修 士 論 文 の 和 文 要 旨

研究科・専攻	大学院情報システム学研究科　社会知能情報学専攻　博士前期課程		
氏　　名	関井　祐介	学籍番号	1551015
論 文 題 目	オートエンコーダを利用した任意話者の声質変換手法の提案		

要　旨

声質変換は，入力音声を目的話者の声質に変換する技術である．声質変換手法として，従来は Gaussian Mixture Model (GMM) を用いた手法がよく用いられていたが, 近年の Deep Learning に関する技術の台頭により, Deep Neural Network (DNN) を用いた声質手法が注目されている．しかし，GMM や DNN を用いた手法の多くは一対一の声質変換手法を提案しており，任意話者の入力に対応した研究は少なく，従来の任意話者の声質変換手法は，一対一声質変換と比べ変換精度が劣ってしまうという問題がある．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，多くの訓練データが必要となり，かつ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，これらの問題を解決するため，オートエンコーダおよびスパース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，提案手法と従来手法を比較し，オートエンコーダを用いた手法は従来手法よりも若干高い精度でスペクトル変換を行い，変換時間を短縮することができた．スパースオートエンコーダを用いた手法では，オートエンコーダを用いた提案手法と比べ，スペクトル変換精度の向上および変換した音声の自然性を改善し，任意話者の声質変換精度を向上させることができた．

平成 28 年度修士論文

オートエンコーダを利用した
任意話者の声質変換手法の提案

電気通信大学 大学院情報システム学研究科
社会知能情報学専攻

学籍番号　：　1551015
氏名　：　関井 祐介

主任指導教員　：　田原 康之 准教授
指導教員　：　大須賀 昭彦 教授
指導教員　：　石川 冬樹 客員准教授

提出年月日　：　平成 29 年 2 月 22 日 (水)

概要

声質変換は，入力音声を目的話者の声質に変換する技術である．声質変換手法として，従来は Gaussian Mixture Model（GMM）を用いた手法がよく用いられていたが，近年の Deep Learning に関する技術の台頭により，Deep Neural Network（DNN）を用いた声質手法が注目されている．しかし，GMM や DNN を用いた手法の多くは一対一の声質変換手法を提案しており，任意話者の入力に対応した研究は少なく，従来の任意話者の声質変換手法は，一対一声質変換と比べ変換精度が劣ってしまうという問題がある．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，多くの訓練データが必要となり，かつ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，これらの問題を解決するため，オートエンコーダおよびスパース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，提案手法と従来手法を比較し，オートエンコーダを用いた手法は従来手法よりも若干高い精度でスペクトル変換を行い，変換時間を短縮することができた．スパースオートエンコーダを用いた手法では，オートエンコーダを用いた提案手法と比べ，スペクトル変換精度の向上および変換した音声の自然性を改善し，任意話者の声質変換精度を向上させることができた．

目次

図目次

表目次

第 1 章

はじめに

1.1 背景

近年，入力音声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．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あり，現在も研究が行われている [1, 2, 3]．しかし，近年，DNN（deep neural network）を用いた声質変換手法が GMM 等の従来の声質変換手法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 [4]．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いるが，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 [5]．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 [6] や RBM を拡張した DBN（deep belief network）を用いた変換手法 [7]，CRBM（conditional restricted Boltzmann machine）を用いた変換手法 [8], LSTM-RNN（Long Short-Term Memory based Recurrent Neural Network）を用いた変換手法 [9] 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 [10, 11].

声質変換では，主に声道特性（声質等の話者性および韻律情報）を表すスペクトル特徴量，声の高さを表す基本周波数（F0），および声のかすれや雑音を表す非周期性指標の 3 つの特徴量を用いる．また，従来の声質変換の研究の多くは，スペクトル特徴量としてメル周波数ケプストラム係数（MFCC）や MCEP（mel-cepstrum）を用いているが，高解像度なスペクトル特徴であるスペクトル包絡の変換を行う方が声質の類似性が高くなるという結果が報告されている [6, 12]．MFCC や MCEP といったスペクトル特徴量は，スペクトル包絡から人間の知覚に基づき抽出された特徴量であり，低周波数域の情報を強調した

ものである．類似性に差が生じる原因として，MFCC 等の変換では，MFCC 等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うため，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ためであると考えられる [13]．このため，声質変換に用いるスペクトル特徴量としてスペクトル包絡を選択すべきだと言える．

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．しかし，スペクトル包絡は MFCC 等に比べ次元数が大きく，声質変換器の作成に多くのデータを要すると考えられる．また一般的に，声質変換器の作成には入力話者と目的話者の同一内容発話によるパラレルデータが必要となるため，多くのパラレルデータを用意するには高いコストを要する．また，DNN を用いてスペクトル包絡の変換を行う手法 [14, 12] も提案されているが，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るという問題がある．データ収集に高いコストが掛かる，変換時間が長い，という条件のもとでは，アプリケーションの制作が困難になるため，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．

これ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特定の目的話者音声へ変換する多対一変換手法や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が提案されている [15, 11]．ここで言う任意話者とは，声質変換器の作成のために訓練データとして用いられていない話者のことである．任意話者の声質変換が行えるようになると，入力話者のために新たに声質変換器を作成する必要がなくなるため，不特定入力話者の声質変換を行う必要のあるアプリケーションを作成しやすくなるというメリットがある．しかし，現在提案されている任意話者の声質変換手法では，一対一声質変換手法と比べると精度が劣ってしまうという問題があり，任意話者の声質変換を用いた実用的なアプリケーションを作成するためにも，精度向上が求められている [11]．

1.2　声質変換の応用

声質変換技術は，海外映画の吹替音声を役者本人の声質で作成すること [16] や，アニメーション作品において声優の変更による声質に対する違和感を緩和させること [17] などに応用できると考えられる．また，喉頭摘出者の代替発声[*1]を，声質変換技術により自

[*1] がん摘出手術などにより声帯・喉頭を摘出した場合は，調音機能は正常であるにもかかわらず，発声が不可能になる．この場合，障がい者用意思伝達装置でなく，外部から喉に振動を与える電気式人工喉頭を利用し，代替発声を行うケースが多い．

然な音声へリアルタイム変換することへの応用が研究されている [18]．この技術を応用することで，リアルタイム声質変換により，音声を聞き取りやすい声質に変換することで，通話支援などを行うことができると考えられる．声質変換技術を応用したアプリケーションもリリースされており [19]，声質変換技術を用いて様々なアプリケーションが開発されることが期待される．

1.3 研究目的と手法概要

本研究では，一対一の声質変換および任意話者（多対一）の声質変換に要するデータ量の低減および声質変換時間の短縮を行うことを目的とし，オートエンコーダを用いた声質変換手法を提案する．また，任意話者の声質変換では，変換精度向上を目的とし，スパース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，まず入力話者および目的話者のオートエンコーダを作成し，各オートエンコーダから高次特徴量（隠れ層）を抽出する．入力話者のオートエンコーダから得られた高次特徴量を目的話者のオートエンコーダから得られた高次特徴量に近づけるようなDNN を作成し，変換された高次特徴量を得る．変換された高次特徴量に目的話者のオートエンコーダの重みを用いることで，音響特徴量を復元する．最後に，音響特徴量から音声合成を行い変換音声を得る．

1.4 本稿の構成

本稿の構成は以下の通りである．第 2 章では，声質変換システムの全体像や代表的手法である GMM を用いた声質変換，および現在主流である DNN を用いた声質変換について述べる．第 3 章では，最新の声質変換手法について幅広く示す．第 4 章では，本研究で提案する手法の全体像を説明し，オートエンコーダおよびスパースオートエンコーダの仕組みと利点について述べる．第 5 章では，評価実験として，一対一変換および任意話者変換について既存手法との比較実験を行い，結果を記載する．第 6 章では，評価実験結果について考察する．第 7 章では，本稿をまとめ，今後の課題について述べる．

第 2 章

従来技術

2.1　声質変換の全体像

一般的な統計的声質変換システムの概要を図 2.1 に示す．声質変換システムでは，まず，入力音声を音声分析し，基本周波数（F0），スペクトル特徴量（スペクトル包絡，MFCC 等），非周期性指標などの音響特徴量を得る．そして，それぞれの特徴量を予め構築したモデル等を用いて変換し，変換後の特徴量を求める．求められた特徴量を音声合成することで，変換音声が得られる．声質変換では，得られる音響特徴量は変換をした方が良いが，非周期性指標に関しては，変換を行わない場合が多い．

声質変換モデルの構築例を図 2.2 に示した．一般的に声質変換のモデル構築に用いる音声コーパスには，入力話者が発話した音声と，目的話者がそれと同一文を読み上げた音声が必要となる．そして，この入力音声と目的音声の長さを揃えたデータをパラレルデータと呼ぶ．近年では，パラレルデータを必要としない手法も提案されているが [20, 21]，本項ではパラレルデータを用いる場合のモデル構築を考える．また，声質変換モデルでは主に声質を表すスペクトル特徴量の変換を目的とする研究が多いため，スペクトル特徴量を変換するモデル構築を例として，図 2.2 に流れを記す．声質変換のモデル構築では，まず，パラレルデータである入力音声および目的音声をそれぞれ音声分析し，スペクトル特徴量を得る．得られたスペクトル特徴量を動的計画法によりアラインメントを取ることで，同一内容発話の長さを揃える．そして，アラインメントの取れた特徴量を入力とし，モデルを構築する．同様に，F0 や非周期性指標についても必要であればモデルを構築し，構築したモデルを図 2.1 の＜変換＞で用いることで，声質変換を実現できる．

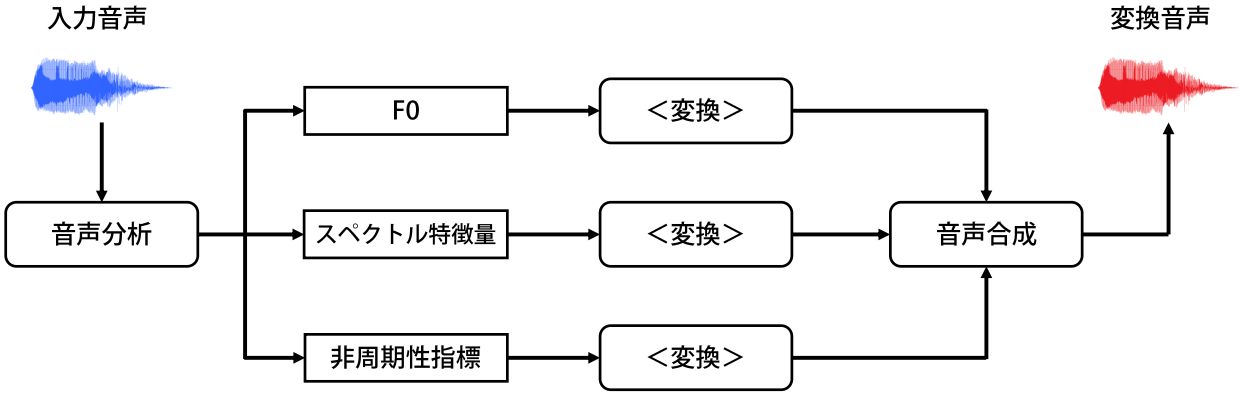

図 2.1　一般的な声質変換システム

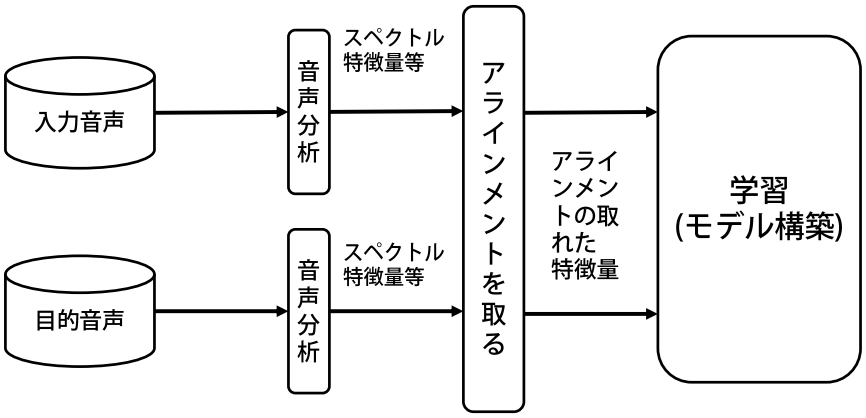

図 2.2　声質変換のモデル構築例

2.2　GMM を用いた声質変換

GMM を用いた声質変換手法 [2] は，統計的声質変換において高精度に声質変換を行える手法の一つである．この手法が登場する前に主流であったコードブックマッピング法 [22] では，音声の特徴量空間をベクトル量子化により離散的に表現していたが，変換音声も離散的に表現されるという問題があった．一方，GMM を用いる手法では，特徴量

空間を連続的に表現することができ，変換音声も連続的に表現することができるようになった．

この手法では，GMM を推定することで，入力話者の発話 X と目的話者の発話 Y の同時分布をモデル化する．学習時には，EM アルゴリズムを用いて同時確率を最大化する．

$$P(X,Y) = P(Z) = \sum_{m=1}^{M} w_m \mathcal{N}(z; \mu_m^{(z)}, \Sigma_m^{(z)}) \tag{2.1}$$

$$z = \begin{bmatrix} x \\ y \end{bmatrix},\ \mu_m^{(z)} = \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix},\ \Sigma_m^{(z)} = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix} \tag{2.2}$$

Z は X と Y の結合特徴量，M は混合数，w_m は重み，$\mathcal{N}(z; \mu_m^{(z)}, \Sigma_m^{(z)})$ は平均 $\mu_m^{(z)}$，共分散 $\Sigma_m^{(z)}$ とするガウス分布である．

変換時は，平均二乗誤差を最小化することで，入力特徴量 x から目的話者の特徴量 $\hat{y}$ を推定する．

$$\hat{y} = F(x) = \sum_{m=1}^{M} p_m(x)[\mu_m^{(y)} + \Sigma_m^{(yx)}(\Sigma_m^{(xy)})^{-1}(x - \mu_m^{(x)})] \tag{2.3}$$

$$p_m(x) = \frac{w_m \mathcal{N}(z; \mu_m^{(x)}, \Sigma_m^{(xx)})}{\sum_{m=1}^{M} w_m \mathcal{N}(z; \mu_m^{(x)}, \Sigma_m^{(xx)})} \tag{2.4}$$

ここで，$p_m(x)$ は m 番目の混合から生成される入力特徴量 x の事後確率である．

2.3 DNN を用いた声質変換

近年，入力話者の音響特徴量を目的話者の音響特徴量に変換する際，DNN を用いて変換を行う声質変換手法が主流となっている．DNN を用いた声質変換の流れを図 2.3 に示した．声質変換の流れについては，ほとんど図 2.1 と変わらないが，F0 やスペクトル包絡といった音響特徴量をまとめて DNN を用いて変換することもできる [14]．この手法では，入力音声から得られた音響特徴量からなる入力ベクトルを DNN の変換ネットワークに通すことで変換を行なう．ここで，このネットワークの変換能力はネットワークの構造（層数および各層の素子数），各素子に掛かる重みやバイアス，そして活性化関数によって決定される．また，重みやバイアスといったパラメータについては，入力話者の音声から得られた特徴量および目的話者の音声から得られた特徴量を用いてネットワークを学習させることで決定する．

