Vol.2014-SLP-104 No.3
2014/12/15

話者特徴量入力を付加したデノイジングオートエンコーダによるクロスリンガル声質変換

伊藤 洋二郎[1]　篠崎 隆宏[1,a)]　能勢 隆[2]

概要：数発話程度のごく少量のラベルなし音声を用いて特定話者の任意の発話を任意話者の声質に変換することを目的として，音声特徴量を音声特徴量に変換するデノイジングオートエンコーダに話者特徴量入力を付加した構造を持つニューラルネットを用いた声質変換手法を提案する．多言語音声コーパスを用いた実験により，提案法の有効性を示す．

1. はじめに

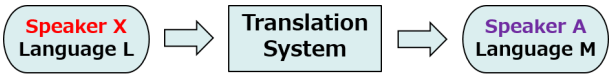
近年自動音声翻訳システムが普及しつつある．これらのシステムでは不特定多数のユーザーの発声を不特定話者音声認識技術によりテキストに変換し，テキストレベルで翻訳を行った後にシステムに備え付けられた特定の話者の声質で合成音声を合成し出力する構成が一般的である．したがって，ユーザーの声質と出力される合成音の声質は別人のものとなってしまう．より自然な音声コミュニケーションを実現するためには，ユーザーの声質で合成音が出力されることが必要である．特定のユーザーのみを対象として音声自動翻訳システムを動作させる場合でもし予めそのユーザーの音声データが学習データとして利用できるときには，ユーザー依存の音声合成器を作成することでそのユーザーに合わせた声質の音声を合成できる．しかし，このアプローチでは事前に翻訳先言語でのユーザーの音声モデルを用意することが前提となるため，不特定のユーザーを対象とした実時間の音声自動翻訳には適用できない．

そのようなシステムを対象としてユーザーの声質で翻訳先言語の音声を得るための方法として，戸田らは GMM ベースの声質変換器と Eigen Voice 話者適応化技術を組み合わせた一対多声質変換手法を提案している [1]．これは，特定の声質で合成音声を出力する音声自動翻訳器の後段として声質変換器を導入するアプローチである．Eigen Voice 話者適応化は本来であれば翻訳先言語でのユーザー音声を必要とするものであるが，圧縮された空間でのモデル推定となるため翻訳先言語のモデルをユーザーの母語側音声を用いて適応化することでも有効な適応化が期待できる．実際に戸田らは，少量の入力発話を用いた入力言語側の音声によるラベル無し適応化により翻訳先言語での音声の声質がユーザーの声質に近づくことを示している．

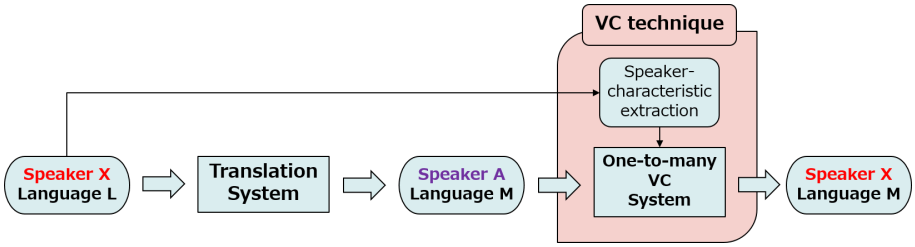
GMM ベースの声質変換器と Eigen Voice 話者適応化を組み合わせた手法は同様に多対一声質変換にも利用できる．Desai らはニューラルネットを用いた多対一声質変換手法を提案し GMM を用いた手法との比較を行い，ニューラルネットを用いた手法が有望であることを示している [2], [3]．しかし，ニューラルネットのパラメタを少量のデータを用いて適応化することは難しいこのこともあり，少量のラベル無しユーザー音声をもとに一対多の声質変換をニューラルネットを用いて実現する方法はこれまで存在していない．

