Vol.2017-SLP-117 No.1
2017/7/27

DNNを用いた時変線型変換とその音声変換への応用

小谷 岳[1,a)]　齋藤 大輔[1,b)]　峯松 信明[1,c)]

概要：音声や画像といったメディア情報分野において，特徴量空間を混合ガウス分布 (Gaussian mixture model; GMM) でモデル化する手法は広く用いられてきた．GMM に基づいたメディア変換を行う際，GMM の各要素分布により特徴量空間を領域分割し局所線型性に基づいた変換が行われる．これに対し近年，複雑な対応関係を持つ特徴量間マッピングをディープニューラルネットワーク (Deep neural networks; DNNs) によりモデル化する研究が盛んに行われている．しかし，従来の DNN に基づくマッピング手法は非常に柔軟性が高い一方で，入力-出力特徴量の変換過程で非線型な特徴量変換が繰り返し適用されるために，同一ドメイン内変換の場合においては，そのマッピング関数は非現実な対応関係を学習しうると考えられる．本研究では，この問題に対し，入力-出力特徴量変換が同一ドメイン内の変換であるという制約を効果的にモデル化することを検討する．具体的には，DNN による特徴量間の変換過程に対し時変線型変換という制約を設けることで，より合理的にモデルを学習する新しい DNN アーキテクチャを提案する．また，声質変換というタスクにおいて実験的にその性能を評価し，従来の DNN に基づく手法と比べ，より自然な音声変換が実現できることを示す．

キーワード：声質変換，Deep Learning，時変線型変換

Time-variant linear transformation using deep neural networks and its application to voice conversion

Gaku Kotani[1,a)]　Daisuke Saito[1,b)]　Nobuaki Minematsu[1,c)]

1. はじめに

音声や画像といったメディア情報分野において，特徴量空間を GMM でモデル化する手法は広く用いられてきた [1], [2], [3], [4]．GMM に基づいたメディア変換を行う際には，GMM の各要素分布により特徴量空間を領域分割し局所線型性に基づいた変換が行われる [1], [2]．GMM に基づく特徴量変換は，入力特徴量が与えられた場合の出力特徴量の条件付き確率に基づいて行われる．この際，局所的には入力特徴量が属する GMM の要素分布の識別とその要素分布に従った線型変換が行われていると解釈できる．一方で近年，入力-出力特徴量間のマッピングに DNN を用いる研究が盛んに行われている [5], [6], [7], [8]．音声認識や画像認識といった異なるドメイン間の特徴量の対応関係を DNN で記述することにより，精度の向上が見られている [6], [8]．声質変換のような同一ドメイン内における変換に対しても，DNN に基づく変換モデルは研究されており，訓練データ量が多い条件下では GMM に基づく手法を上回る変換精度が得られている [5]． DNN に基づく手法は GMM に基づく手法と比べ，入力-出力特徴量間のより複雑な対応関係を記述することが出来ると考えられる．しかし，従来の DNN に基づくマッピング手法は非常に柔軟性が高い一方で，入力-出力特徴量変換が同一ドメイン内の変換であるという制約を効果的にモデル化する試みは少ない．従来の DNN に基づく変換モデルは入力特徴量から出力特徴量への変換過程において，非線型な特徴量変換が繰り返し適用されるために，同一ドメイン内での変換のように，変換の解釈がある程度明確である場合には合理的では

[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

ない．我々は，特にメディア変換における入力-出力特徴量が同一ドメイン内にある場合に，DNN による特徴量間の変換過程に対し時変線型変換という制約を設けることで，より合理的にモデルを学習する新しい DNN アーキテクチャを提案する．また，声質変換というタスクにおいて実験的にその性能を評価し，従来の DNN に基づく手法と比べ，より自然な音声変換が実現できることを示し，さらに学習した DNN 変換モデルの機能に対し分析的な検討を行う．

2. GMM を用いた従来の話者変換

本章では，結合確率密度 GMM を用いた従来の話者変換手法について簡単に説明する [1]．入力-出力話者のパラレルデータに対し，時刻 t における入出力話者の D 次元特徴量ベクトルをそれぞれ $\boldsymbol{x}_t = [x_1, x_2, ..., x_D]^\top$，$\boldsymbol{y}_t = [y_1, y_2, ..., y_D]^\top$ と表す．このとき，結合ベクトル $\boldsymbol{z}_t = [\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top]^\top$ を GMM を用いて式 (1) のようにモデル化する．

$$P(\boldsymbol{z}_t|\boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}). \tag{1}$$