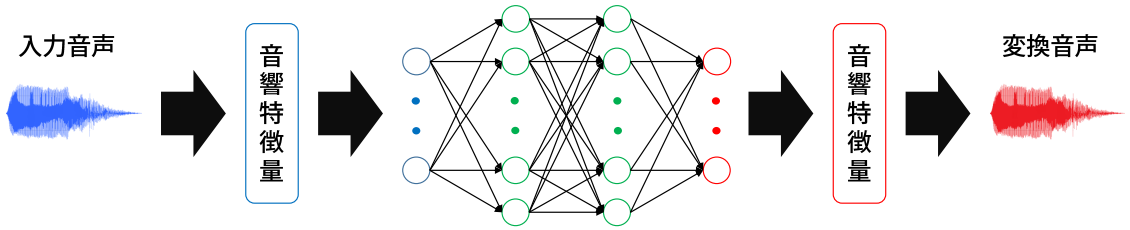

図 2.3　DNN を用いた声質変換

学習を行っていないネットワークでは，重みやバイアスのパラメータにランダムに初期値を与える．または，RBM や Denoising Autoencoder といった教師なし学習の手法を用いて pre-training（事前学習）を行なうことで，より学習が上手く行えるような初期値を与える．入力話者の特徴量および目的話者の特徴量を用いた教師あり学習では，誤差逆伝播法により各層の素子の誤差を計算し，DNN のモデルパラメータの更新を行なう．そして，これをネットワークの出力とそれに対応する正解データの誤差が最小になるまで繰り返し行なうことで，入力話者の特徴量を目的話者の特徴量に変換するネットワークが構築される．

第 3 章

関連研究

3.1 GMM や NMF を用いた声質変換

Yang ら [3] は，マルチフレーム特徴量と行列変量混合正規分布モデル（MV-GMM）を用いた声質変換手法を提案した．隣接したフレームの特徴量を各話者に関して並べた行列の結合行列の確率分布をモデル化した．また，MV-GMM を用いることで，特徴量空間と話者空間で独立した分散構造が得られる．評価実験では，従来の動的特徴量を用いる GMM よりも客観評価においても主観評価において優れていた．Yang らの手法は，声質変換の時系列特性のモデル化に有効であると考えられる．

Aihara ら [23] は，Semi-NMF（Semi-Non-negative Matrix Factorization）を用いた声質変換手法を提案した．NMF を用いた声質変換では，GMM 等を用いた声質変換と比べ，自然性の高い声質変換が可能である．しかし，多くのメモリを使うことや変換に時間を要するという問題がある．Arihara らの手法では，小さいスペクトル特徴量を使い，ADMM（Alternating Direction Method of Multipliers）を利用した Semi-NMF を用いて声質変換を行うことで，メモリの使用量を低減させ，変換時間の短縮を行った．評価実験では，従来の NMF と同程度の変換精度で変換時間を短縮したが，GMM を用いた声質変換よりも変換に時間を要するという結果であった．

3.2 RBM を用いた声質変換

Nakashika ら [5] は，話者依存型 CRBM を用い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 CRBM を学習させ，入力話者 CRBM より得られる高次特徴量を目的話者 CRBM により得られる高次特徴量へ NN を用いて変換し，変換で得られた高次特

徴量を目的話者 CRBM の逆射影を用いることで音響特徴量へ戻し，音声信号を得た．評価実験では，従来手法である GMM，RBM や RNN を用いた手法に比べ，変換精度が高かった．

Chen ら [24] は，スペクトル特徴量，F0 と非周期性指標を総合的に変換する話者変換手法を提案した．スペクトル変換では，RBM と BBAM（Bernoulli bidirectional associative memories）を結合した GTDNN を用いた．F0 変換では LSTM-RNN，非周期性指標の変換では DNN をそれぞれ用いた．評価実験では，GMM を用いたベースライン手法より類似性も自然性も優れているという結果が得られた．

3.3 オートエンコーダを利用した声質変換

Nguyen ら [12] は，スペクトル包絡，F0 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，変換器の作成には多くのデータ量を要し，声質変換には長い時間を要すると考えられる．

Mohammadi らはオートエンコーダを用いた声質変換手法を提案している [10, 25]．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では，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 と次元が小さいため，DNN の隠れ層の層数が増えても，少ないデータ数で変換器を作成でき，比較的短い時間で声質変換を実現できると考えられる．また，Stacked Joint-Autoencoder（SJAE）を用いた声質変換手法では，入力話者，目的話者それぞれのオートエンコーダを作成する際，それぞれの誤差関数として，オートエンコーダの出力誤差に 2 つのオートエンコーダの中間層の誤差を加えたものを用いる．このような誤差関数を用いて学習を行い，作成された入力話者オートエンコーダのエンコード重み，目的話者オートエンコーダのデコード重みを結合させた DNN を Joint-Autoencoder とした．さらに，これを Stacked　Autoencoder のように積み上げた SJAE を用いて声質変

換を行った．加えて，SJAE を fine-tuning した手法も提案したが，fine-tuning したものは SJAE で抽出した特徴を平滑化してしまうため，fine-tuning を行わない SJAE の方が主観評価の結果が良かったと報告されている．

3.4 任意話者の声質変換

Liu ら [11] 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 を用いた手法には劣るものの，大差ない精度で任意話者の声質変換を実現した．また，音響特徴量に MCEP を用いているため，スペクトル包絡を用いることで精度向上が期待できる．

3.5 パラレルデータフリーな声質変換

近年，パラレルデータを必要としない声質変換手法が提案されている．パラレルデータが必要となる手法では，学習に用いる音声コーパスの発話内容に制約がかかるため，学習データの準備が容易ではない．一方，パラレルデータを必要としない声質変換が行えると，入力話者や目的話者の発話内容に制約がないため，異言語間の声質変換が比較的容易にできると考えられる．本節では，このようなパラレルデータフリーな声質変換手法を示す．

Xie ら [20] は，話者非依存 DNN を用いた自動音声認識（ASR）と KL ダイバージェンスを用いた KLD-DNN を利用したパラレルデータフリーな声質変換手法を提案した．話者非依存 DNN ASR は音素空間において入力話者と目的話者の違いを平均化するために用い，KL ダイバージェンスは，入力話者フレームを目的話者の TTS Senone（音素よりも細かいレベルの音声要素）と音素クラスタに変換するために用いた．評価実験では，客観評価，主観評価ともにパラレルデータを必要とする従来手法より大幅に精度を改善したと報告されている．

Nakashika ら [21] は，ボルツマンマシンに基づいた話者適応モデルを用いたパラレルデータフリーな声質変換手法を提案した．提案されたモデルでは，スペクトル特徴量を話者に依存する項と依存しない項に分離する．入力話者，目的話者のモデルをそれぞれ構築

し，入力音声を話者に依存する項と依存しない項に分離し，話者に依存する項を目的話者の話者に依存する項に置き換えることで，パラレルデータを用いない声質変換を実現した．評価実験では，パラレルデータを用いる GMM を用いた声質変換手法には及ばなかったが，既存のパラレルデータフリーな声質変換手法より優れていた．

第 4 章

提案手法

4.1 提案手法の全体像

本研究で用いた声質変換システムを図 4.1 に記載した．本システムでは，まず入力音声から TANDEM-STRAIGHT[26] により F0，スペクトル包絡と非周期性指標のパラメータを得る．次に，F0 は線形変換，スペクトル包絡は本稿で提案する手法を用いてそれぞれ変換を行い，非周期性指標は変換を行わず入力音声から得られたものを用いる．ここで，F0 およびスペクトル包絡の変換器作成には，目的話者の音声を STRAIGHT 分析することで得られる F0 およびスペクトル包絡を用いる．そして，変換した F0 およびスペクトル包絡と入力音声の非周期性指標を用いて，音声合成を行うことで変換音声を得る．なお，本システムにおいて F0 の変換は以下の式で行った．

$$\hat{y}_t = \frac{\sigma^{(y)}}{\sigma^{(x)}}(x_t - \mu^{(x)}) + \mu^{(y)} \tag{4.1}$$

x_t，$\hat{y}_t$ は対数尺度での入力音声，目的音声の F0 である．$\mu^{(x)}$，$\sigma^{(x)}$ はそれぞれ入力話者音声の対数 F0 の平均および標準偏差である．同様に，$\mu^{(y)}$，$\sigma^{(y)}$ はそれぞれ目的話者音声の対数 F0 の平均および標準偏差である．

本研究ではスペクトル包絡の変換を以下の流れで行う（図 4.2）．

Step 1: 入力音声からスペクトル包絡を抽出する

Step 2: 入力話者音声から得られるスペクトル包絡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

Step 3: 入力話者オートエンコーダから得られる高次特徴量を DNN を用いて目的話者オートエンコーダの高次特徴量へ変換する

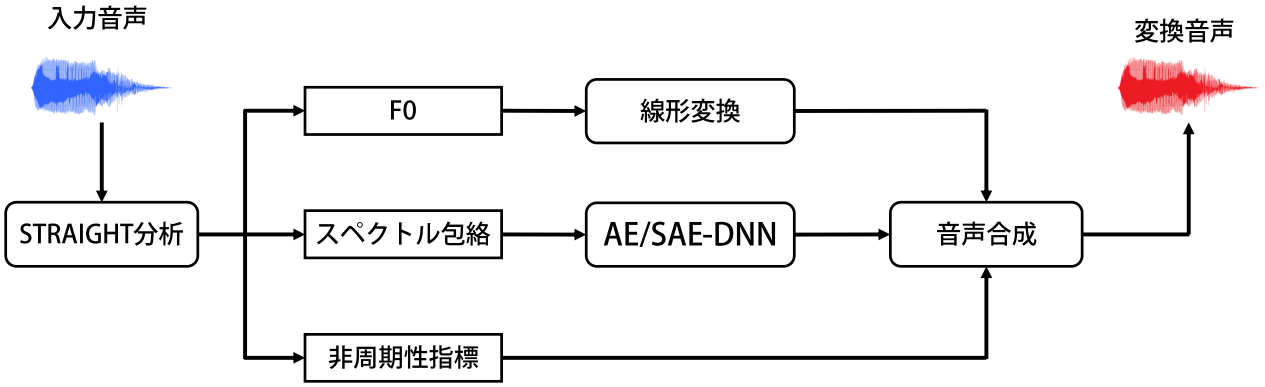

図 4.1　提案する声質変換システム

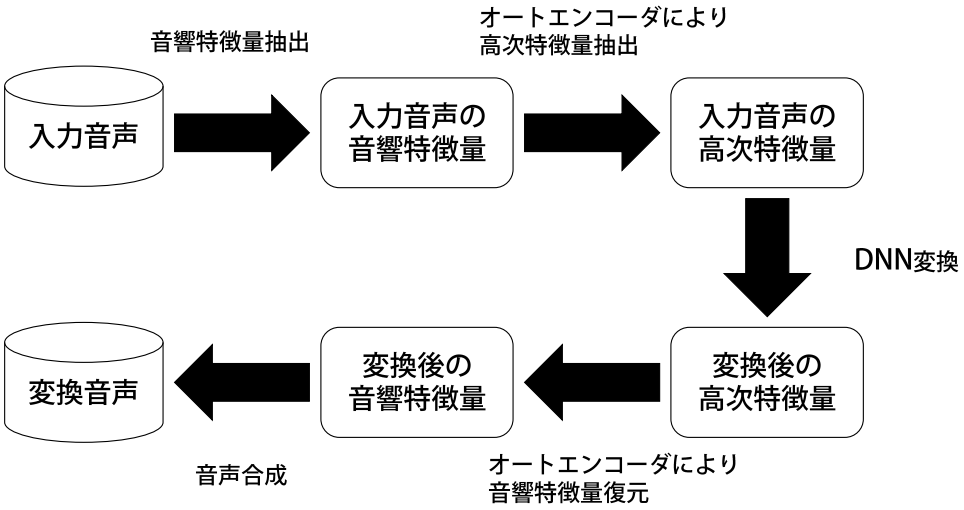

図 4.2　声質変換の流れ

Step 4:　変換された高次特徴量を目的話者音声から得られるスペクトル包絡で訓練されたオートエンコーダを用いてスペクトル包絡へ復元する

Step 5:　得られたスペクトル包絡を元に，音声合成によって変換音声を求める

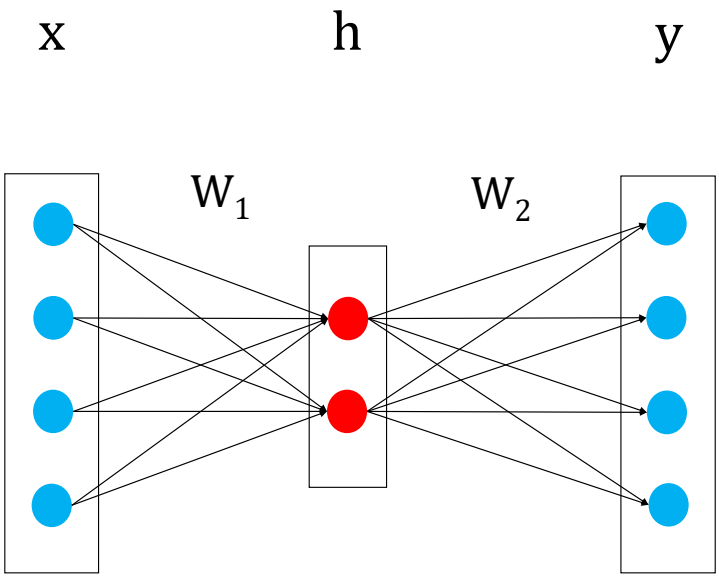

図 4.3　オートエンコーダ

4.2　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;AE）[27] 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 NN であるため，入力値のみを必要とする．

図 4.3 のように入力層，隠れ層，出力層の 3 層からなる NN を考え，入力素子 x，隠れ素子 h，出力素子 y とした時，オートエンコーダを以下のように表す．

$$\mathrm{h} = f(\mathrm{W}_1\mathrm{x} + \mathrm{b}_1) \tag{4.2}$$

$$\mathrm{y} = g(\mathrm{W}_2\mathrm{h} + \mathrm{b}_2) \tag{4.3}$$

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 (4.2)，(4.3) 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．

$$\mathrm{y} = g(\mathrm{W}_2 f(\mathrm{W}_1\mathrm{x} + \mathrm{b}_1) + \mathrm{b}_2) \tag{4.4}$$

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 $\mathrm{W}_1, \mathrm{W}_2$ とバイアス $\mathrm{b}_1, \mathrm{b}_2$ 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．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．

$$\mathrm{E} = ||\mathrm{x} - \mathrm{y}||^2 \tag{4.5}$$

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 DNN 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 [10]．また，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．