他方で音声認識の分野では近年ニューラルネットに対する話者適応化が精力的に研究され，様々な手法が提案されている．それらのうちの一つに，話者コードを用いた手法がある [4]．これはニューラルネットに含まれる大量のパラメタを利用時に変更する代わりに，話者の特性を表すコードを入力データの一部としてニューラルネットに入力することで，話者適応を実現しようとするものである．ニューラルネット自体は大量の多数話者データを用いて不特定話者モデルとして学習される一方で話者に依存した特性を話者コードとして取り込むことから，極少量のデータを用いた話者適応が可能となる．本論文ではこの話者コードを用いた話者適応化技術を応用した，ニューラルネットによる一対多声質変換器の提案を行う．

[1] 東京工業大学 大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
[2] 東北大学 工学研究科
School of Engineering, Tohoku University
[a)] www.ts.ip.titech.ac.jp



(a) Conventional system



(b) Combination with voice conversion

図 1 声質変換を取り入れた音声翻訳システム

2. デノイジングオートエンコーダ (DAE) によるクロスリンガル声質変換

不特定のユーザーを対象とした既存の音声自動翻訳システムの一般的な構成を図 1(a) に示す．この構成では，システムから音声合成により出力される言語 M に翻訳された音声の声質はユーザー X のものではなく，システムの音声合成器にあらかじめ備えられた特定の話者 A のものとなる．それに対して同図 (b) の様に音声自動翻訳システムの後段に声質変換システムを接続し，ユーザー X の音声入力から抽出した話者性の情報を用いることで音声翻訳システムから出力される話者 A の声質の合成音を瞬時に話者 X の声質に変換できれば，全体としてユーザー X の声質で翻訳された音声が出力されるシステムを構成できる．このようなシステムを実現するためにはシステム利用時において，一対多声質変換をパラレルデータを使用せずに極少量のラベルなし目標話者データを用いてクロスリンガルで動作させる必要がある．ただし，システムの構築時にパラレルデータを用いることは可能である．以下では，提案するニューラルネットによる一対多声質変換器について説明する．ニューラルネットはデノイジングオートエンコーダ (Denoising Auto Encoder:DAE) に話者特徴量入力を付加した構成となっており，上記の要請を満たすことが可能である．

2.1 DAE を用いた声質変換

DAE はあるドメインでの雑音の重畳した信号を同じドメインのクリーンな信号に変換するために用いられる多階層ニューラルネットの一種である．オートエンコーダ (AE) はくびれた中間層を通してあるドメインの信号をまた同じドメインに再構成するようにネットワークを学習することで中間層に情報を圧縮したコードが生成されることを期待したものであるが，DAE では再構成操作を通して雑音を除去することが目的となっている．DAE は画像処理はもとより，音声情報処理においても加算性雑音や乗算性雑音，残響の除去に応用され，良い結果が報告されている [5]．

DAE による雑音除去は，特定の種類の雑音に依存したものではなく，学習データさえ用意すれば様々な種類の雑音に適用可能である．さらには，発話音声の話者性を音韻情報に対する雑音ととらえることで，声質変換にも応用できる．すなわち，入力としてスペクトルやケプストラムなどの形である話者からの発話信号を受け取り，音韻情報をそのままに声質を他の話者のものに変換することが可能である．ただし，DAE をそのまま用いて声質変換を実現する場合には，ネットワークの学習のために予め変換元と変換先の音声のパラレルデータを大量に用意することが必要となる．少量のデータを効果的に用いる手法も提案されているが [6]，学習データに含まれない話者の音声に対応してシステム利用時にネットワークのパラメタを変化させることが困難なことから，基本的には特定の話者対か多対一を想定した声質変換手法である．

本研究ではこの問題を解決しニューラルネットを用いて一対多の声質変換を実現するために，図 2 に示すように DAE の入力側に話者特徴量を入力するためのネットワークを追加した構造を提案する．すなわち，ネットワークは（音声翻訳システムの出力合成音声などの）ある特定の話者の声質による発話内容を表す入力ベクトルと，目的とする話者の話者性を表す入力ベクトルを受け取り，音韻情報はそのままで声質を目的話者のものに置き換えた出力ベクトルを出力する．ネットワークは学習時において，多数の話者のパラレルデータを用いて入出力の関係を学習する．学習の完了した後，利用時においては全てのパラメタは固定であり，ユーザーからの音声を用いてパラメタを調整することは行わない．その代わりに学習データに含まれない任意のユーザーについて学習時と同じ方法で話者特徴量を抽出し，それをネットワークへの入力として用いることでユーザーの声質に変換された出力を得る．