ここで，w_m と $\boldsymbol{\mu}_m^{(z)}$，$\boldsymbol{\Sigma}_m^{(z)}$ はそれぞれ GMM の m 番目の要素分布に対する重みと平均ベクトル，分散行列にあたる．また，$\boldsymbol{\mu}_m^{(z)}$ と $\boldsymbol{\Sigma}_m^{(z)}$ は入出力話者の特徴量ベクトルに対する平均及び分散行列を用いて，それぞれ式 (2) のように表される．

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}. \tag{2}$$

結合確率密度 GMM を用いた変換モデルにおいて，入力特徴量 $\boldsymbol{x}_t$ を出力特徴量 $\boldsymbol{y}_t$ に変換するマッピング関数 $\mathcal{F}(\cdot)$ は，$\boldsymbol{x}_t$ が与えられた場合の $\boldsymbol{y}_t$ の条件付き確率に基づく．この条件付き確率は結合確率密度 GMM のパラメータを用いて表すことができ，最小二乗誤差基準で学習されたマッピング関数 $\mathcal{F}(\cdot)$ は式 (3) のように表される．

$$\mathcal{F}(\boldsymbol{x}_t) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)})(\boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)})). \tag{3}$$

式 (3) において，初項 $P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)})$ は時刻 t における入力特徴量 $\boldsymbol{x}_t$ を GMM の特定の要素分布に割り当てる役割をし，残りの第二項は各要素分布に対応する線型変換を行っていると解釈することが出来る．言い換えると，マッピング関数 $\mathcal{F}(\cdot)$ は，領域分割による局所線型変換，つまり時変線型変換として表すことが出来る．ただし，式 (3) による変換は，各混合の重み付け和により計算されるため，最終的な変換は離散的ではなく連続的な変換となる．

3. 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 [5]．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 $\boldsymbol{h}^{(l)}$ 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型変換と活性化関数 $g(x)$ の組み合わせによって以下のように表される．

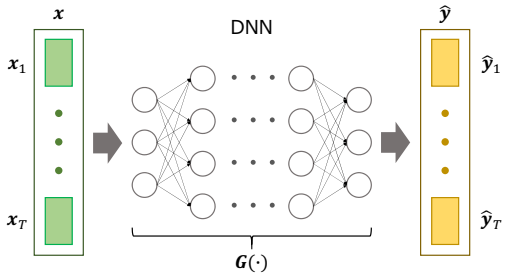


図 1　DNN を用いた従来の話者変換
Fig. 1　Traditional DNN-based VC

$$\boldsymbol{h}^{(l)} = g(\boldsymbol{W}^{(l)} \boldsymbol{h}^{(l-1)} + \boldsymbol{b}^{(l)}). \tag{4}$$

活性化関数 $g(x)$ としては，シグモイド関数や $g(x) = tanh(x)$（双曲線関数），$g(x) = max(x, 0)$（Rectified Linear Unit; ReLU）などが用いられ，本研究では ReLU を用いる．最終層の活性化関数については，声質変換のような連続値に対する回帰問題では線型写像が広く持ちいられており，本研究についても同様である．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を表す特徴量（メルケプストラム）を入出力音声特徴量として用い，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．

従来手法では図 1 に示すように，入力音声特徴量系列 $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_T^\top]^\top$ から，DNN を用いて変換音声の特徴量系列 $\hat{\boldsymbol{y}} = \boldsymbol{G}(\boldsymbol{x})$ を推定し，音声波形を合成する [2]．まず，別途推定した音源特徴量系列を用いて混合励信源波形を合成する．これに対して，$\hat{\boldsymbol{y}}$ によるフィルタを適用することで最終的な変換音声を得る．また，音声波形を合成する他の手法として，差分スペクトル推定に基づく手法がある [9]．差分スペクトル推定に基づく手法では，入力音声に対し，入力-出力声道スペクトル特徴量系列の差分によるフィルタを適用することで，最終的な変換音声を得る．差分スペクトル推定に基づく手法により，ボコーダ処理による音質劣化を回避することができるが，F_0 や非周期成分といった特徴量の変換は困難となる．本研究では，声道スペクトル特徴量系列の差分 $\hat{\boldsymbol{y}} - \boldsymbol{x}$ から得られるフィルタを用いて，差分スペクトル推定に基づく手法により変換音声を得る．また，本研究では実験において異性間の変換も扱うため，F_0 の変換を行う必要性がある．予備実験において，ボコーダ処理による音声波形合成手法を用いて F_0