4.3 スパースオートエンコーダ

スパースオートエンコーダ（Sparse Autoencoder; SAE）は，オートエンコーダの隠れ層がスパースになるよう制約を加えたオートエンコーダである．誤差関数は以下のようになる．

$$\mathrm{E} = ||\mathrm{x} - \mathrm{y}||^2 + \sum_{i=1}^{D_\mathrm{h}} \mathrm{KL}(\theta||\mathrm{h}_i) \tag{4.6}$$

ここで，θ は正則化を制御するためのパラメータ，h_i は隠れ層 h の i 番目の素子，D_h は h の次元数である，また，通常はスパース項として L1 ノルムを使用するが，L1 ノルムは微分不可能のため，代わりに $\sum_{i=1}^{D_\mathrm{h}} \mathrm{KL}(\theta||\mathrm{h}_i)$ を用いた．このスパース項は θ と h_i をパラメータとする 2 つのベルヌーイ分布の KL 情報量であることから以下のように記述できる [28].

$$\mathrm{KL}(\theta||\mathrm{h}_i) = \theta \log \frac{\theta}{\mathrm{h}_i} + (1-\theta) \log \frac{1-\theta}{1-\mathrm{h}_i} \tag{4.7}$$

スパースオートエンコーダは誤差関数にスパース項を加えたことにより，汎化能力が高まり，より抽象的な高次特徴量が得られる．これにより，複数話者音声のスペクトル包絡からなる入力に対し，汎化された高次特徴量が得られることで，任意話者の入力に対応しやすくなると考えられる．

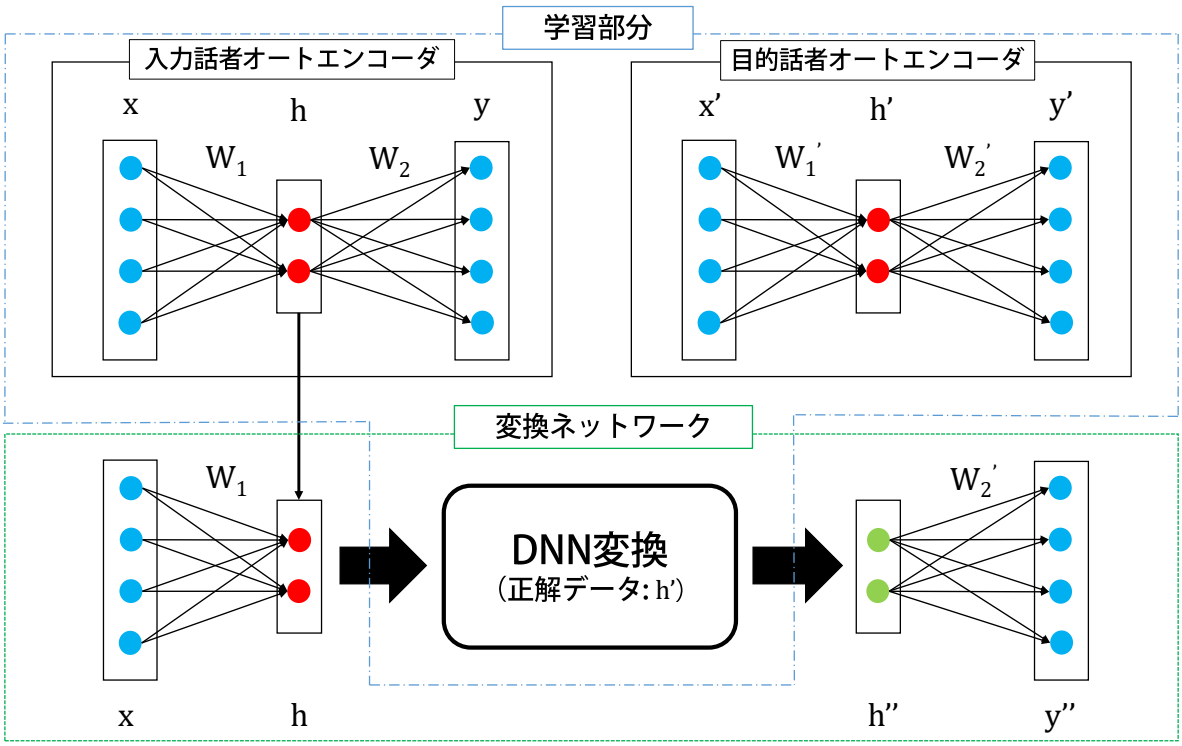

図 4.4　特徴量変換の全体構造

4.4　スペクトル特徴量変換

本手法では，オートエンコーダおよびスパースオートエンコーダから抽出した高次特徴量を利用する．オートエンコーダを用いて次元圧縮された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データ量を低減し，変換時間を短縮することが目的である．また，スパースオートエンコーダを用いて得られた高次特徴量は汎化されているので，従来手法よりも任意話者の声質変換の精度を向上させることができると考えられる．

本研究では，図 4.4 のようなスペクトル特徴量（本稿では，スペクトル包絡）変換構造を提案する．まず，各話者のオートエンコーダと高次特徴量変換を行う DNN の学習を行う．入力話者のスペクトル特徴量 x，目的話者のスペクトル特徴量 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．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 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．次に，作成したオートエンコーダ

と DNN を用いて声質変換器の作成を行う（図 4.4 下のフロー）．まず，入力音声のスペクトル特徴量を入力話者オートエンコーダのエンコーダ重み W_1 を利用することで高次特徴量を得る．得られた高次特徴量を DNN を用いて目的話者の高次特徴量に変換する．そして，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' を利用することでスペクトル特徴量を復元し，変換されたスペクトル特徴量 y'' が得られる．そして，このオートエンコーダと高次特徴量変換 DNN を結合した DNN 全体を fine-tuning する．この一連の流れのネットワークを構築することで，1 つのネットワークによりスペクトル特徴量変換が行える．

任意話者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．複数話者から構成される訓練データを用いることで，オートエンコーダおよびスパースオートエンコーダはより一般化された高次特徴量を与えると考えられる．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することが期待される．

第 5 章

評価実験

5.1　予備実験

オートエンコーダを用いた提案手法及び比較手法の最適なパラメータを特定するため，予備実験を行った．また，訓練に用いるデータ量を変化させることで，スペクトル変換の精度がどのように変化するのか検証を行った．予備実験には，ソリッドスフィア社が作成した音声データセット[*1]より，入力話者に男性話者（YMGT），目的話者に女性話者（RDY）を用いた．パラメータ調整では，訓練データとして 450 発話，テストデータとして 50 発話を用いた．

5.1.1　パラメータ調整

提案手法の最適なパラメータの特定実験では，100 次元の高次特徴量を隠れ層の層数と隠れ層の素子数をパラメータとする様々な DNN を用いて変換を行った．学習回数については，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．スペクトル変換の評価には，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）を用いた．つまり，LSD の値が小さいほどスペクトルの変換精度が良いということを表している．また，LSD の評価式は以下の通りである．

[*1] 男性話者 4 名，女性話者 6 名，各話者 500 発話で構成．
日常会話の発話内容を感情を排して読み上げた音声を収録．

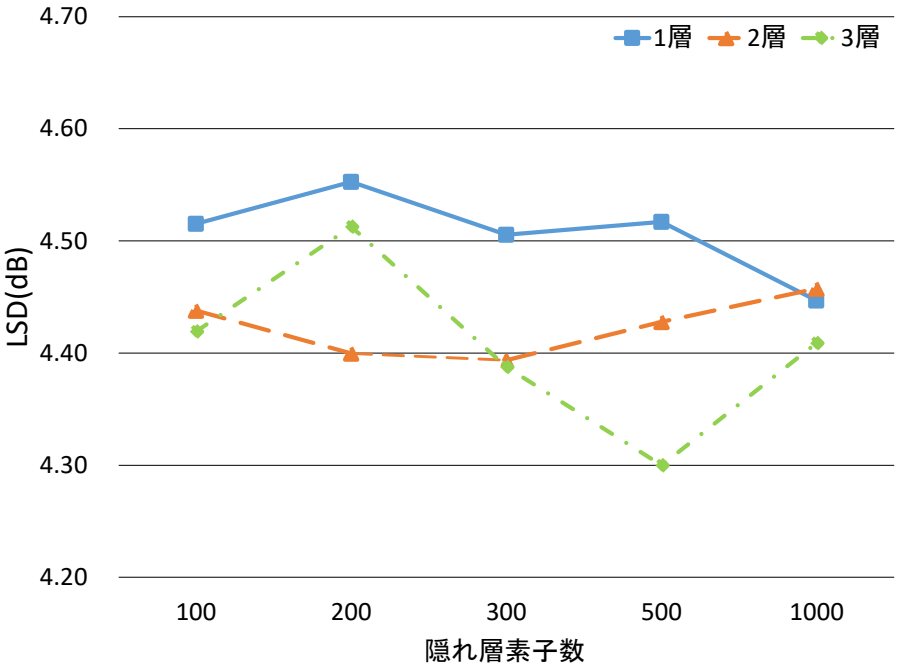

図 5.1　高次特徴量を用いた手法のパラメータ変化による LSD 変化

$$\mathrm{LSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(10\log_{10}\frac{x_i}{y_i}\right)^2} \tag{5.1}$$

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，n はスペクトルの次元数（今回は 513）である．パラメータを変化させ，スペクトル変換した結果を図 5.1 に示した．図 5.1 より，本実験には最も精度の高かった 3 層 500 素子のものを用いる．また，本実験ではこれより簡易な DNN として 50 次元の高次特徴量を用いた手法を合わせて提案手法として用いる．

比較手法の最適なパラメータの特定実験では，Nguyen らの手法 [12] を参考とし，513 次元対数スペクトル包絡を DNN を用いて変換する手法を用いた．513 次元対数スペクトル包絡を，隠れ層の層数と隠れ層の素子数をパラメータとする様々な DNN を用いて変換させた．学習回数は 30 とした．スペクトル変換の評価には LSD を用いた．パラメータを変化させ，スペクトル変換した結果を図 5.2 に示した．図 5.2 より，3 層 3000 素子の DNN が最も精度が高く，次に 3 層 100 素子の DNN の精度が高かったため，本実験ではこの 2 つの DNN を用いた手法を比較手法として用いた．

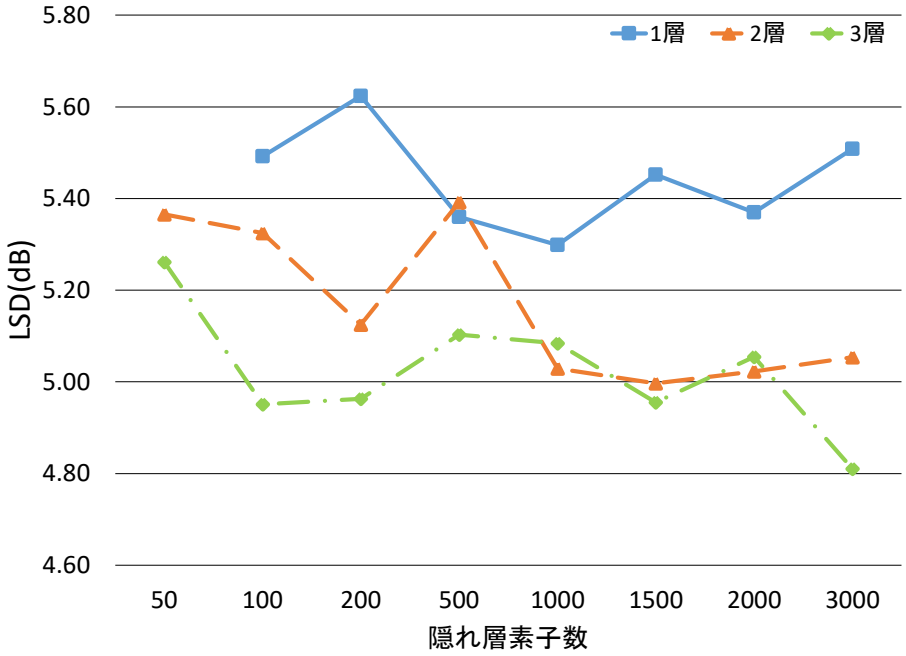

図 5.2　スペクトル包絡 DNN 変換手法のパラメータ変化による LSD 変化

5.1.2　データ量変化

訓練に用いるデータ量の変化により，スペクトル変換の精度がどのように変化するのか検証を行った．スペクトル変換の手法には，提案手法として 50 次元の高次特徴量を用いた手法（AE50），100 次元の高次特徴量を用いた手法（AE100），比較手法として，513 次元対数スペクトル包絡を DNN を用いて変換する手法で，DNN の構造を変えた 2 種（SPEC3000，SPEC100）を用いた．AE50，AE100，SPEC3000，SPEC100 における DNN の隠れ層及び隠れ層素子数はそれぞれ，2 層 200 素子，3 層 500 素子，3 層 3000 素子，3 層 100 素子とした．学習回数については，AE50 と AE100 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，SPEC3000 と SPEC100 の学習回数は 30 とした．スペクトル変換の評価には LSD を用いた．データ量を変化させ，スペクトル変換した結果を図 5.3 に示した．訓練に用いるデータ数が少ない時は，AE50 の精度が高く，データ数が多くなると，AE100 や SPEC3000 の精度が高くなった．このことから，簡易な DNN を用いた手法（AE50）では，データ数が少ない時に高い精度が得られ，複雑な DNN を用いた手法（AE100，SPEC3000）はデータ数が多くなると高い精度を得られ

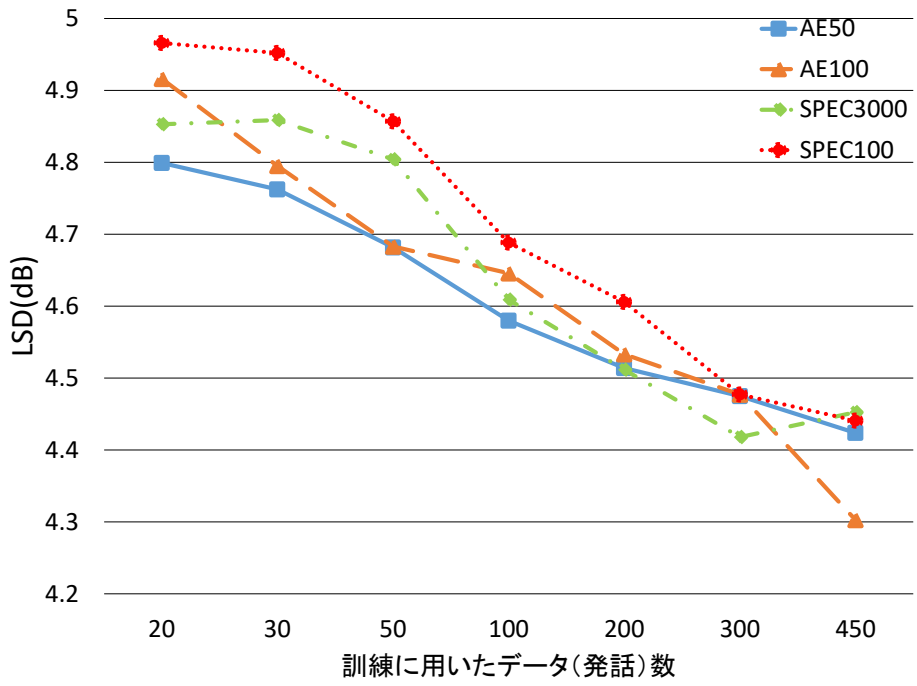

図 5.3　訓練データ数変化による LSD 変化

るという仮説が立てられる．しかし，比較的簡易な DNN を用いている SPEC100 に関しては，データ数が少ない時は精度が低く，データ数が多くなると，他手法と同程度の精度が得られた．この仮説が正しいかどうかを確かめるため，本実験ではデータセットを 2 種用意し，検証を行う．