2.2 ネットワークの学習

提案するネットワークは通常の DAE の構成に加えて話

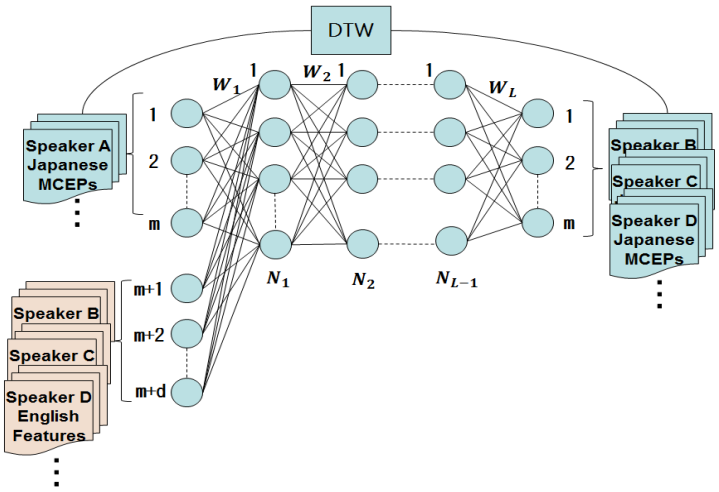


図 2 DAE の構成と，その学習に用いる訓練・教師データ．日本語・英語間のクロスリンガル声質変換の例.

者特徴量の入力を備えるため，それらを含めたパラメタの学習が必要となる．いくつかの方法が考えられるが，本論文では以下の手順に示すように制約付きボルツマンマシン (Restricted Boltzmann Machine:RBM) を用いたプレトレーニングと，それに続くバックプロパゲーションによるファインチューニングを元にしている．

Step 1: 変換元言語での発話内容を表す特徴量ベクトル集合を用い，1 レイヤーづつ RBM を教師なし学習することで積み上げる．

Step 2: 一定数のレイヤーが学習できたらそれらを連結することでエンコーダ側のネットワークを作成するとともに，その入出力を反転させたデコーダ側のネットワークを作成する．それらを連結することで，DAE の初期モデルを得る．

Step 3: 最下層の隠れ層に，話者特徴量の入力ベクトルからのコネクションを増設する．コネクションの結合重みは乱数で初期化する．

Step 4: 学習用に用意された入力側の音声発話と出力側の音声発話の対（ともに翻訳先言語の音声であり，入力側が固定話者，出力側が様々な話者を想定する）について，動的時間伸縮法（Dynamic Time Warping:DTW）によりフレームレベルの対応をとる．また出力側音声と同一の話者による他言語での音声 (翻訳元の言語の音声に相当する) から，発話内で固定の話者特徴量を抽出する．

Step 5: フレームレベルで発話内容を表す特徴量ベクトルと話者性を表す特徴量ベクトルを連結したものと，対応する発話内容を表す出力特徴量ベクトルを対として一つの学習サンプルとする．学習を効果的に行うために，学習データ全体でサンプルをシャッフルする．

Step 6: 学習サンプルを用いて，ネットワークをバックプロパゲーションによりファインチューニングする．

2.3 話者特徴量

話者特徴量としては話者の声質を再現する情報を表現したものであると同時に，1 から数発話程度の少ないラベルなし音声データから安定して抽出できるものが望ましい．様々なものが考えられるが，以下の実験では話者識別の分野で有効性が認められている GMM スーパーベクトル（GMM SuperVector: GSV） [7] を利用している．GMM スーパーベクトルの作成は以下のように行っている．