のみを変換した場合の音質劣化は比較的小さいことを確認したため，本研究ではまずボコーダ処理により F_0 のみを変換した変換音声を作り，その後差分スペクトル推定に基づく手法により声道スペクトル特徴量を変換する．F_0 の変換は次式で定義される線型変換を行った．

$$\hat{y}_t = \frac{\sigma^{(y)}}{\sigma^{(x)}}(x_t - \mu^{(x)}) + \mu^{(y)} \quad (5)$$

ここで，$\hat{y}_t$ と x_t はそれぞれ時刻 t における対数化した F_0 である．$\mu^{(x)}$,$\mu^{(y)}$ は学習データから求めた入出力話者それぞれの対数化した F_0 の平均であり，$\sigma^{(x)}$,$\sigma^{(y)}$ は学習データから求めた入出力話者それぞれの対数化した F_0 の標準偏差である．

DNN に基づく従来の話者変換手法では，入力特徴量から出力特徴量への変換過程において，式 (4) で表される非線型な特徴量変換が繰り返し適用される．テキストから音声への合成，音声からテキストへの認識など，異なるドメイン間の対応付けを学習する際には，DNN を用いた手法は優れた性能を発揮している [6], [7], [8]．しかし，声質変換のような同一ドメイン内における変換では，DNN を用いた従来手法はその性能を十分に発揮しているとは言えない．同一ドメイン内の特徴量変換では，特徴量変換の意味付けが明確であることが多く，このトップダウンの知識を活用することが望ましい．入力-出力特徴量変換が同一ドメイン内の変換であることを明示的に表現した DNN アーキテクチャとして，Residual Networks (ResNet) を考えることが出来る [8]．ResNet は，式 (6) で表されるように，入力-出力特徴量間の差分を学習する．

$$\hat{\boldsymbol{y}}_t = \boldsymbol{x}_t + \boldsymbol{R}(\boldsymbol{x}_t). \quad (6)$$

しかし，声質変換というタスクにおいて，ResNet はケプストラムドメイン内における変換という制約を十分には活かせていない．このことを 4 章において説明し，実験によって提案手法が ResNet を用いた変換手法よりも高い変換精度を実現できることを示す．

4. 提案手法

4.1 DNN を用いた時変線型変換

本節では，提案手法として DNN を用いた時変線型変換について述べる．提案手法の DNN アーキテクチャを図 2 に示す．提案する DNN アーキテクチャは 2 つのサブネットワークとその結合部で構成される．2 つのサブネットワークでは，入力特徴量 $\boldsymbol{x}_t$ からそれぞれ変換行列 $\boldsymbol{A}(\boldsymbol{x}_t)$ とバイアス項 $\boldsymbol{b}(\boldsymbol{x}_t)$ が推定される．推定されたパラメータを用いて，入力特徴量 $\boldsymbol{x}_t$ から出力特徴量 $\boldsymbol{y}_t$ への変換を行う（式 (7)）．この時，変換パラメータ $\boldsymbol{A}(\boldsymbol{x}_t)$ 及び $\boldsymbol{b}(\boldsymbol{x}_t)$ は各時刻 t において変化し，DNN を用いた時変線型変換を実現する．

$$\hat{\boldsymbol{y}}_t = \boldsymbol{A}(\boldsymbol{x}_t)\boldsymbol{x}_t + \boldsymbol{b}(\boldsymbol{x}_t). \quad (7)$$