5.2　オートエンコーダを用いた声質変換手法の評価実験

5.2.1　評価方法

一対一変換と任意話者変換（多対一変換）においてオートエンコーダを用いた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（YMGTtoKJM，KJMtoHM，TKtoYMGT，HMtoTK）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 人に対し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．一対一変換と任意話者変換ともに，訓練データには大規模コーパスとして 300 発話，小規模コーパスとして 20 発話を用いた．ここで，任意話者変換に用いたコーパスの総発話数は，訓練話者数に関わらず一定である．また，訓練データに用いていない 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 4 つの比較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（AE50）と 100 次元の高次特徴量を用いた手法（AE100）の 2 つの手法を用いた．なお，AE50 および AE100 は fine-tuning を行わず，オートエンコーダと高次特徴量変換 DNN を結合したものとした．比較手法には，GMM を用いた手法（JDGMM）[2]，MFCC を DNN を用いて変換した手法（MFCC-DNN）[4]，対数スペクトル包絡を DNN を用いて変換した手法（SPEC3000，SPEC100）[12] の 4 つの手法を用いた．入出力音響特徴量として，TANDEM-STRAIGHT により 1 フレーム 40ms，フレームシフト 5ms として求めた 513 次元の対数スペクトル包絡を AE50，AE100，SPEC3000，SPEC100 で用い，スペクトル包絡より計算された 25 次元 MFCC を JDGMM，MFCC-DNN で用いた．JDGMM における混合数は 64 とし，MFCC-DNN，SPEC3000，SPEC100，AE50，AE100 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，3 層 100 素子，2 層 200 素子，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 [29]，ADAM（α=0.0001）[30] を用いた．学習回数については，MFCC-DNN，SPEC3000，SPEC100 はそれぞれ 200，20，20 とし，AE50，AE100 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は，MFCC-DNN は Desai らの手法 [4] を参考にし，他は予備実験により決定した．

一般的に声質変換手法の評価は客観評価と主観評価の 2 つで行われる．客観評価基準として，LSD と音響特徴量の変換所要時間の 2 つを用いた．主観評価では，MOS（mean opinion score）に基づき評価を行った．MOS とは，人間の知覚に基づき音声の質を評価するための指標である．MOS は 1 から 5 の間の数値で表され，1 が最も音質が悪く，5 が最も音質が良いことを示す．被験者（20 代男女）9 人に目的話者音声と変換音声を試聴させ，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い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

表 5.1　小規模コーパスにおける一対一変換の LSD（dB）

target	AE50	AE100	JDGMM	MFCC-DNN	SPEC3000	SPEC100
YMGT → KJM	4.53	**4.39**	5.80	5.44	4.44	4.74
KJM → HM	4.71	4.66	6.77	6.31	**4.62**	4.78
TK → YMGT	4.56	**4.33**	5.66	5.30	4.45	4.67
HM → TK	4.18	**4.12**	5.12	4.85	4.14	4.28
average	4.50	**4.38**	5.84	5.48	4.41	4.61

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いないランダムに選択した 1 発話を評価に用いた．被験者にはどの変換手法で変換した音声か分からないよう，変換音声の試聴順をランダムに並べ試聴させた．

5.2.2　一対一変換の結果

小規模コーパスを用いた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 5.1 に示した．LSD 値を比較すると，提案手法である AE50，AE100 および先行研究手法である SPEC3000，SPEC100 は JDGMM および MFCC-DNN 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-DNN 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．しかし，MFCC は人間の音声知覚を考慮した特徴量であるため，図 5.4 の主観評価においては，手法間に LSD で見られた差ほど大きな差はないことが観測できる．また，AE100，SPEC3000 のスペクトル変換精度に大きな差はなかったが，これらの 2 手法に比べ AE50 と SPEC100 は少し精度が下がるという結果であった．AE50 より AE100 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．大規模コーパスを用いた一対一変換の結果を示した表 5.2 において，表 5.1 と似た傾向の結果が得られたが，AE50 と SPEC100 の変換精度と AE100 と SPEC3000 の変換精度差が小さくなっている．AE50 と SPEC100 は AE100 や SPEC3000 よりも簡易な DNN を用いていることから，簡易な DNN を用いた手法は訓練に多くのデータ量を要すると考えられる．よって，予備実験において立てた，”簡易な DNN を用いた手法では，データ数が少ない時に高い精度が得られ，複雑な DNN を用いた手法はデータ数が多くなると高い精度を得られる”という仮説は否定された．

図 5.4 と図 5.5 では，それぞれ小規模コーパスを用いた実験と大規模コーパスを用いた

表 5.2　大規模コーパスにおける一対一変換の LSD（dB）

target	AE50	AE100	JDGMM	MFCC-DNN	SPEC3000	SPEC100
YMGT → KJM	4.08	**4.04**	5.06	5.19	4.06	4.17
KJM → HM	4.29	**4.20**	4.72	4.96	4.21	4.35
TK → YMGT	4.02	**3.96**	5.04	5.10	**3.96**	4.12
HM → TK	3.88	**3.82**	4.55	4.50	3.88	3.97
average	4.07	**4.01**	4.84	4.94	4.03	4.15

実験において，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．小規模コーパスを用いた実験では，類似性は SPEC3000 が最も高く，自然性は AE100 が最も高かったが，どちらも大きな差はなかった．大規模コーパスを用いた実験では，類似性と自然性ともに手法間に大きな差はなかったが，類似性は提案手法である AE50 が最も MOS 値が高かった．小規模コーパスを用いた実験において，自然性については AE100 と SPEC100 で有意差が見られたが，自然性の他の組合せや類似性については有意差はなかった．大規模コーパスを用いた実験では，類似性，自然性共に一要因の分散分析を行ったが，有意差はなかった．なお，大規模コーパスにおいて，特に自然性では提案手法である AE50 と AE100 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．

図 5.6 では，各手法によるスペクトル変換に要する時間比較を行った．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 1 発話分（約 1 秒）のスペクトルを変換するのに要した時間である．表 5.1 と表 5.2 では，AE100，SPEC3000 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 AE100 が変換に要した時間が 0.36 秒であるのに対し，先行研究手法である SPEC3000 は変換に要する時間が 1.48 秒と 4 倍以上の時間を要するという結果となった．本研究において，声質変換に要する時間の目標値としては，入力音声から変換音声を求めるまで 2 秒程度とした．現在の音声合成技術では，約 2 秒間の音声を特徴量に分解し音声合成で音声を求めるには，およそ 1.9 秒要する[*2]ため，目標値を達成するには特徴量変換は 0.1 秒程度で行う必要があるが，これを満たす手法は，MFCC-DNN のみである．しかし，MFCC-DNN は変換精度があまり良

[*2] TANDEM-STRAIGHT を利用した場合

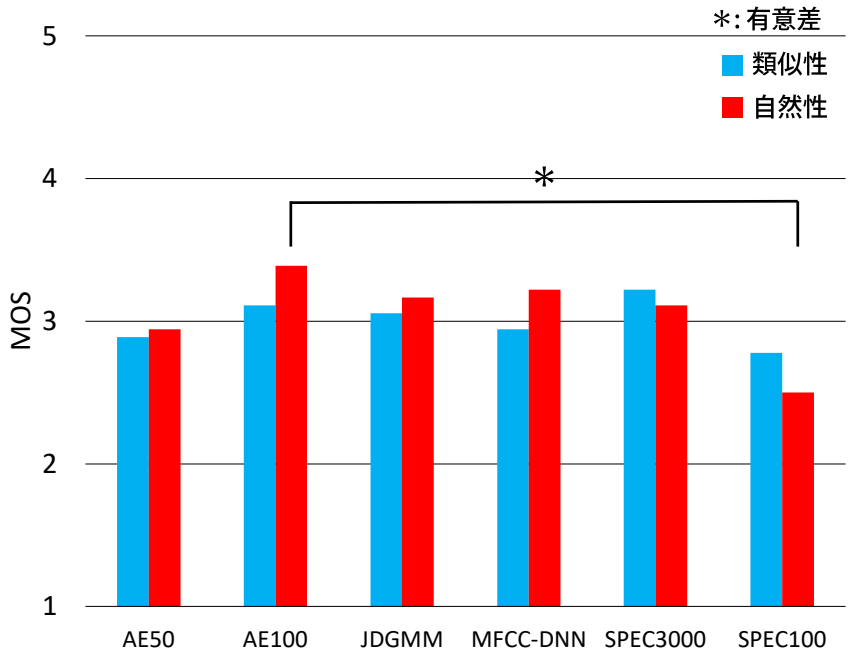

図 5.4　小規模コーパスにおける一対一変換の主観評価

くないため，変換精度が良くかつ変換時間が短いものを考えると，AE100 となり，声質変換に要する時間は 2.3 秒と少々目標値を超えてしまうが，同程度の精度で変換が行える SPEC3000 と比較すると，大幅に時間は短縮された．

また，大規模コーパスにおいて，従来手法である JDGMM は MFCC-DNN より変換精度は高かったが，MFCC-DNN の変換時間が約 0.13 秒に対し，JDGMM は変換に 0.8 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．一方，小規模コーパスにおいては，変換精度と変換時間ともに JDGMM より MFCC-DNN の方が優れていたという結果が得られた．

以上の結果より，一対一変換において提案手法である AE100 が変換精度，変換時間共に優れていたことがわかる．

5.2.3　任意話者変換の結果

任意話者変換実験では，JDGMM は多対一変換を行えないため，JDGMM を他の多対一変換手法と比較するための一対一変換手法として用いた．以下，JDGMM については，各変換組（HMtoKRT，KRTtoHM）の一対一変換を行った結果を記載した．

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

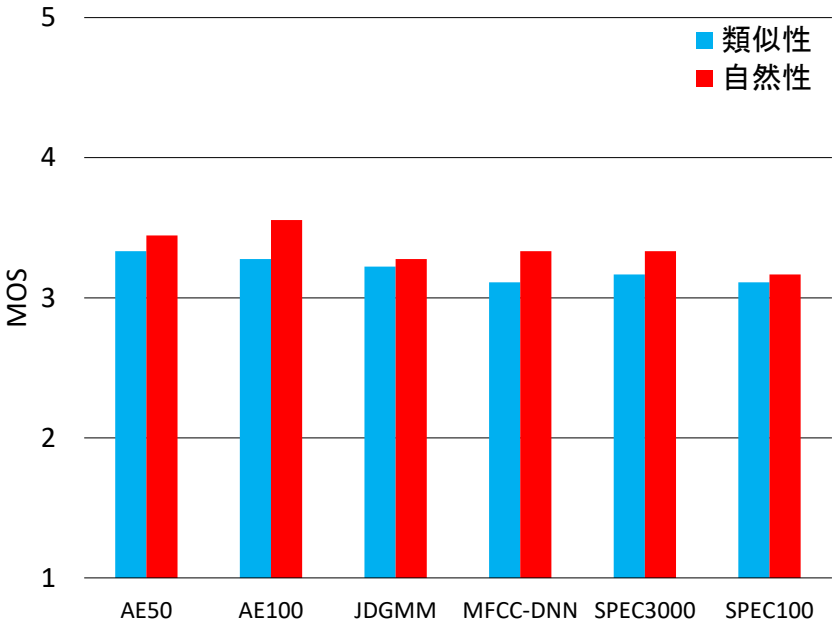

図 5.5　大規模コーパスにおける一対一変換の主観評価

積である 8 通りに対し各手法の LSD 評価を行い，訓練話者数毎にまとめた結果を表 5.3，表 5.4（それぞれ小規模コーパス，大規模コーパスを用いた実験に対応）に示した．ここで，mix2 というのは，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．同様に，mix4，mix6，mix8 はそれぞれ，訓練データの話者が 4 人，6 人，8 人であることを意味する．表 5.1 や表 5.2 と同様，JDGMM と MFCC-DNN 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．任意話者変換においても，大規模コーパスと小規模コーパス共に AE100 の手法が最も高い精度を得た．また，大規模コーパスにおいては AE50 も SPEC3000 よりも良い精度が得られた．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 5.3，表 5.4 における訓練話者数の変化に対する LSD の変化を見た時，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．ただし，表 5.3 の SPEC100 に関しては訓練話者数が増えると変換精度も向上するという結果となった．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

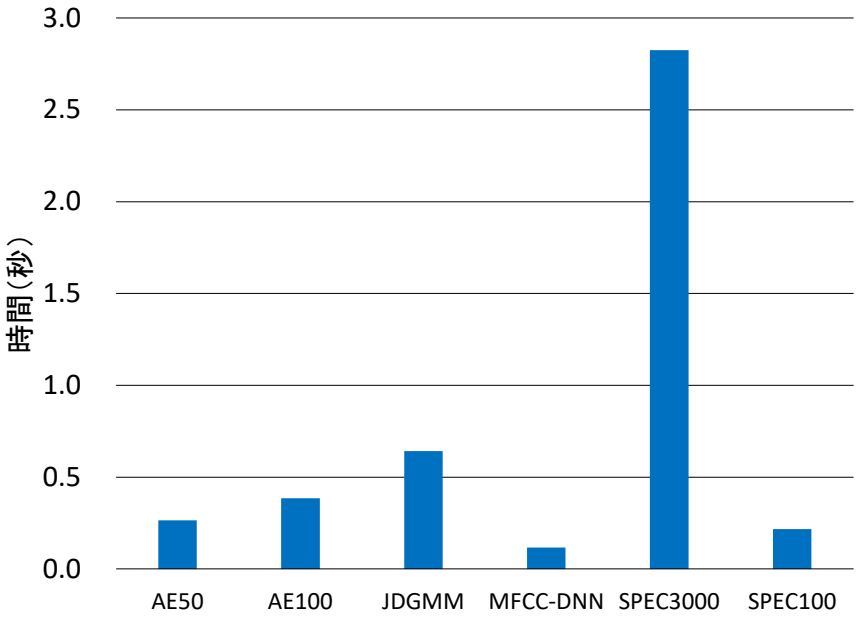

図 5.6　変換時間（秒）

表 5.3　小規模コーパスにおける訓練話者数に対する LSD（dB）

	AE50	AE100	JDGMM	MFCC-DNN	SPEC3000	SPEC100
mix2	4.60	**4.52**	–	5.33	4.61	4.91
mix4	4.49	**4.44**	–	5.39	4.48	4.81
mix6	4.55	**4.47**	–	5.35	4.51	4.70
mix8	4.55	4.48	–	5.41	**4.47**	4.66
average	4.55	**4.48**	5.65	5.37	4.52	4.77