Step 1: 全学習用目標話者の発話を用いて話者非依存 GMM を学習する．

Step 2: 目標話者ごとに，少量の発話を用いて，MAP 適応 [8] により GMM の平均ベクトルを更新する．

Step 3: 更新された GMM の各コンポーネントの平均ベクトルを連結し，目標話者ごとの GSV を得る．

2.4 特徴量正規化について

入力特徴量として連続値ベクトルを用いるニューラルネットを RBM によるプレトレーニングを用いて学習する場合，特徴量を平均 0 分散 1 となるように正規化して用いることが一般的である．これは初段に用いるガウシアンベルヌーイ RBM の教師なし学習において，一般には分散項の学習が難しく，分散を 1 に固定して平均のみを学習させた方が性能が良いという知見に基づいたものである [9]．しかし，ニューラルネットを声質変換に用いる場合，この操作について注意が必要となる．

特徴量ベクトルの正規化はニューラルネットの入力側のデータに対して行うものであるが，AE や DAE ではエンコード用ネットワークの入出力を反転させて出力用ネットワークの初期値とすることから，出力側のデータについても同様の正規化を行うのが一貫性のある手順と考えられる．さらに，プレトレーニングの後はバックプロパゲーションによるファインチューニングを行うことになるが，その際

もプレトレーニングで初期化したパラメタを意味のあるものとするためプレトレーニングと同じ正規化を行ったデータを入出力に用いることになる．バックプロパゲーションでは 2 乗誤差最小化などの基準でパラメタの更新を行う．このとき，正規化を行った場合と行わない場合で，次元間の影響度が異なってくることがある．

すなわち，特徴量の種類によっては次元間でのダイナミックレンジの違いが大きい．バックプロパゲーションの目的関数として 2 乗誤差を用いる場合，正規化をしない場合はダイナミックレンジの大きい次元の方がダイナミックレンジの小さい次元に比べて相対的に 2 乗誤差への貢献度が大きく，全体の誤差を減らす上でより重視される傾向になると予想される．一方，正規化を行う場合には各次元のダイナミックレンジが揃えられるため，どの次元も一様に考慮されると予想される．どちらが良いかはタスクによるが，声質変換の目的でメルケプストラムを場合には低次の次元が高次の次元に比べてより重要である．加えて，メルケプストラムでは一般に低次元のダイナミックレンジの方が高次元のものよりも大きい．このことから，RBM によるプレトレーニングにおいて一般的な特徴量の正規化をそのまま適用してしまうと，ニューラルネットの学習において効果的な声質変換の実現のために期待される最適化と実際の目的関数との間で不整合が発生してしまう可能性が考えられる．

対処法としてはいくつかの方法が考えられるが，本論文では以下の方法について比較検討を行なった．

w/-norm　通常どおりの正規化を行う

w/o-norm　正規化を行わない

w/-m-norm　平均のみ正規化を行い，分散については正規化しない

w/-norm-wgt　通常どおりの正規化を行うとともに, バックプロパゲーションを行う際の目的関数を変更し次元間で 2 乗誤差の貢献度に重みを指定できるようにする．重みは元のデータにおける分散値を用いる．

正規化したデータを用いて学習を行なった場合，ニューラルネットの出力は正規化した特徴量に相当したものとなる．そこで，ニューラルネットによる声質変換の後処理として逆正規化を行なったものを MCD などの評価尺度の計算や，音声波形を再構成するモジュールへの入力として用いる．なお，話者特徴量についてはプレトレーニングの対象としていないため正規化を行なわずに用いている．

3. 実験的評価

3.1 実験条件

声質変換器の学習および評価には，高度言語情報融合フォーラム ALAGIN の提供する「日英・日中バイリンガル独話音声データベース」に含まれる音声データを使用した．使用したのは全て女声データであり，データベース中 1 名の話者を「特定の元話者」，9 名を「学習用目標話者」，3 名を「評価用目標話者」として用いた．評価用目標話者はテストセットとして用いるものであり，学習に用いた話者は含まれていない．特定の元話者は，音声自動翻訳システムの出力を想定したものである．実際に音声合成器により出力された音声データを声質変換器の入力側学習データとして使うことも考えられるが，ここではデータベース中の話者によるデータを用いた．実験で使用する特定の元話者と目標話者の発話は同一である必要があり，データベース中で対応している発話を抜き出して使用した．声質変換の対象となる音声は日本語発話であり，話者特徴量は対応する英語発話から抽出した．すなわち音声自動翻訳システムにおいて，英語音声を日本語音声に翻訳する状況を想定している．