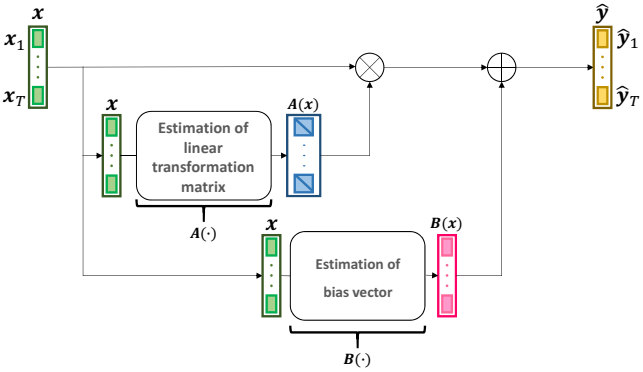
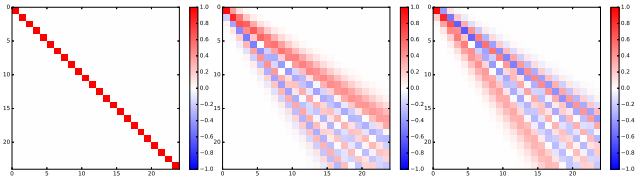


図 2 DNN を用いた時変線型変換モデル
Fig. 2 Proposed framework of DNN-based time-variant linear conversion



(a) α=0.0 (b) α=0.2 (c) α=−0.2
図 3 [10] における線型変換行列 $\boldsymbol{A}$ を可視化した例．
Fig. 3 Visualization of several examples of matrix $\boldsymbol{A}$ in [10].

DNN のパラメータの学習は，従来の DNN を用いた手法と同様に，変換特徴量 $\hat{\boldsymbol{y}}_t$ と出力特徴量 $\boldsymbol{y}_t$ 間の二乗誤差を最小化する基準で行われる．また，提案手法の変換行列 $\boldsymbol{A}(\boldsymbol{x}_t)$ として常に単位行列を用いた場合，提案した DNN アーキテクチャは ResNet と等価である（3 章）．入力-出力特徴量間の対応関係がケプストラム空間上での回転成分を持たない場合，ResNet を用いたモデル化は提案手法に近い性能を有すると考えられる．

提案手法は，入力-出力特徴量間の変換過程を時変線型変換に制約することで，非線型な特徴量変換を繰り返す従来手法と比べ，同一ドメイン内の変換に対しより効果的な学習が期待出来る．2 章で GMM に基づいた変換手法が時変線型変換と解釈できることについて述べたが，GMM に基づいた時変線型変換は変換の自由度として $\boldsymbol{\Sigma}_m^{(yx)}\boldsymbol{\Sigma}_m^{(xx)-1}$ の重み付け和しか許されていない．提案手法の DNN を用いた時変線型変換は，より効果的で柔軟な同一ドメイン内における変換を実現できると考えられる．

提案手法がケプストラムドメインにおける線型変換を実現すると，その変換行列 $\boldsymbol{A}(\boldsymbol{x}_t)$ は少なくとも，入力-出力話者間の声道長の違いを反映すると考えられる．これについて，次節でより詳細に説明する．

4.2 Vocal tract length normalization

本節では，提案手法による変換モデルが，話者間の声道長変換を明示的に表現することについて述べる．話者間の声道長の違いは，話者間の声質の違いの一要素として広く知られている．音声認識において，話者間の声道長の違い

は著しく認識精度を低下させる要因の一つであり，声道長正規化 (Vocal Tract Length Normalization; VTLN) による話者正規化技術が広く用いられている [11]．以下では，[10] で報告されている声道長変化の定式化とケプストラムの声道長依存性について述べる．

話者の声道長の単調な変化は，音声のスペクトル表現における周波数ウォーピングとして考えることができる．今，周波数ウォーピングにおける変換前後の正規化角周波数を $\omega, \hat{\omega} (0 \leq \omega, \hat{\omega} \leq \pi)$ とする．このとき $z = e^{j\omega}$，$\hat{z} = e^{j\hat{\omega}}$ とし，周波数ウォーピングとして以下の 1 次全域通過関数を考える．

$$\hat{z}^{-1} = m(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \quad (-1 < \alpha < 1). \tag{8}$$

$\alpha < 0$ の場合，周波数軸が低域に変換され声道長は長くなる．一方 $\alpha > 0$ の場合，周波数軸は高域に変換され声道長は短くなる．

以下，前述のスペクトルドメインにおける周波数ウォーピングをケプストラム空間における記述に置き換える．パワーを表現するケプストラムの 0 次項を考慮しない場合，周波数ウォーピングは以下の式でケプストラム空間における線型変換として表現される．

$$\hat{\boldsymbol{c}} = \boldsymbol{A}\boldsymbol{c}, \tag{9}$$

$$\hat{\boldsymbol{c}} = (\hat{c}_1\ \hat{c}_2\ \hat{c}_3\ \hat{c}_4\ \cdots)^\top, \tag{10}$$