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 5.7 と図 5.8 では，それぞれ小規模コーパス，大規模コーパスを用いた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 5.3 や表 5.4 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た高品質な音声を作成できることがわかる．しかし，MFCC-DNN については訓練話者数が 2 人より 8 人の方が結果が良いという傾向は見られなかった．手法間の差については，表 5.3 や表 5.4 の結果に反し，AE100 よりも SPEC3000 の MOS 値が高

表 5.4　大規模コーパスにおける訓練話者数に対する LSD（dB）

	AE50	AE100	JDGMM	MFCC-DNN	SPEC3000	SPEC100
mix2	4.38	**4.35**	–	5.26	4.40	4.43
mix4	4.32	**4.28**	–	5.11	4.29	4.30
mix6	4.30	**4.26**	–	5.08	4.33	4.35
mix8	4.32	**4.27**	–	5.12	4.36	4.31
average	4.33	**4.29**	4.87	5.14	4.34	4.35

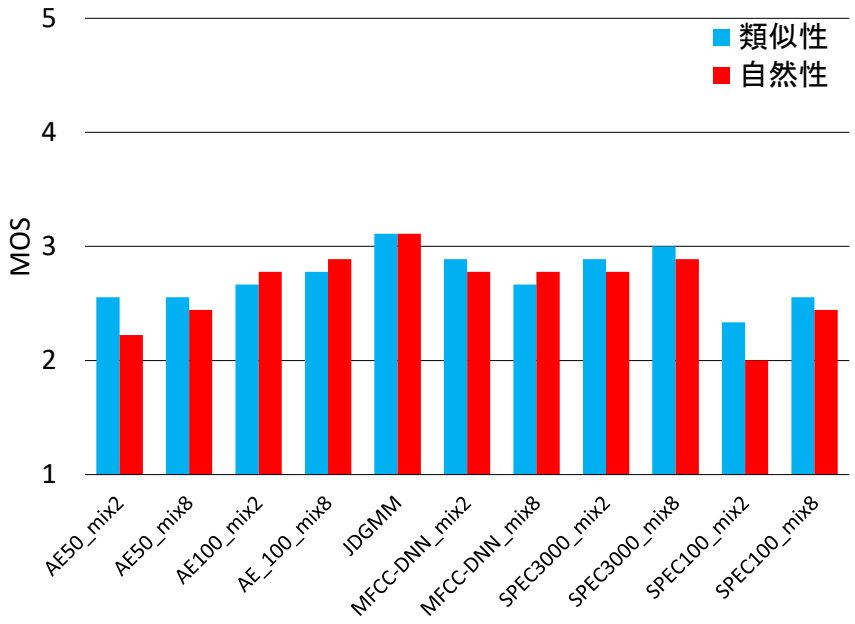

図 5.7　小規模コーパスにおける任意話者変換の主観評価結果

かった．ただし，小規模コーパスにおける実験，大規模コーパスにおける実験ともに類似性，自然性で一要因の分散分析を行ったが，有意差はなかった．また，大規模コーパスを用いた実験では，一対一変換である JDGMM より任意話者変換を行う SPEC3000 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．SPEC3000 の主観評価結果が優れていた理由として，SPEC3000 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．

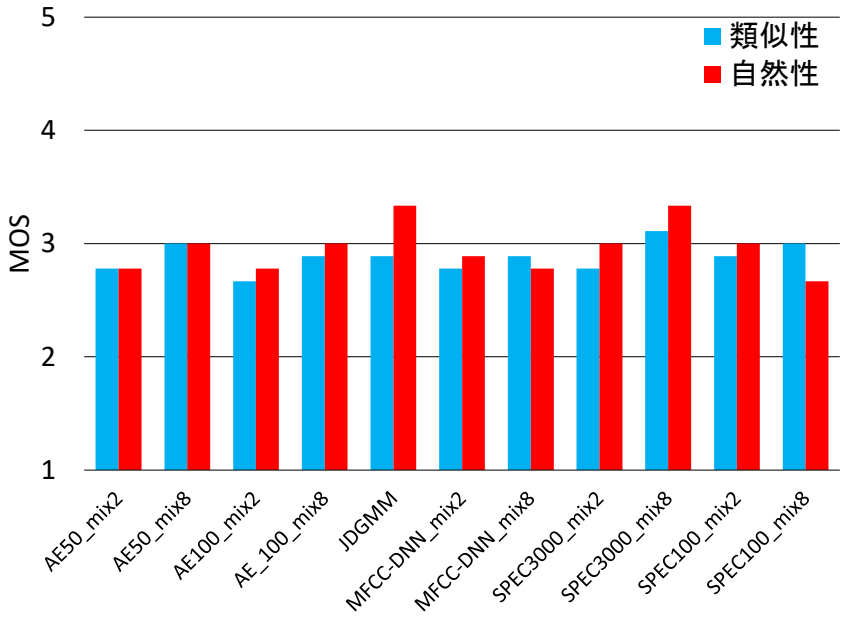

図 5.8　大規模コーパスにおける任意話者変換の主観評価結果

5.3　スパースオートエンコーダを用いた声質変換手法の評価実験

5.3.1　評価方法

評価実験では，一対一声質変換と任意話者声質変換においてスパースオートエンコーダを用いた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（YMGTtoKJM，KJMtoHM，TKtoYMGT，HMtoTK）作成し実験を行った．任意話者変換では，声質変換器の学習に 8 人の訓練話者（男性話者 3 人，女性話者 5 人）音声データを用いた．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．一対一変換と任意話者変換ともに，訓練データに 300 発話，訓練に使用していない検証用データに 50 発話，訓練および検証に使用していないテストデータに 50 発話を用いた．ここで検証用データとは，DNN 学習時に非学習データに対する誤差値を算出するために用いたデータセットのことである．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，提案手法であるスパースオートエンコーダの高次特徴量を用いた手法（SAE-DNN）と 3 つの比較手法を用いて声質変換精度比較を行った．比較手法には，ベースライン手法として GMM を用いた手法（JDGMM）[2]，MFCC を DNN を用いて変換した手法（MFCC-DNN）[4]，そして，オートエンコーダの高次特徴量を用いた手法（AE-DNN）を用いた．入出力音響特徴量として，TANDEM-STRAIGHT より 1 フレーム 40ms，フレームシフト 5ms として求めた 513 次元の対数スペクトル包絡を AE-DNN と SAE-DNN で用い，スペクトル包絡より計算された 25 次元 MFCC を JDGMM，MFCC-DNN で用いた．また，一般的に音響特徴量の変換では時系列を考慮するために，動的特徴量やマルチフレームを入力データとすることから，JDGMM では 25 次元 MFCC に動的特徴量を合わせた計 50 次元，MFCC-DNN では 11 フレームの MFCC（計 275 次元），AE-DNN および SAE-DNN では 11 フレームの対数スペクトル包絡（計 5643 次元）を入力として実験を行った．JDGMM における混合数は 64 とし，MFCC-DNN の隠れ層数を 3，隠れ層素子数は 550 とした．AE-DNN および SAE-DNN はともに，オートエンコーダの隠れ層素子数は 1000 とし，高次特徴量変換 DNN の隠れ層数を 3，隠れ層素子数を 3000 とした．活性化関数については，オートエンコーダと各 DNN は ReLU 関数，スパースオートエンコーダはシグモイド関数，学習最適化アルゴリズムはいずれも ADAM（α=0.0001）を用いた．学習回数については，MFCC-DNN は 100，AE-DNN と SAE-DNN はともにオートエンコーダは 200，高次特徴量変換 DNN は 30 とした．また，過学習を防ぐため，学習過程において検証用データによる誤差値が最も小さかったモデルを利用した．

一般的に声質変換手法の評価は客観評価と主観評価の 2 つで行われる．本実験では客観評価基準として，LSD を用いた．主観評価では，被験者（20 代男女）12 人に目的話者音声，変換音声 A，変換音声 B を試聴させ，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，それぞれどちらの変換音声が優れているかについて，「変換音声 A，変換音声 B，どちらも変わらない（N/P）」の 3 つの選択肢から 1 つ選択させた．また，一対一変換における主観評価には 4 つの変換組（YMGTtoKJM，KJMtoHM，TKtoYMGT，HMtoTK）の声質変換器を用い，各組のテストデータから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，KRT と HM を目的話者とした声質変換器を用い，各組のテストデータからランダムに選択した 1 発話を評価に用いた．被験者にはどの変換手法で変換した音声か分からないよう，変換音声をランダムに並べ試聴させた．

表 5.5　一対一変換の LSD（dB）

target	JDGMM	MFCC-DNN	AE-DNN	SAE-DNN
YMGTtoKJM	5.06	5.34	3.95	**3.92**
KJMtoHM	4.72	4.85	4.10	**4.02**
TKtoYMGT	5.04	4.71	3.81	**3.77**
HMtoTK	4.55	5.08	3.87	**3.86**
average	4.84	5.00	3.93	**3.89**

5.3.2　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 5.5 に示した．LSD 値を比較すると，スペクトル特徴量にスペクトル包絡を用いた AE-DNN と SAE-DNN は JDGMM と MFCC-DNN より高いスペクトル変換精度が得られた．これは，前節でも述べたように，JDGMM および MFCC-DNN はスペクトル特徴量に MFCC を利用しているためである．AE-DNN と SAE-DNN の LSD を比較すると，僅差ではあるが，全ての変換組で SAE-DNN が最も高い変換精度を得た．このことから，スパースオートエンコーダから得られた高次特徴量は，通常のオートエンコーダより抽象度が高く，目的スペクトル特徴量へ変換しやすい特徴量となっているのではないかと考えられる．

表 5.6 では，変換した音声の自然性と類似性について，人間の聴覚に基づき評価した結果を示した．ランダムに選択した音声を各手法を用いて変換した音声から 2 つを選び，どちらの音声がより自然性，類似性が高かったかを選択するという実験を行った．各手法の数値は得票率を表しており，N/P はどちらの音声も優劣がつけられない場合に選択された．また，p は T 検定を行った結果（p 値）を記載した．自然性，類似性ともに提案手法である AE-DNN および SAE-DNN が，ベースライン手法である JDGMM および MFCC-DNN より優れているという結果が得られた．しかし，AE-DNN と SAE-DNN の間には自然性，類似性ともに有意差は見られなかった．よって，一対一変換において AE-DNN と SAE-DNN の性能に差はないと言える．

5.3.3　任意話者変換の結果

任意話者変換実験では，JDGMM は任意話者変換を行えないため，JDGMM を他の任意話者変換手法と比較するための一対一変換手法として用いた．以下では，JDGMM は

表 5.6　一対一変換の主観評価結果（%）

	JDGMM	MFCC-DNN	AE-DNN	SAE-DNN	N/P	p
自然性	**47.9**	12.5	–	–	39.6	0.000
	18.8	–	**50.0**	–	31.3	0.001
	22.9	–	–	33.3	43.8	0.261
	–	12.5	**52.1**	–	35.4	0.000
	–	16.7	–	**52.1**	31.3	0.001
	–	–	27.1	27.1	45.8	1.000
類似性	14.6	20.8	–	–	64.6	0.428
	8.3	–	**47.9**	–	43.8	0.000
	10.4	–	–	**33.3**	56.3	0.006
	–	14.6	**47.9**	–	37.5	0.000
	–	20.8	–	**41.7**	37.5	0.028
	–	–	27.1	20.8	52.1	0.478

表 5.7　任意話者変換の LSD（dB）

target	JDGMM	MFCC-DNN	AE-DNN	SAE-DNN
to KRT	5.01	5.17	3.93	**3.89**
to HM	4.72	5.57	4.40	**4.38**
average	4.87	5.37	4.17	**4.13**

一対一変換を行った結果を記載している（変換組は HMtoKRT および KRTtoHM）.

任意話者変換における LSD の評価結果を表 5.7 に示した．表 5.5 同様，AE-DNN と SAE-DNN に比べ JDGMM と MFCC-DNN の LSD 値が大きいが，理由は前節で述べた通りである．任意話者変換においても，僅差であるものの SAE-DNN が最も高い変換精度を得た．よって一対一変換の結果と同様に，スパースオートエンコーダから得られた高次特徴量は抽象度が高く，目的スペクトル特徴量へ変換しやすい特徴量となっていると考えられる．

表 5.8 では，任意話者変換における主観評価結果を示した．自然性，類似性ともに一対一変換を行った JDGMM が他手法よりも優れていることが確認できる．このことから，今回提案した任意話者変換手法は，従来の一対一変換の精度には及ばないことが分かった．また，自然性については AE-DNN と SAE-DNN がともに，類似性については AE-DNN

表 5.8　任意話者変換の主観評価結果（%）

	JDGMM	MFCC-DNN	AE-DNN	SAE-DNN	N/P	p
自然性	**79.2**	12.5	–	–	8.3	0.000
	75.0	–	4.2	–	20.8	0.000
	66.7	–	–	20.8	12.5	0.001
	–	12.5	**54.2**	–	33.3	0.002
	–	12.5	–	**54.2**	33.3	0.002
	–	–	8.3	**37.5**	54.2	0.016
類似性	**79.2**	8.3	–	–	12.5	0.000
	54.2	–	20.8	–	25.0	0.017
	66.7	–	–	16.7	16.7	0.000
	–	8.3	**54.2**	–	37.5	0.000
	–	20.8	–	33.3	45.8	0.340
	–	–	16.7	25.0	58.3	0.488

が MFCC-DNN よりも優れていた．AE-DNN と SAE-DNN との比較では，SAE-DNN は自然性において，AE-DNN より優れていたことが確認できた．よって，SAE-DNN は，AE-DNN と比較し，任意話者変換において精度改善することができたと言える．

第 6 章

考察

6.1　変換精度

5.3.3 の任意話者変換の主観評価では，提案手法である SAE-DNN と AE-DNN 間に，自然性において有意差が見られたが，類似性では有意差は見られなかった．この理由として，任意話者変換の主観評価に用いたデータ数が少なかったことと，SAE-DNN の手法において fine-tuning を行ったことが考えられる．本節ではこれらの原因について議論する．