表 1　目標話者ごとの発話数．1 発話あたりの音声時間は平均して 6 秒程度である．

Speaker	Number of utterances
EJF01	23
EJF02	698
EJF03	89
EJF05	115
EJF05	791
EJF08	100
EJF10	71
EJF101	495
EJF102	168
Sum	2550

ニューラルネットの学習で使用した学習用目標話者の発話数を表 1 に示す．元話者の発話数は目標話者毎に対応した発話を用いるために，その発話数と同数である．評価には，文章としては学習セットに同じ発話文が含まれているものから抽出した 11 発話 (Train-data) と，学習セットに全く含まれない発話文の中から抽出した 12 発話 (Test-data) を用いた．どちらも，音声としては学習セットに含まれない評価用話者による発話である．前者で評価した結果を学習セットスコア，後者で評価した結果をテストセットスコアとしている．評価スコアとしては，ネットワークから出力された音声と目標話者の音声の間の MCD を用いた．なお，実験では MCD 算出のために評価時においても日英パラレルデータを使用するが，実際の使用に際しては利用時にはパラレルデータは不要である．

特徴量抽出に用いた音声データのサンプリング周波数は 16kHz であり，データベースに収録されている 48kHz の音声ファイルをダウンサンプリングしたものである．発話内容を表す音声特徴量としては，25 次元のメルケプストラムを用いた．フレームシフトは 5ms としている．話者特徴量として用いる GSV を作成するための GMM につ

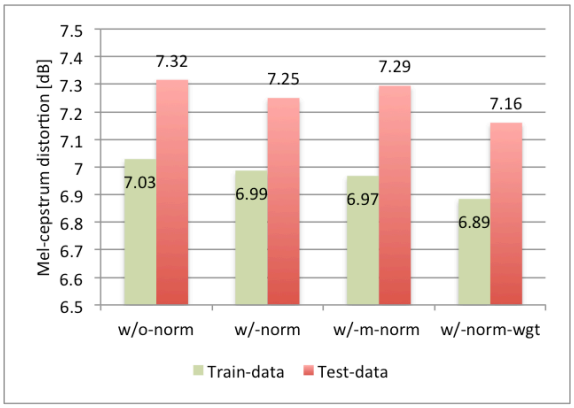


図 3 特徴量の正規化と MCD.

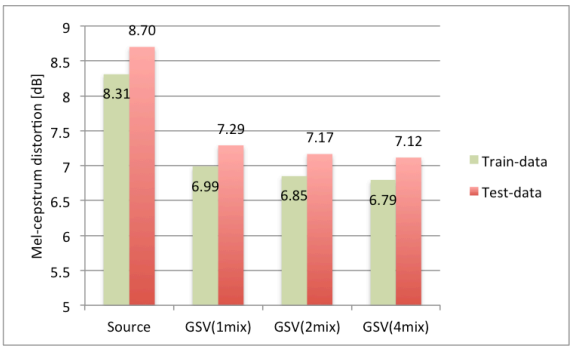


図 4 GSV 話者特徴量の抽出に用いる GMM の混合数と MCD.

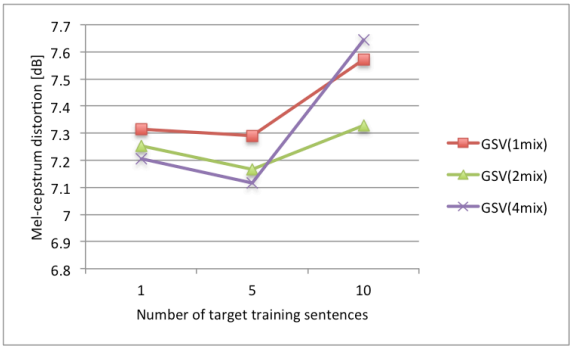


図 5 GSV 話者特徴量の推定単位と MCD.