$$\boldsymbol{A} = \begin{pmatrix} 1-\alpha^2 & 2\alpha - 2\alpha^3 & \cdots & \cdots \\ -\alpha + \alpha^3 & 1 - 4\alpha^2 + 3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}, \tag{11}$$

$$\boldsymbol{c} = (c_1\ c_2\ c_3\ c_4\ \cdots)^\top. \tag{12}$$

また，[12] では，式 (11) の変換行列が強い回転性を持ち，その性質は音韻による影響が見られることを実験的に示している．

本研究で提案する DNN アーキテクチャには，入力-出力特徴量（メルケプストラム）間の変換に時変線型変換という制約を設けている．話者間の声道長変換がケプストラム空間での線型変換で表現できることに着目すると，提案手法による変換モデルの変換行列は，少なくとも式 (11) の変換行列を内包し，さらにその音韻による変化を捉えていると解釈できる．ここで，提案手法における $\boldsymbol{A}(\boldsymbol{x}_t)$（4.1 節の式（7））は一般行列であり，式 (11) のパラメータ α から求まる行列 $\boldsymbol{A}$ と同一ではないことには注意が必要である．

5. 実験

5.1 実験条件

提案手法の性能を評価するために，3 章で議論した DNN を用いた従来手法及び ResNet を用いた手法と比較実験を行った．実験データには JNAS を用いた [13]．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の中から，同性話者ペアと異性話者ペアを一組ずつ実験に用いた．実験に用いた話者ペアを表 1 に示す．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 分析に基づいた 0 次から 24 次のメルケプストラム係数を用いた [14]．メルケプストラム係数の 1 ∼ 24 次を DNN の入出力特徴量とし，パワーを表す 0 次項については入力音声のものをそのまま用いた．

表 1　実験に用いた話者ペア
Table 1　Speaker pairs used for experiments

Speaker pairs	Input speaker	Output speaker
male to male	m002	m080
male to female	m038	f071

各変換モデルの学習には，各話者 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，変換精度の評価に用いた．実験に用いた手法のモデルパラメータについて説明する．まず，全ての手法について，隠れ層の素子数は 128 とした．隠れ層数は，3 章の DNN に基づいた変換手法と ResNet を用いた手法，提案手法の変換行列を推定するサブネットワーク及びバイアス項を推定するサブネットワークについてそれぞれ 5, 4, 6, 3 とした．最適化手法として，学習率 0.0005 の Adam を用いた [15]．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．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．また，メルケプストラム特徴量の前処理としてカットオフ周波数 50 Hz のトラジェクトリスムージングを施した [16].

評価指標として，メルケプストラム歪みに基づく客観評価と，変換音声の自然性と話者性に関して AB テスト及び ABX テストによる主観評価を用いた．客観評価は，評価データ 10 文に対する変換音声と出力音声間のメルケプストラム歪みの平均を用いた．主観評価は，変換音声の自然性を評価するために，3 手法間で 3 通りの AB テストを行った．また，変換音声の話者性を評価するために，3 手法間で 3 通りの ABX テストを行った． 主観評価の被験者数は 11 名であり，評価データ 10 文を自然性と話者性の評価に用いた．

5.2 実験結果・考察

実験の客観評価結果を表 2 に，主観評価結果を図 4 に示す．まず，DNN を用いた従来の変換手法（Baseline）と

表 2　メルケプストラム歪みによる客観評価結果 [dB]
Table 2　Results of objective evaluations by mel-cepstral distortion [dB]

Speaker pairs	Baseline	Residual	Proposed
male to male	4.503	4.631	4.561
male to female	4.262	4.369	4.334