6.1.1　主観評価に用いるデータ数

任意話者変換の主観評価に用いたデータは，被験者 12 人に 2 組の変換組からそれぞれランダムに選択した 1 発話を試聴させたもので，計 24 の評価データによって評価を行った．一方，一対一変換の主観評価に用いたデータ数は，被験者 12 人に 4 組の変換組からそれぞれランダムに選択した 1 発話を試聴させたもので，計 48 の評価データを用いた．他の先行研究では 120 の評価データを用いているものもあり [31]，今回の主観評価実験に用いたデータが少なかったため，類似性において有意差が見られなかった可能性が考えられる．また，5.2.2 や 5.2.3 の主観評価に用いたデータ数についても，一対一変換では 18 の評価データ，任意話者変換では 9 の評価データによって統計的有意差検定を行っているため，有意差が見られ難かったと考えられる．

6.1.2　Fine-tuning の影響

本稿で提案した SAE-DNN の手法では，最後に fine-tuning を行なうものであるが，Mohammadi らの研究 [25] では，オートエンコーダを結合した DNN を fine-tuning した

ものより，fine-tuing を行わないものの方が主観評価の結果が良いと報告されている．また，5.2 の実験において，オートエンコーダを用いた提案手法 AE50，AE100 ではどちらも fine-tuning を行わないものであったが，任意話者の主観評価では，一対一変換を行った JDGMM との有意差が見られなかった．このことから，スパースオートエンコーダと高次特徴量 DNN を結合し fine-tuing を行わない手法を用いることで，スパースオートエンコーダにより得られた抽象的な特徴量が平滑化されず，話者の特徴を捉えた変換が行える可能性がある．

6.2 訓練話者数の影響

5.3.3 の任意話者変換の主観評価において，一対一変換を行った JDGMM にいずれの任意話者変換手法も及ばなかったという結果が得られた．本実験では，任意話者変換のための訓練データに複数話者（8 人）の発話データを用いたが，様々な声質の入力があることを考えると，8 人では多様な声質に対応するのは難しいと考えられる．オートエンコーダを用いた手法の評価実験では，訓練データに用いる複数話者の人数を変化させ，変換精度にどのように影響するか検証を行ったが，訓練話者が 2 人の場合は任意話者の入力に対応できず，訓練話者が 4〜8 人の時場合は 2 人の時と比較すると任意話者の入力に対応できるという結果であった．複数話者による大規模コーパスの入手が困難であるため，10 人以上の訓練話者による実験が行えていないが，任意話者の入力に対応するためには，より多くの訓練話者を用いて学習を行なう必要があると考えられる．そのため，十分な話者数の訓練データセットを用いることにより，本手法は一対一変換の JDGMM と大差ない結果が得られる可能性があると考えられる．

表 5.8 を目的話者別にまとめたものを表 6.1 に示した．目的話者が HM のときは JDGMM がどの任意話者変換手法に対しても有意差があったが，目的話者が KRT のときは有意差が見られないものもあった．これは，KRT への変換評価は入力話者に HM を用いており，KRT の変換器作成のための訓練話者 8 人の中に HM と近い声質を持つ話者がいたため，任意話者変換手法でも比較的高い精度で変換が行えたと考えられる．一方，HM への変換評価では入力話者に KRT を用いたが，訓練話者 8 人の中に KRT と近い声質を持つ話者がいなかったため，一対一変換と比べ変換精度が劣ってしまったと考えられる．そこで，入力話者と訓練話者の声質の類似度（LSD）を計測した結果を表 6.2 に示した．KRT と最も声質が近い話者は male_3 で，LSD は 5.24 であるのに対し，HM と最も声質が近い話者は female_5 で，LSD は 4.87 と KRT よりも声質の近い話者の音声データが訓練データに使われていたことがわかった．このことから，入力話者に近い声質を持つ

表 6.1　目的話者別任意話者変換の主観評価結果（%）

		JDGMM	MFCC-DNN	AE-DNN	SAE-DNN	N/P	p
toHM	自然性	**91.7**	0.0	–	–	8.3	0.000
		83.3	–	0.0	–	16.7	0.000
		75.0	–	–	8.3	16.7	0.000
		–	0.0	**75.0**	–	25.0	0.000
		–	8.3	–	**75.0**	16.7	0.000
		–	–	16.7	33.3	50.0	0.368
	類似性	**83.3**	0.0	–	–	16.7	0.000
		58.3	–	16.7	–	25.0	0.036
		58.3	–	–	16.7	25.0	0.036
		–	0.0	**66.7**	–	33.3	0.000
		–	0.0	–	**50.0**	50.0	0.003
		–	–	25.0	16.7	58.3	0.633
toKRT	自然性	**66.7**	25.0	–	–	8.3	0.042
		66.7	–	8.3	–	25.0	0.002
		58.3	–	–	33.3	8.3	0.237
		–	25.0	33.3	–	41.7	0.670
		–	16.7	–	33.3	50.0	0.368
		–	–	0.0	**41.7**	58.3	0.010
	類似性	**75.0**	16.7	–	–	8.3	0.003
		50.0	–	25.0	–	25.0	0.223
		75.0	–	–	16.7	8.3	0.003
		–	16.7	41.7	–	41.7	0.193
		–	41.7	–	16.7	41.7	0.193
		–	–	8.3	33.3	58.3	0.143

訓練話者のデータを含むコーパスを用いて声質変換器が作成することで，変換精度が向上することが期待される．よって，任意話者声質変換器を作成する際に用いる音声コーパスは，多様な声質からなる複数話者の音声データから構成されているものを用いるのが望ましいと言える．

表 6.2　KRT および HM と訓練話者間の LSD（dB）

	KRT	HM
male_1	5.59	6.32
male_2	5.63	6.23
male_3	**5.24**	5.53
female_1	6.23	5.69
female_2	6.27	5.70
female_3	5.84	5.27
female_4	5.72	5.31
female_5	5.99	**4.87**

6.3　変換時間

5.2.2 の変換時間の実験結果では，既存手法である SPEC3000 と比べ提案手法が短時間で変換することが可能になったが，音声から音声へ変換を行う場合はどの手法も 2 秒以上要してしまう．声質変換を用いた通話など，リアルタイム性を求められるアプリケーションへ用いる場合はより早い変換が必要とされる．提案手法のスペクトル変換自体は 0.4 秒程度であるのに対し，音声分析と音声合成に要する時間は約 1.9 秒であるので，音声合成および音声合成に要する時間の短縮が行われなければ，リアルタイム性を求められるアプリケーションへの応用は難しいと考えられる．本稿では，音声分析ツールとして TANDEM-STRAIGHT を用いたが，他の音声分析ツールとして SPTK[*1]や WORLD[32]（D4C edition [33]）などがあり，TANDEM-STRAIGHT に比べ合成分析に要する時間が短くなるものがあるので，アプリケーションを作成する際には，アプリケーションの制約によってどの音声分析ツールを用いるべきか検討する必要がある．

6.4　マルチフレームによる精度変化

DNN を用いた手法において，マルチフレームによる変換精度変化を表 6.3 に示した．入力に用いるフレーム数を増やすことで，LSD の精度が改善されていることが確認できる．ただし，MFCC-DNN は 5.2 でも述べたように，LSD では精度の改善ができているか

[*1] http://sp-tk.sourceforge.net/

表 6.3　フレーム数による LSD 変化（dB）

	1 フレーム	3 フレーム	5 フレーム	11 フレーム
MFCC-DNN	5.08	5.05	5.07	5.12
AE-DNN	4.12	4.08	4.09	4.01
SAE-DNN	4.15	4.08	4.04	3.97

確認できない．1 フレームと 11 フレームとの LSD 差は AE-DNN で 0.1，SAE-DNN では 0.2 程度となった．主観評価による比較を行っていないため，音声にどの程度差が生じるかは不明だが，本実験での AE-DNN と SAE-DNN の LSD 差が 0.04 で一対一変換の主観評価では特に差がなかったことから，大きな差は生じないと考えられる．入力に用いるフレーム数が増えると，変換に要する DNN の構造が複雑になり，変換に長い時間掛かる．よって，本手法において変換精度と変換時間はトレードオフの関係にあり，本手法を用いてアプリケーションを開発する場合には，アプリケーションの制約によって入力フレーム数を調整するべきだと考える．例えば，映画の吹替音声の作成などを行なうアプリケーションであれば，時間的制約がないので，11 フレームを用いて変換精度を優先できる．声質変換を用いて通話を行なうようなアプリケーションでは，リアルタイムの声質変換が必要になるため，1 フレームや 3 フレームなどフレーム数を少なくし，変換時間を短縮するよう調整が可能である．

6.5　スペクトログラム

5.3.3 で用いた，HM の発話，KRT の発話，HM の発話を KRT の声質へ JDGMM を用いて変換した音声，任意話者（HM）の発話を KRT の声質へ SAE-DNN を用いて変換した音声におけるスペクトログラムの一例をそれぞれ，図 6.1，図 6.2，図 6.3，図 6.4 に示した．図 6.1 と図 6.2 を比較すると，HM（女性話者）の発話は高周波数帯（3kHz 以上）まで強い成分が出ているのに対し，KRT（男性話者）の発話は低周波数帯（0-2kHz）に強い成分が集中している．図 6.3 と図 6.4 では，図 6.2 で見られたように，どちらも低周波数帯に強い成分が集中していることが見受けられる．このことから，どちらの手法も目的話者の特徴を捉えた変換が行えていることがわかる．また，どちらの図もほとんど差がないように見受けられるが，JDGMM で変換した音声スペクトログラムの方が，SAE-DNN で変換したものより若干低周波数帯成分が強く出ていると見て取れる．この違いが，5.3.3 における JDGMM と SAE-DNN の主観評価の差に影響したと考えられる．

