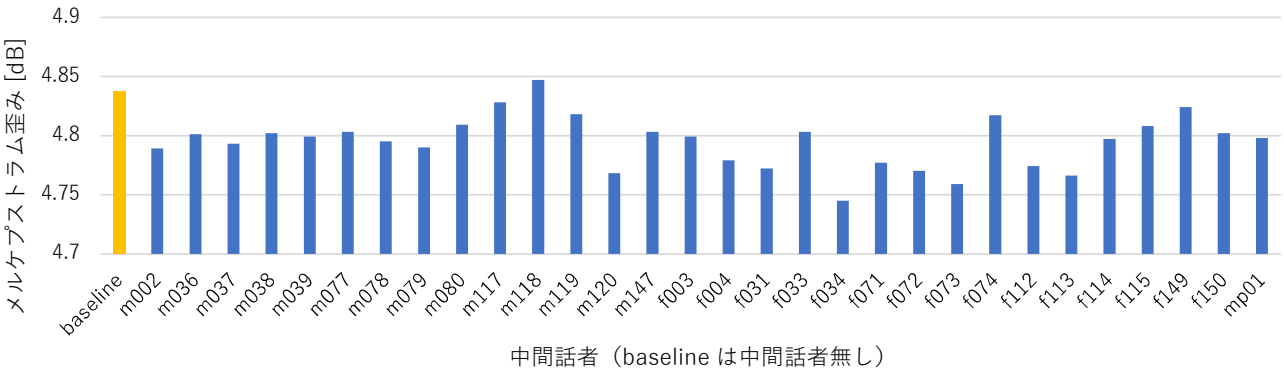

図 6 選択した中間話者ごとの変換精度 (入力話者は m001，出力話者は f032)

話者間の関係をメルケプストラム歪みを用いた 1 次元の話者空間で表現しているが，これは話者間の関係性を記述するには不十分だと考えられる．特に，3 章の議論より 2 話者間では自然音声間のメルケプストラム歪みに基づく話者距離と変換精度の間に一定の相関が得られたが，本章の実験結果を鑑みるに中間話者を加えた 3 話者間の関係性は捉えていないと言える．

5. まとめと今後の課題

本研究では，従来の話者変換手法において未だアプローチが十分ではない，入力-出力話者間の話者距離と変換精度の間にある関係を分析することを試み，話者空間内の変換経路を考慮した話者変換を検討した．特に入力-出力話者間の距離が遠い場合に，第三話者（中間話者）のデータを用い，入力音声から出力音声への変換過程で中間話者の音声をガイドとする変換モデルの構築を提案した．実験の結果，メルケプストラム歪みを用いた客観評価において提案手法の有効性が示された．

本研究では自然音声間のメルケプストラム歪みに基づく話者距離で話者空間を構築したが，実験の結果，2 話者間のメルケプストラム歪みに基づく話者距離を用いることでは決定論的に中間話者を選択することが出来ないことが分かった．今後の課題として，中間話者の選択と話者変換精度の関係性をより明らかにするために，Eigenvoice などの話者空間を構築し更なる分析を行う [9]．

参考文献

[1] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara: Voice conversion through vector quantization, *Journal of the Acoustical Society of Japan (E)*, Vol. 11, No. 2, pp. 71–76 (1990).
[2] A. Kain and M. W. Macon: Spectral voice conversion for text-to-speech synthesis, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 285–288 (1998).
[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] S. Desai, E. V. Reghavendra, B. Yegnanarayana, A. W. Black and K. Prahalled: Voice conversion using artificial neural networks, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 3893–3896 (2009).
[5] M. Wester, Z. Wu and J. Yamagishi: Analysis of the Voice Conversion Challenge 2016 Evaluation Results, pp. 1637–1641 (2016).
[6] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano and S. Itahashi: JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research, Vol. 20, No. 3, pp. 199–206 (1999).
[7] H. Kawahara, I. Masuda-Katsuse and A. De Cheveigne: Restructuring speech representations using a pitch-adaptive time—frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds, Vol. 27, No. 3, pp. 187–207 (1999).
[8] D. Kingma, J. Ba: Adam: A Method for Stochastic Optimization, *arXiv:1412.6980 [cs.LG]* (2009).
[9] T. Toda, Y. Ohtani, and K. Shikano: Eigenvoice conversion based on Gaussian mixture model, *Proceedings of the International Conference on Spoken Language Processing*, pp. 2446–2449 (2006).