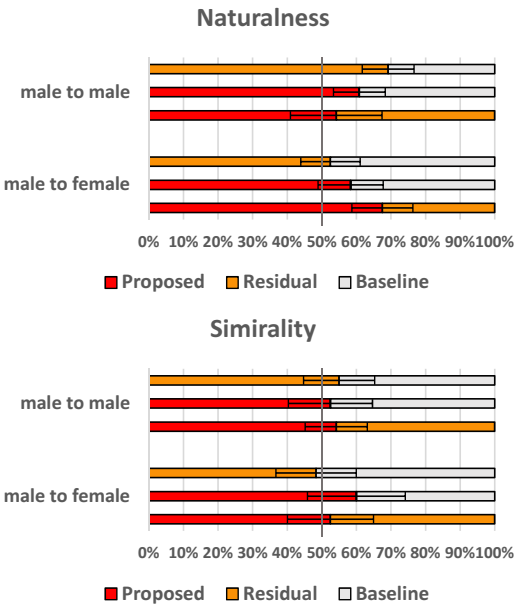
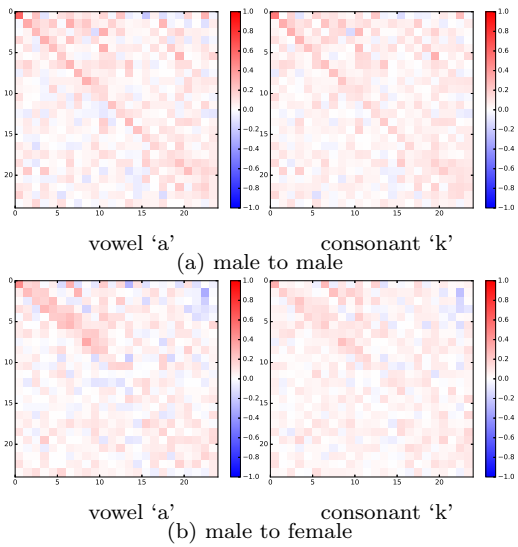


図 4　主観評価結果
Fig. 4　Results of subjective evaluations



(a) male to male
(b) male to female

図 5　提案手法による変換行列 $\boldsymbol{A}(\boldsymbol{x}_t)$ を，各音韻（モノフォン）についての時間平均により可視化した例.
Fig. 5　Visualization of results of $\boldsymbol{A}$, which is the averaged $\boldsymbol{A}(\boldsymbol{x}_t)$ along the time axis for each phoneme in the test sentences.

提案手法（Proposed）を比較すると，客観評価においてはBaseline の方が若干精度が良い．しかし，音声の自然性に関する主観評価においては同性間及び異性間変換で共にProposed の方が変換精度が良い．これは，3, 4 章で議論したように，Baseline では非現実的な特徴量マッピングをモデルが学習していることに起因すると考えられる．音声の話者性に関する主観評価においても，自然性のために話者性を損なうといったことなく変換が行えており，妥当な結果が得られている．提案手法は，時変線型変換という同一ドメイン内変換の明示的な制約を設けることで，より自然な音声変換を実現している．また，ResNet を用いた変換手法（Residual）と提案手法（Proposed）を比較すると，客観評価においては提案手法の方が若干精度が良い．これは，提案手法の変換行列が同性間及び異性間変換の両方で機能していることを示している．音声の自然性に関する主観評価においては同性間と異性間で結果が異なる．同性間変換の場合，Proposed は Residual の変換精度を上回ってはいるがその差は小さい．しかし，異性間変換の場合は Proposed は Residual の変換精度を大きく上回っている．これは，同性間変換では入力-出力話者間の声道長の違いが小さい，つまり声道長変換を表す変換行列が単位行列に近くなる（図 3 (a)）ために，入力-出力特徴量間の変換は回転成分が小さく，ResNet を用いることでもその対応関係を学習出来ていることを示している．しかし，異性間変換においては声道長の違いが大きい，つまり声道長変換を表す変換行列は帯行列に近くなる（図 3 (b)）ために，入力-出力特徴量間の変換は回転成分が大きく，ResNet を用いた変換モデルはその対応関係を効果的に学習できていないことを示している．音声の話者性に関する主観評価結果については，Baseline と Proposed を比較した時と同様，妥当な結果が得られている．実験結果から，提案手法はケプストラムドメインにおける柔軟な変換を実現していることが分かる．