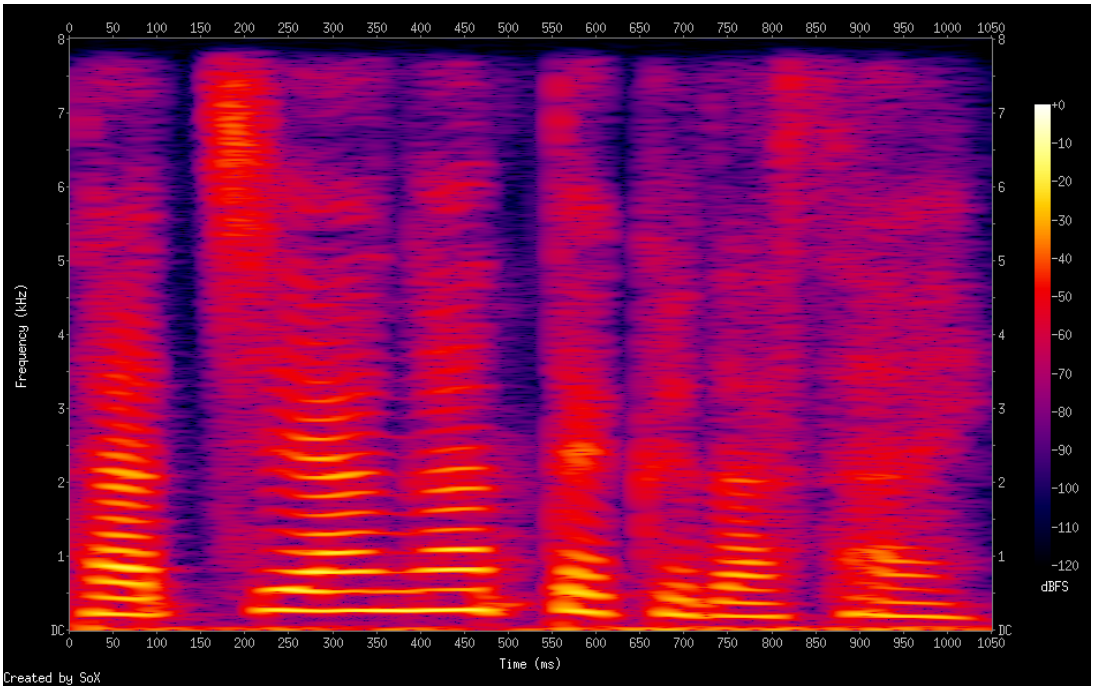

図 6.1　HM の発話におけるスペクトログラム

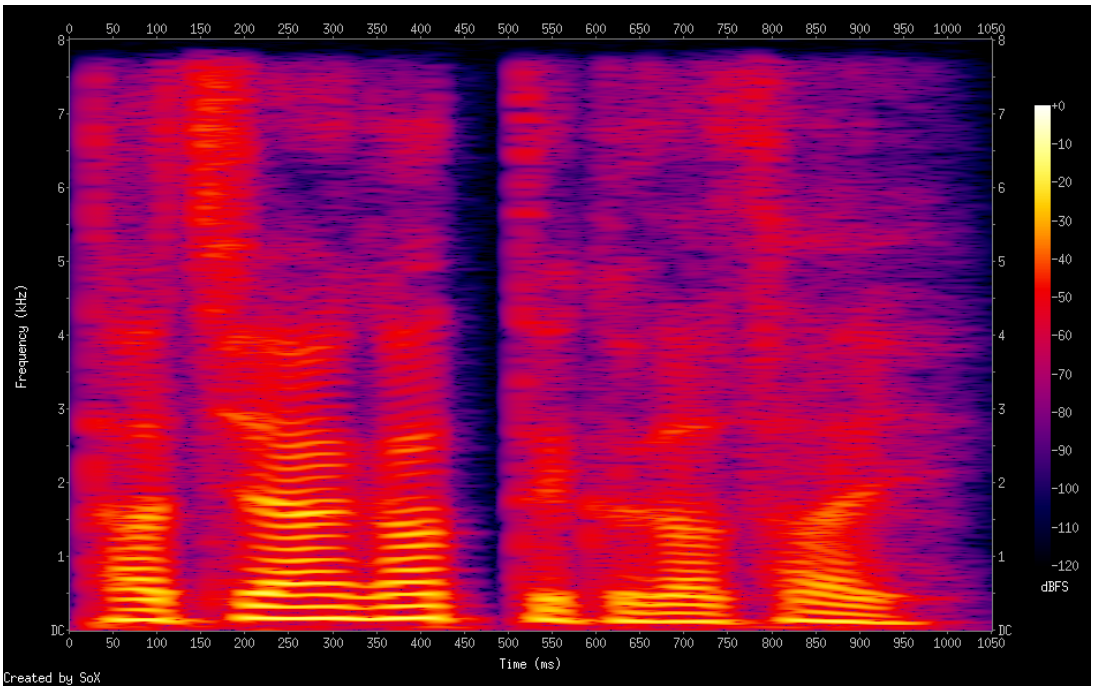

図 6.2　KRT の発話におけるスペクトログラム

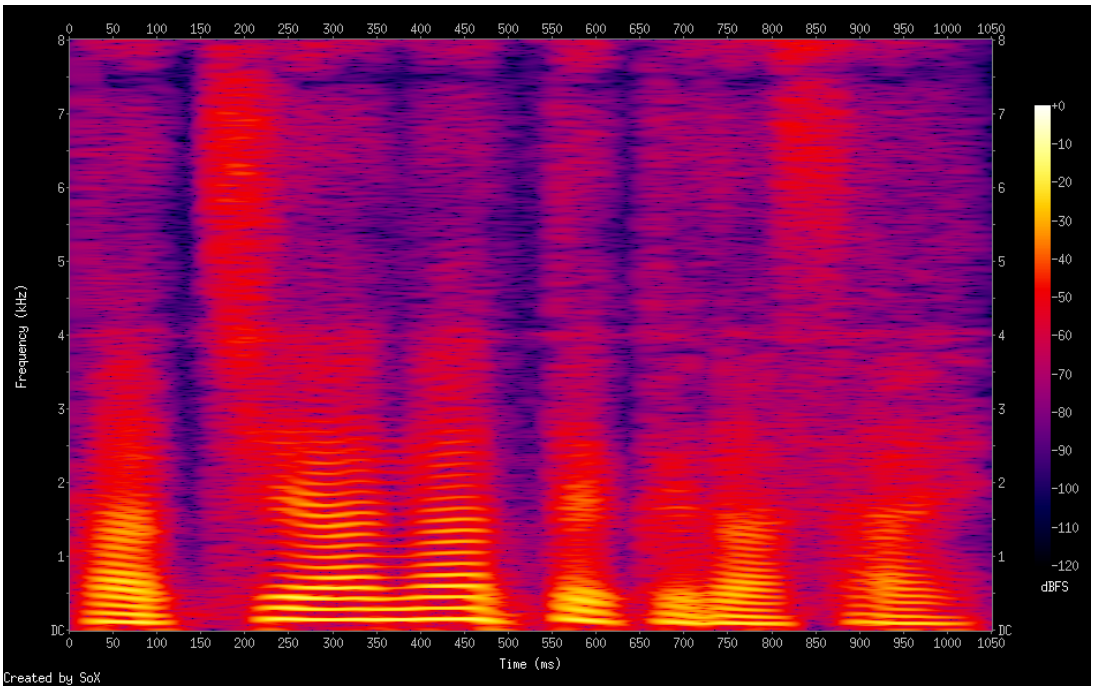

図 6.3　HM の発話を KRT の声質へ JDGMM を用いて変換した音声のスペクトログラム

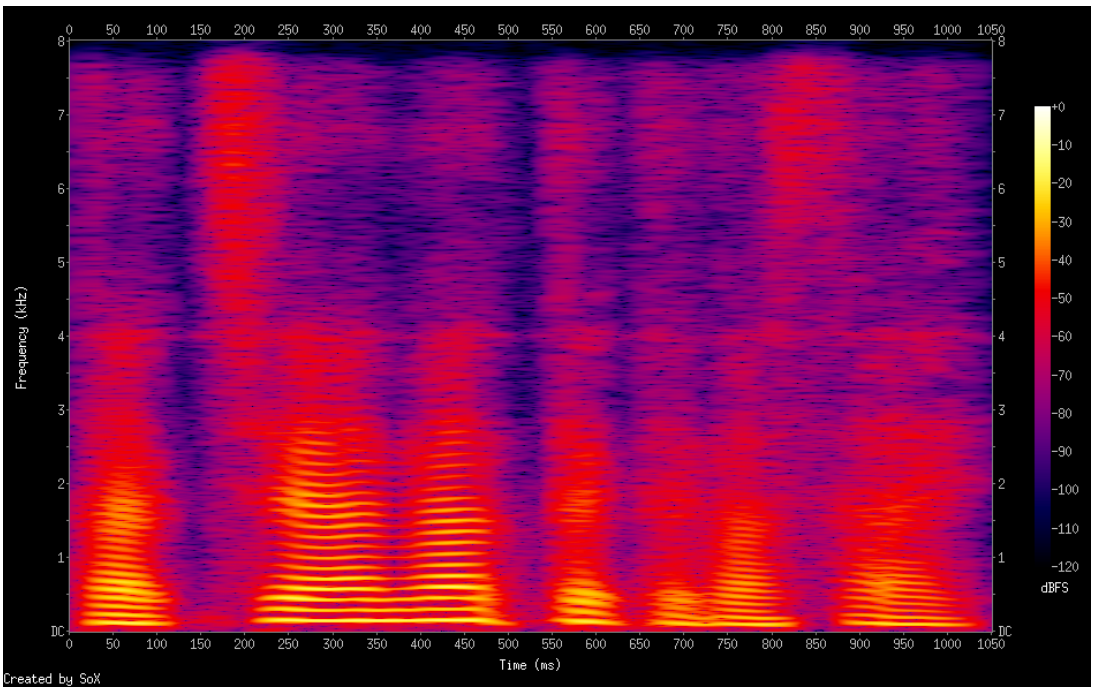

図 6.4　任意話者（HM）の発話を KRT の声質へ SAE-DNN を用いて変換した音声のスペクトログラム

第 7 章

まとめ

本研究では，訓練に用いるデータ量を低減させること，および声質変換に要する時間を短縮することを目的とし，オートエンコーダを用いた声質変換手法を提案した．また，任意話者の声質変換精度を向上させるため，スパースオートエンコーダを用いた任意話者声質変換手法を提案した．オートエンコーダを用いた声質変換手法の評価実験では，客観評価において一対一変換，任意話者変換ともに提案手法が従来手法よりも変換精度が優れていた．変換時間については，従来手法と比べ大きく短縮することができた．主観評価では，一対一変換については提案手法が優れていたが，任意話者変換では従来手法には及ばなかった．また，スペクトル包絡を変換する DNN を用いた手法が従来の GMM を用いた一対一変換手法と同程度の精度で変換が行えたことから，手法や用いる音声コーパス次第で，任意話者変換を一対一変換と大差ない精度で行えることがわかった．スパースオートエンコーダを用いた声質変換手法の評価実験では，客観評価において提案手法が最も高い精度でスペクトル変換を行った．主観評価では，提案手法は一対一変換手法の精度には及ばなかったものの，他の任意話者変換手法に比べ自然性は最も優れていたという結果が得られた．よって，スパースオートエンコーダを用いることで話者に依存しない特徴量を抽出し，その特徴量を DNN を用いて変換することで，任意話者変換の精度向上を行うことができたと考えられる．

今後の課題として，従来の一対一変換手法と差がなくなることを目的とし，任意話者の入力により対応できるよう手法を改良する．また，本手法における fine-tuning の影響を検証し，fine-tuning するべきか否かを判断する必要がある．訓練データに利用する話者数を増やすことで，様々な声質を持つ話者からの入力に対応できるか確認をし，一般的に任意話者の声質変換において，訓練データに用いる話者数は何人程度必要になるか検証を行う．同時に，任意話者変換に適する音声コーパスの設計法の検討を行いたい．また，本稿

で提案した声質変換手法を用いたアプリケーションの開発を行い，アプリケーションと本手法を総合した評価を行いたい．

付録 A

スペクトログラム

6.5 で記載しなかった任意話者（入力話者は HM）から KRT へ AE-DNN と MFCC-DNN を用いて変換された音声のスペクトログラム（図 A.1，図 A.2），および任意話者（入力話者は KRT）から HM へ変換された音声のスペクトログラム（図 A.3，図 A.4，図 A.5，図 A.6）を記載する．