いても，同じ特徴量を用いた．特徴量の抽出には SPTK (Speech Signal ToolKit) を用いた．

声質変換に用いる DAE は入力層と出力層の間に 3 層の隠れ層を設けたものとした．隠れ層のユニット数はいずれも 50 とした．本論文の実験では入力フレームのコンテキスト拡張は行なっておらず，発話内容を表す入力および出力特徴量は 25 次元である．ネットワークの入力側ではそれに加えて GSV による話者特徴量が入力される．話者特徴量の次元数はメルケプストラムの次元数と GMM の混合数を掛けたものであることから，例えば 2 混合の GSV を用いる場合はネットワークの入力側は発話内容を表す特徴量を含めて 75 次元の信号が入力される．

3.2 特徴量正規化についての事前実験

第 2.4 節で説明したように，ニューラルネットを声質変換に用いる場合特徴量の正規化の有無が結果に影響することが考えられる．そこでまず正規化の有無や方法について比較した結果を図 3 に示す．話者特徴量としては 1 混合の GSV(実際には特徴量の単純な一発話毎の平均) を用いている．発話内容を表す特徴量の正規化を行なわない場合 (w/o-norm) と，行なった場合 (w/-norm) を比較すると，学習セットに同じ発話文（音声としては別話者）が含まれる学習セットスコアと発話文も話者も同じものは学習セットに含まれないテストセットスコアの両方とも正規化を行なった場合の方が MCD がやや減少している．平均だけを正規化した場合 (w/-m-norm) も，おおよそ w/-norm と同様の MCD となった．一番よい結果は，平均と分散を正規化しつつ，バックプロパゲーションにおける目的関数で元の信号の分散を重みとして使用することで低次次元を強調した場合 (w/-norm-wgt) に得られた．これは，音声の再構成において重要で元々のダイナミックレンジの大きい低次次元を強調した最適化を行なった効果と考えられる．

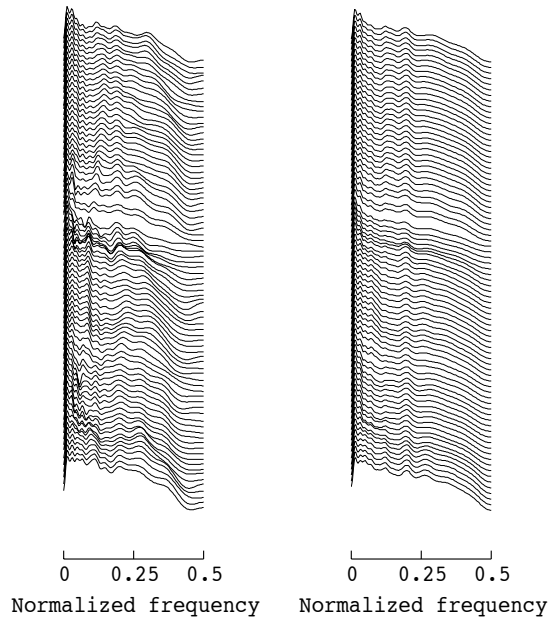
3.3 声質変換器の諸条件での MCD による評価

図 4 に話者特徴量の抽出に用いる GMM の混合数を変えた場合の結果を示す．話者特徴量は 5 発話毎に求め，それらに対応する発話で共通の値を用いている．発話内容を表す特徴量は，正規化しないものを用いている．Source として示したのは，声質変換を行なわずに元話者と目標話者の発話の間で MCD を評価したものである．混合数を 1 とした場合よりも 2 とした場合の方が学習セットスコアとテストセットスコアの両方が僅かであるが改善している．さらに，2 とした場合と 4 混合とした場合を比較しても，混合数を増やした方が僅かではあるが良い結果となった．GSV に 4 混合の GMM を用いた場合について声質変換を行なわない場合と比較すると，学習セットスコアで 1.52dB，評価セットスコアで 1.58dB の改善が得られた．