さらに，提案手法がケプストラムドメイン内の変換を合理的に実現していることを確認するために，提案手法における変換行列 $\boldsymbol{A}(\boldsymbol{x}_t)$ の可視化を行った（図 5）．図 5 は，提案手法における変換行列 $\boldsymbol{A}(\boldsymbol{x}_t)$ に対し，評価データ内の各音素（モノフォン）毎に時間平均を取った行列の例（母音 ‘a’, 子音 ‘k’）を可視化した．音素ラベルは Julius を用いた強制アラインメントによって得た [17]．図 5 を見ると，音素によって変換行列の様子が異なることが分かる．まず，共鳴音である母音 ‘a’ について同性間および異性間の変換行列を見ると，入力-出力話者間の声道長の違いに起因した違いが見て取れる．母音 ‘a’ の場合，大きい正の値を持つ行列成分が対角要素の近くに集まっていることが分かるが，同性間変換では異性間変換の場合よりも，対角成分がはっきりと見える．この様子は，声道長変換を表す理想的な変換行列 (図 3 (a)(b)) と比べることで，声道長の違いに起因していることが分かる．また，子音 ‘k’ について見てみると，母音 ‘a’ の場合と比べて声道長の違いによる影響が小さくなっていることが見て取れる．つまり，提案手法はケプストラムドメイン内における時変線型変換を実現し

ていると言える．

6. まとめと今後の課題

本論文では，同一ドメイン内における合理的な変換を実現する新しい DNN アーキテクチャを提案した．具体的には，入力-出力特徴量変換が同一ドメイン内の変換である場合に，その変換過程を時変線型変換に制約することを提案した．声質変換というタスクにおいて，提案した DNN を用いた時変線型変換を適用することで，従来の DNN を用いた変換手法を上回る変換精度が得られることを実験的に示した．また，提案手法において学習されたモデルの線型変換行列を可視化することで，提案手法の機能に対して分析的な検討を行った．

提案手法は，同一ドメイン内における変換に対して，そのドメインの知識を効果的に導入した一例と解釈でき，その適用範囲は広い．今後の展望の一つとして，提案手法を Recurrent neural networks といった時系列モデルと組み合わせることが挙げられる．提案手法に対し複数の時間フレームを用いることで，変換精度の更なる向上が見込まれる．

参考文献

[1] 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).

[2] T. Toda, A. Black and K. Tokuda: Voice conversion based on maximum likelihood estimation of spectral parameter trajectory, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2222–22352 (2007).

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet: Front-End Factor Analysis for Speaker Verification, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 4, pp. 788–798 (2011).

[4] D. Lee: Effective Gaussian Mixture Learning for Video Background Subtraction, *IEEE Transactions on Pattern Aanalysis and Machine Intelligence*, Vol. 27, No. 5, pp. 827–832 (2005).

[5] S. Desai, E. Reghavendra, B. Yegnanarayana, A. Black and K. Prahalled: Voice conversion using artificial neural networks, *in Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 3893–3896 (2009).

[6] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel and Y. Bengio: End-to-end attention-based large vocabulary speech recognition, *in Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 4945–4949 (2016).

[7] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu: WaveNet: A Generative Model for Raw Audio, *arXiv:1609.03499* (2016).

[8] K. He, X. Zhang, S. Ren and J. Sun: Deep Residual Learning for Image Recognition, *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016).

[9] S. Takamichi, T. Toda, A. Black, G. Neubig, S. Sakti and S. Nakamura: Postfilters to modify the modulation spectrum for statistical parametric speech synthesis, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 24, No. 4, pp. 755–767 (2016).

[10] M. Pitz and H. Ney: Vocal tract length normalization equals linear transformation in cepstral space, *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 5, pp. 930–944 (2005).

[11] E. Eid and H. Gish: A parametric approach to vocal tract length normalization, *in Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 346–348 (1996).

[12] D. Saito, R. Matsuura, S. Asakawa, N. Minematsu and K. Hirose: Directional dependency of cepstrum on vocal tract length, *in Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 4485–4488 (2008).

[13] 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, *Journal of the Acoustical Society of Japan (E)*, Vol. 20, No. 3, pp. 199–206 (1999).

[14] H. Kawahara, I. Masuda-Katsuse and A. 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, *Speech communication*, Vol. 27, No. 3, pp. 187–207 (1999).

[15] D. Kingma, J. Ba: Adam: A Method for Stochastic Optimization, *arXiv:1412.6980 [cs.LG]* (2009).

[16] K. Kobayashi, S. Takamichi, S. Nakamura and T. Toda: The NU-NAIST Voice Conversion System for the Voice Conversion Challenge 2016, *in Proceedings of INTERSPEECH*, pp. 1667–1671 (2016).

[17] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro and K. Shikano: Free Software Toolkit for Japanese Large Vocabulary Continuous Speech Recognition, *in Proceedings of the International Conference on Spoken Language Processing*, Vol. 4, pp. 476–479 (2000).