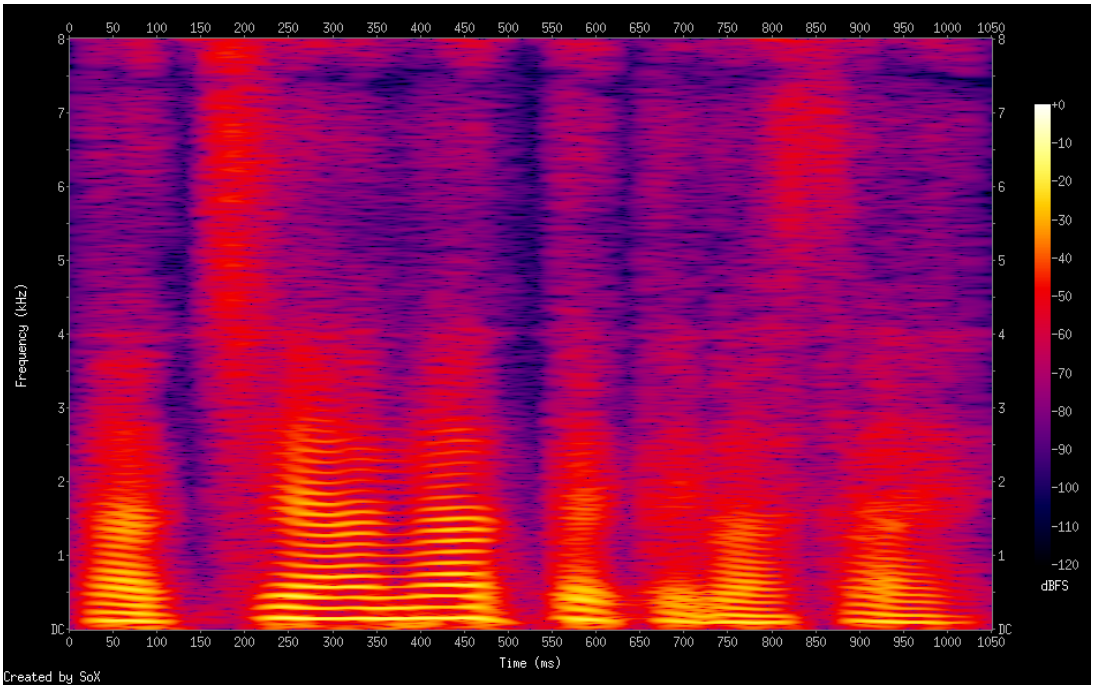

図 A.1　任意話者（HM）の発話を KRT の声質へ MFCC-DNN を用いて変換した音声のスペクトログラム

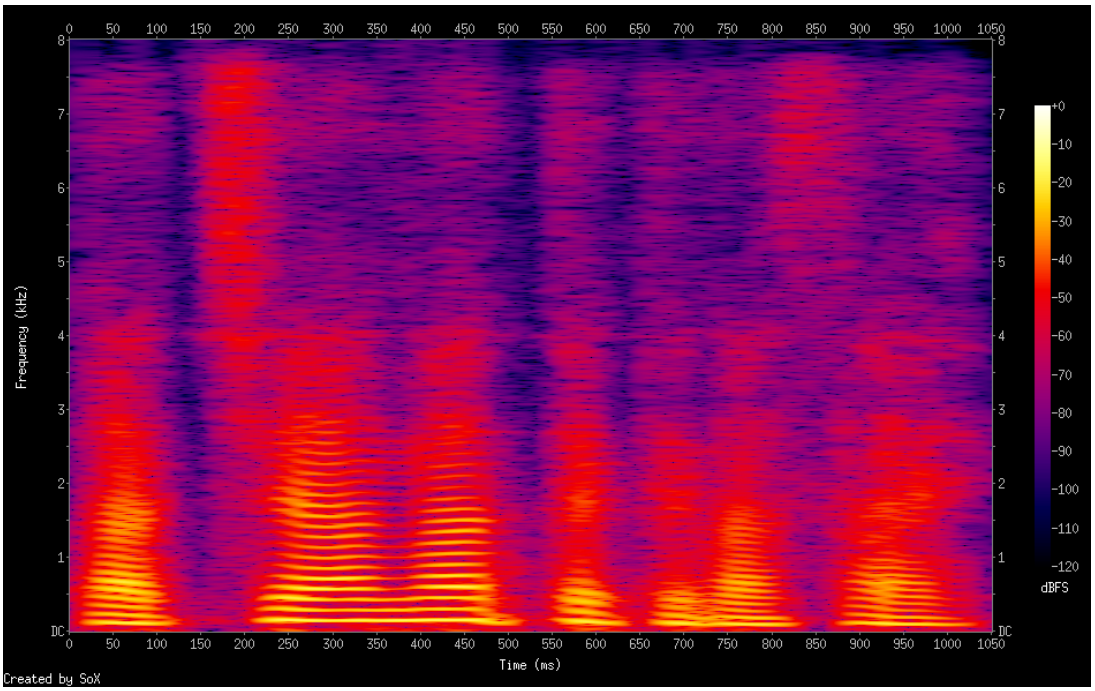

図 A.2　任意話者（HM）の発話を KRT の声質へ AE-DNN を用いて変換した音声のスペクトログラム

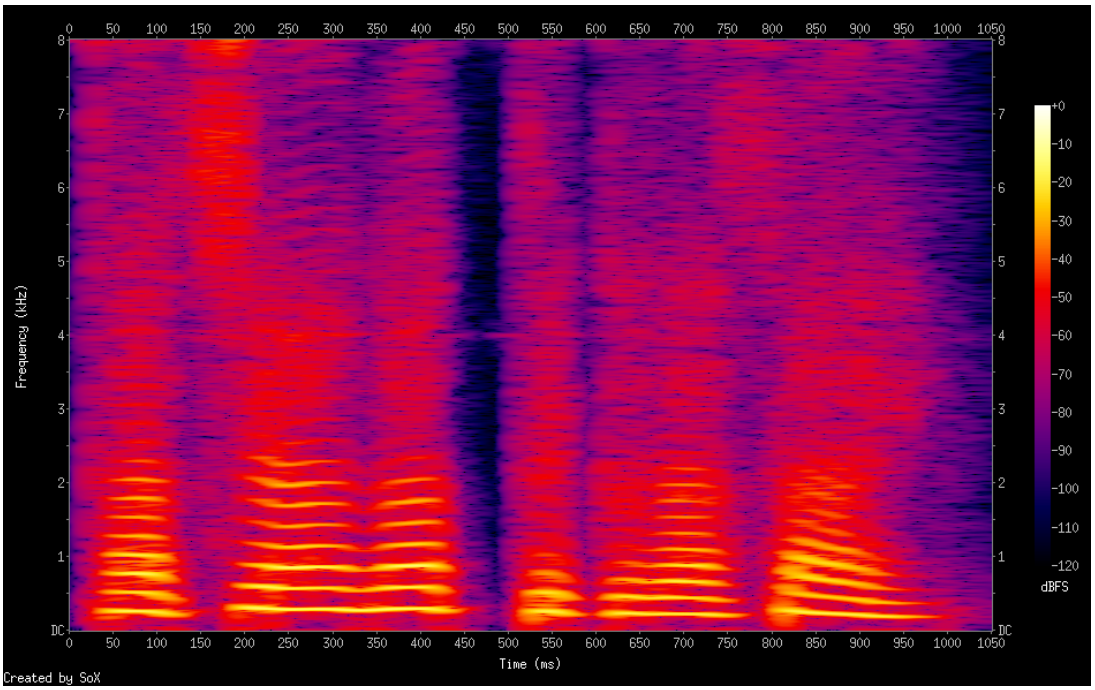

図 A.3　KRT の発話を HM の声質へ JDGMM を用いて変換した音声のスペクトログラム

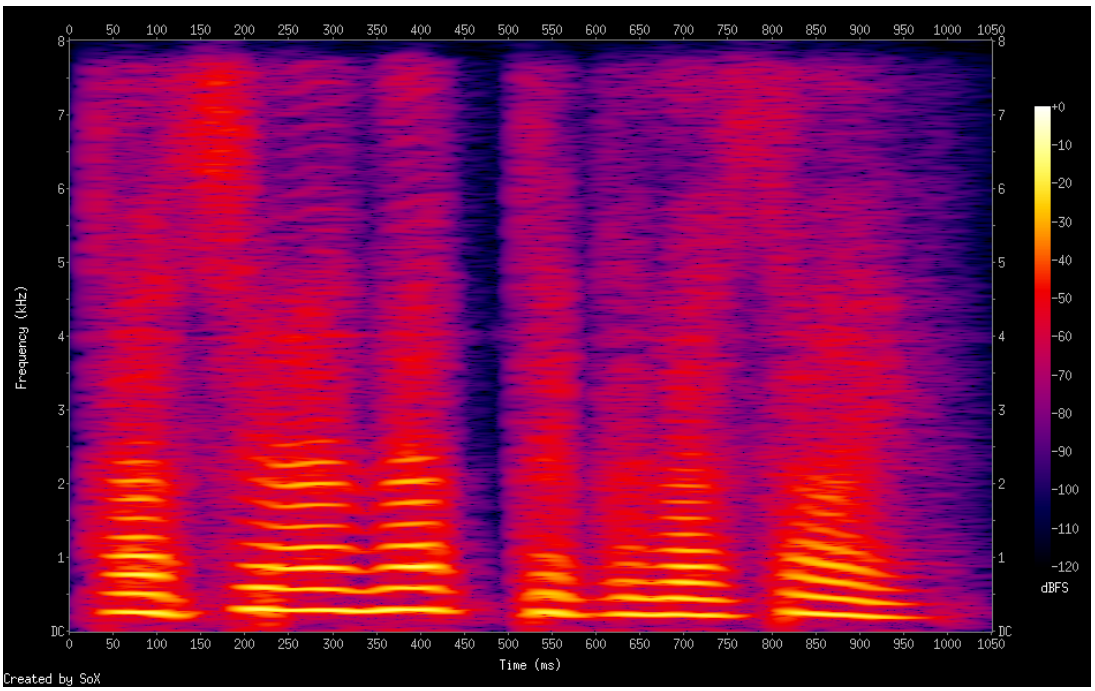

図 A.4　任意話者（KRT）の発話を HM の声質へ SAE-DNN を用いて変換した音声のスペクトログラム

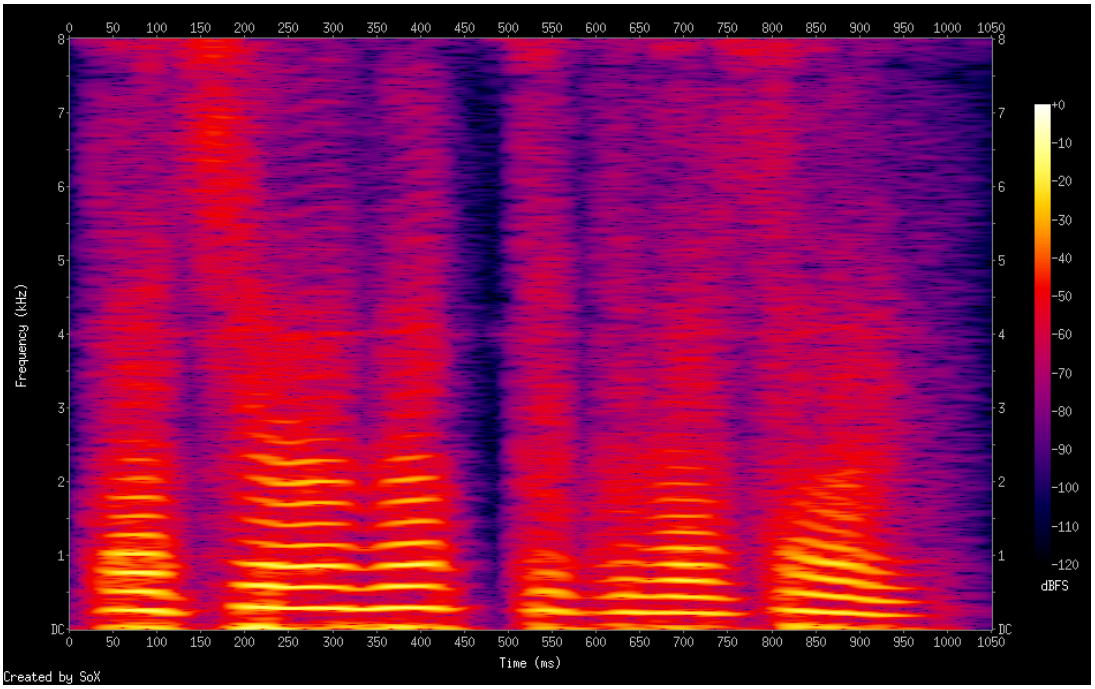

図 A.5　任意話者（KRT）の発話を HM の声質へ MFCC-DNN を用いて変換した音声のスペクトログラム

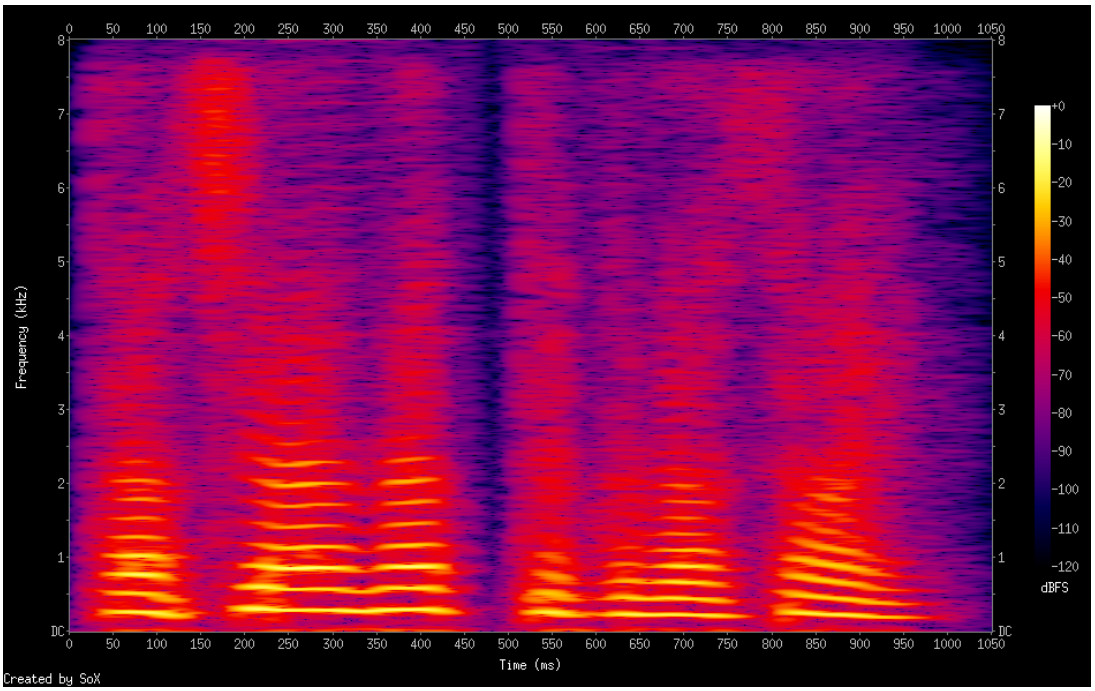

図 A.6　任意話者（KRT）の発話を HM の声質へ AE-DNN を用いて変換した音声のスペクトログラム

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, Mar 1998.

[2] T. Toda, A. W. 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. 2222–2235, Nov 2007.

[3] Yi Yang, Hidetsugu Uchida, Daisuke Saito, and Nobuaki Minematsu. Voice conversion based on matrix variate gaussian mixture model using multiple frame features. In *Proc. INTERSPEECH*, pp. 302–306, 2016.

[4] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad. Voice conversion using artificial neural networks. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 3893–3896, April 2009.

[5] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki. Voice conversion using speaker-dependent conditional restricted boltzmann machine. *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 2015, No. 1, pp. 1–12, 2015.

[6] L. H. Chen, Z. H. Ling, Y. Song, and L. R. Dai. Joint spectral distribution modeling using restricted boltzmann machines for voice conversion. In *Proc. INTERSPEECH*, pp. 3052–3056, 2013.

[7] Toru Nakashika, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. Voice conversion in high-order eigen space using deep belief nets. In *Proc. INTERSPEECH*, pp. 369–372, 2013.

[8] Z. Wu, E. S. Chng, and H. Li. Conditional restricted boltzmann machine for voice conversion. In *Proc. IEEE China Summit International Conference on Signal and Information Processing* (*ChinaSIP*), pp. 104–108, July 2013.

[9] L. Sun, S. Kang, K. Li, and H. Meng. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 4869–4873, April 2015.

[10] S. H. Mohammadi and A. Kain. Voice conversion using deep neural networks with speaker-independent pre-training. In *Proc. Spoken Language Technology Workshop* (*SLT*), pp. 19–23, Dec 2014.

[11] L. J. Liu, L. H. Chen, Z. H. Ling, and L. R. Dai. Spectral conversion using deep neural networks trained with multi-source speakers. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 4849–4853, April 2015.

[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*, Vol. 75, No. 9, pp. 5265–5285, 2016.

[13] 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, Vol. 114, No. 365, pp. 165–170, dec 2014.

[14] Feng-Long Xie, Yao Qian, Yuchen Fan, Frank K. Soong, and Haifeng Li. Sequence error (SE) minimization training of neural network for voice conversion. In *Proc. INTERSPEECH*, pp. 2283–2287, 2014.

[15] T. Toda, Y. Ohtani, and K. Shikano. Eigenvoice conversion based on gaussian mixture model. In *Proc. INTERSPEECH 2006 - Ninth International Conference on Spoken Language Processing (ICSLP)*, pp. 2446–2249, 2006.

[16] 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, Vol. 109, No. 308, pp. 55–60, nov 2009.

[17] 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, Vol. 2014, No. 1, pp. 495–496, mar 2014.

[18] 山岸順一. 音声の障がい者のための最先端音声合成技術. 情報管理, Vol. 57, No. 12, pp. 882–889, 2015.

[19] 藤本健の“DTM ステーション”. 自分の声をキャラクターの声にリアルタイム変換する SF のような技術、リアチェン voice が楽器フェアに登場！ http://www.

dtmstation.com/archives/51986665.html, (2017.1.26 アクセス).

[20] Feng-Long Xie, Frank K. Soong, and Haifeng Li. A KL Divergence and DNN-Based Approach to Voice Conversion without Parallel Training Sentences. In *Proc. INTERSPEECH*, pp. 287–291, September 2016.

[21] T. Nakashika and Y. Minami. Speaker adaptive model based on boltzmann machine for non-parallel training in voice conversion. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 5530–5534, March 2016.

[22] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Voice conversion through vector quantization. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 655–658 vol.1, Apr 1988.

[23] R. Aihara, T. Takiguchi, and Y. Ariki. Semi-non-negative matrix factorization using alternating direction method of multipliers for voice conversion. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5170–5174, March 2016.

[24] Ling-Hui Chen, Li-Juan Liu, Zhen-Hua Ling, Yuan Jiang, and Li-Rong Dai. The USTC System for Voice Conversion Challenge 2016: Neural Network Based Approaches for Spectrum, Aperiodicity and F0 Conversion. In *Proc. INTERSPEECH*, pp. 1642–1646, September 2016.

[25] Seyed Hamidreza Mohammadi and Alexander Kain. A voice conversion mapping function based on a stacked joint-autoencoder. In *Proc. INTERSPEECH*, pp. 1647–1651, 2016.

[26] 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. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 3933–3936, March 2008.

[27] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, Vol. 313, No. 5786, pp. 504–507, 2006.

[28] A. Ng. Sparse autoencoder. In *CS294A Lecture notes*, 2011.

[29] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. International Conference on Machine Learning* (*ICML*), pp. 807–814. Omnipress, 2010.

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization.

In *Proc. International Conference for Learning Representations* (*ICLR*), 2015.

[31] L. H. Chen, Z. H. Ling, L. J. Liu, and L. R. Dai. Voice conversion using deep neural networks with layer-wise generative training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 22, No. 12, pp. 1859–1872, Dec 2014.

[32] Masanori MORISE, Fumiya YOKOMORI, and Kenji OZAWA. World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, Vol. E99.D, No. 7, pp. 1877–1884, 2016.

[33] Masanori Morise. D4c, a band-aperiodicity estimator for high-quality speech synthesis. *Speech Communication*, Vol. 84, pp. 57 – 65, 2016.

謝辞

本研究を遂行するにあたり，ご多忙の中，終始適切かつ丁寧なご指導を下さった田原康之准教授，大須賀昭彦教授，石川冬樹客員准教授に深謝致します．折原良平客員教授，清雄一助教にはご多忙の中，週１回のゼミを初めとして熱心な研究指導を賜り，貴重な勉学の機会を与えて下さったことに厚く御礼申し上げます．また，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／東京大学の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，大須賀・田原研究室の皆様に感謝致します．さらに，本研究はソリッドスフィア株式会社との共同研究であり，データセットの提供および研究の議論，御意見を頂いたソリッドスフィア株式会社小島圭介氏に感謝の意を表します．

博士課程中林寿文氏には，ソフトウェア工学か声質変換のどちらを研究とするかで心がゆれていた時，相談に乗っていただいたほか，研究についての重要な議論や貴重な意見をいただきました．博士課程江上周作氏には，研究生活における助言を始め，研究の進め方から日々の雑談までお世話になりました．同期・後輩諸氏のおかげで，楽しい研究生活をおくることができました．また実験協力もしていただきました．研究室の皆様にこの場をりて心からお礼申し上げます．

研究業績

国際会議

1. **Yusuke Sekii**, Ryohei Orihara, Keisuke Kojima,Yuichi Sei, Yasuyuki Tahara and Akihiko Ohsuga: Fast Many-to-One Voice Conversion Using Autoencoders, Proc. International Conference on Agents and Artificial Intelligence (ICAART 2017), (2017.2)

論文誌

1. 関井祐介，折原良平，小島圭介，清雄一，田原康之，大須賀昭彦：スパースオートエンコーダによる次元圧縮を用いた任意話者の高品位声質変換，情報処理学会論文誌「エンタテインメントコンピューティング」特集, **査読中**

査読付き国内シンポジウム

1. 関井祐介，折原良平，小島圭介，清雄一，田原康之，大須賀昭彦：オートエンコーダを利用した複数話者の声質変換，合同エージェントワークショップ＆シンポジウム 2016 (JAWS 2016), pp.270-277 (2016.9), **Long 採択**（Long 採択率 43%） **優良論文賞，優秀発表賞**

研究会

1. 関井祐介，折原良平，小島圭介，清雄一，田原康之，大須賀昭彦：Deep Learning を利用した任意話者の声質変換，情報処理学会研究報告. SLP, 音声言語情報処理，

Vol. 112, No. 3, pp. 1-6 (2016.7)