図 5 に話者特徴量ベクトルを作成する際に使用する発話数と MCD の関係を示す．MCD は全てテストセットについて求めたものである．N 発話を用いて GSV を得るための GMM を推定する場合，その N 発話では同じ値の話者特徴量を用いている．話者特徴量の推定は 1 発話毎でも効果はあるものの，5 発話毎とする方が MCD はより小さな値となった．しかし，10 発話毎とすると逆に MCD の値は上昇してしまった．これは，学習データにおける GSV の値の異なり数が減少したためと考えられる．

3.4 変換後音声の分析

図 6 に，声質変換の前と後の音声のランニングスペクトルの例を示す．話者特徴量には 1 混合の GSV を用いている．分析に用いた音声は，「クツシタノレンコン」の「タノ



(a) Source utterance (b) Voice converted utterance

図 6 メルケプストラムから生成したランニングスペクトルの一部．(a) は変換前，(b) は変換後のもの．

レ」の部分である．声質変換前のスペクトルと比較して声質変換後のスペクトルは単調で平滑化されてしまう傾向にあることが分かる．

実際どのような音声が生成されたのか確認するため，声質変換により得られた特徴量と目標話者の F0 を用いて合成した音声を作成し，受聴を行なった．合成された音声は F0 分析に起因するノイズの影響で元音声より音質が悪化していたが，実際に目標話者の声質に近づいていることが確認できた．

4. おわりに

DAE に話者特徴量入力を付加した構造を持つニューラルネットにより数発話程度のごく少量のラベルなし音声を用いて特定話者の任意の発話を任意話者の声質に変換する声質変換手法を提案した．多言語音声コーパスを用いた実験により，提案法の有効性を示した．今後は従来手法との比較を行なうとともに，ネットワーク構造の最適化やより大量のデータを用いた学習を容易に行なうために学習時においてもパラレルデータを必要としない方法のなどについて研究を進める予定である．

謝辞 本研究は JSPS 科研費 26280055 の助成を受けたものです．

参考文献

[1] Toda, T., Ohtani, Y. and Shikano, K.: One-to-Many and Many-to-One Voice Conversion Based on Eigenvoices, *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, Vol. 4, pp. IV–1249–IV–1252 (2007).

[2] Desai, S., Raghavendra, E., Yegnanarayana, B., Black, A. and Prahallad, K.: Voice conversion using Artificial Neural Networks, *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 3893–3896 (2009).

[3] Desai, S., Black, A. W., Yegnanarayana, B. and Prahallad, K.: Spectral Mapping Using Artificial Neural Networks for Voice Conversion, *Trans. Audio, Speech and Lang. Proc.*, Vol. 18, No. 5, pp. 954–964 (2010).

[4] Xue, S., Abdel-Hamid, O., Jiang, H. and Dai, L.: Direct adaptation of hybrid DNN/HMM model for fast speaker adaptation in LVCSR based on speaker code, *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 6339–6343 (2014).

[5] Ishii, T., Komiyama, H., Shinozaki, T., Horiuchi, Y. and Kuroiwa, S.: Reverberant speech recognition based on denoising autoencoder, *INTERSPEECH'13*, pp. 3512–3516 (2013).

[6] Nakashika, T., Takashima, R., Takiguchi, T. and Ariki, Y.: Voice conversion in high-order eigen space using deep belief nets, *INTERSPEECH'13*, pp. 369–372 (2013).

[7] Campbell, W., Sturim, D., Reynolds, D. and Solomonoff, A.: SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation, *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, Vol. 1, pp. I–I (2006).

[8] Gauvain, J. and Lee, C.-H.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, *Speech and Audio Processing, IEEE Transactions on*, Vol. 2, No. 2, pp. 291–298 (1994).

[9] Hinton, G.: A Practical Guide to Training Restricted Boltzmann Machines, Technical report (2010).