声質変換技術によるデータ拡張を利用した話者認証モデルの学習

声質変換技術によるデータ拡張を利用した話者認証モデルの学習

平塚喬* 小高知宏** 黒岩丈介** 白井治彦*** 諏訪いずみ****

Learning Speaker Authentication Models Using Data Augumentation with Voice Conversion Technology

Kyou HIRATSUKA*, Tomohiro ODAKA, Jousuke KUROIWA**, Haruhiko SHIRAI***, Izumi SUWA******

(Received September 29, 2023)

In this paper, we tried to improve the authentication performance by increasing the training data of the speaker authentication model using voice conversion technology. We used one of the deep learning speaker authentication models, ”x-vector”, and increased the amount of data by incorporating data created by statistical voice conversion techniques as new speaker speech data when training the model. From the experiment, a comparison between models that incorporated data created by the voice conversion technology and those that did not, confirmed that models that incorporated data created by the voice conversion technology performed better, albeit slightly. This suggests that increasing the amount of data using voice conversion techniques is effective in learning speaker authentication models, but its impact was limited; therefore, more detailed study of data generation by voice conversion is needed in the future.

Key words **:** Speaker Authentication, Deep Learning, X-vector, Voice Conversion

1. 緒言

本研究では，音声を用いた生体認証方式である話者認証において，音声の声質を変換する声質変換技術を利用して生成した音声によるデータ拡張を行うことによって，話者認証を実現するためのモデルの性能向上を試みた．

話者認証とは，人間の音声データの中に含まれる音声の話者の固有特徴を用いる認証方式である．話者認証では，音声の入力だけで，システムに登録されているユーザーの認証を行うため，認証を行うユーザーに対する負担が比較的小さく，音声データの入力についてはマイクがあれば容易に実装することできるため，認証システムの導入コストも比較的低いとされ，その活用が期待されている．

その話者認証を実現するための手法は様々なものが提案されているが，近年では，他の分野でも用いられている深層学習を利用した方法が注目を集めている．深層学習を利用した手法では，従来の標準的であった手法と比べて高い性能が発揮するものもあり，その研究と標準化が進んでいる．一方で，深層学習を利用した手法では，モデルの学習に利用する音声データが大量に必要であり，学習に利用するデータの量が性能にも大きな影響をもたらしていることが知られている．

そこで，本研究では，深層学習を利用した話者認証手法の性能向上を目的とし，声質変換技術を用いてモデルの学習データを拡張することによって，深層学習を利用した手法における話者認証のためのモデルの性

*大学院工学研究科 知識社会基礎工学専攻
*Fundamental Engineering for Knowledge-Based Society, Graduate School of Engineering
**知能システム工学講座
**Department of Human and Artificial Intelligent Systems
***工学部 技術部
***Technical Division
****仁愛女子短期大学 生活科学学科
****Jin-ai Women’s College

能を向上させることを試みた．ここで，声質変換技術とは，音声に含まれる音響特徴を変換し，変換前の音声とは異なる声質をもった音声を生成する手法である．本研究では，声質変換技術を用いることによって，モデルの学習データとなる人間の音声データから，元の音声データとは異なる声質を持った音声データを生成し，その音声データを深層学習の学習データとして組み入れて話者認証用のモデルを学習した．そして，学習したモデルについて，声質変換によって生成した音声データを用いずに学習したモデルと比較しながらその性能を検証し，声質変換技術によって生成した音声データを利用してデータ拡張を行うことによってモデルの性能を向上させることができるかを確かめた．

2. 話者認証の手法とその問題点

2.1 話者認証の概要

話者認証は，人間が発声する音声を生体情報データとして利用し，入力された音声を発した話者が，システムに登録されたユーザーの音声か否かを識別することによってそのユーザーの認証を行うという認証方式である．この認証方式では，入力された音声データを解析して，その音声に含まれている話者固有の特徴を何らかの手法によって抽出し，元々登録されていた音声データからも同じ手法で特徴を抽出する．そして，抽出されたそれぞれの固有特徴について比較し，同一のものであるかを判断することによって，入力された音声の話者が元々登録されていた話者と同一話者か否かを決め認証を行う．話者認証を実現するための手法としては，機械学習によって構成される i-vector[1] に基づく手法，深層学習に基づく手法など，様々なものが提案されており，それぞれ高い性能を示している．

一方，近年では，深層学習に基づく話者認証手法の 1 つである x-vector[2] と呼ばれる手法が注目を集めており，盛んに研究が行われている．

2.2 x-vector による話者認証

本研究では，話者認証を実現するための手法として，x-vector に基づく手法を用いた．この手法では，図 1 のような特徴抽出部と識別部に分かれた深層ニューラルネットワーク (DNN) を用いる．

DNN の学習においては，まず，多数の話者によって構成される多数の音声データを用意し，それらの音声から音響特徴量を抽出し，発声した話者を識別するための話者ラベルを付与して学習データとする．次に，用意した学習データと DNN から，話者ラベルを教師データとする教師あり学習を行い，モデルが音声データの

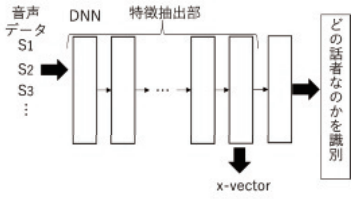

図 1 x-vector の概略図

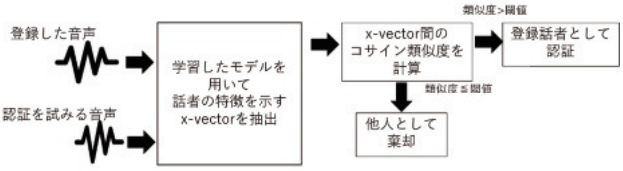

図 2 本研究における話者認証の流れ

発声話者を識別することができるように学習を行う．この学習を行った DNN の特徴抽出部を利用することで，可変長の音声データから話者の特徴を効果的に示した x-vector と呼ばれる固定長のベクトルを抽出できる．これにより，音声データから抽出された x-vector を比較することによって，話者認証を行うことができる．

本研究における x-vector モデルに基づく話者認証の流れを図 2 に示す．本研究では，図 2 のように，登録されている音声 (本人となる話者の音声) と，認証を試みる話者の音声を，学習した x-vector モデルにそれぞれ入力することによって x-vector を抽出し，抽出された x-vector を比較して類似度を求め，閾値によって本人か否かを区別することで認証を行うことを想定する．ここで，本研究において x-vector を比較するための類似度にはコサイン類似度を用いた．

2.3 x-vector の問題点と本研究のアプローチ

x-vector による話者認証は，i-vector などの従来の他の手法を上回る精度も報告されており，現在の話者認証における有用な方法の一つである．一方で，深層学習を用いた手法では，学習データが重要であり，高い性能を発揮するためには，学習を行うためのデータが大量に必要となる．x-vector による手法でも，深層学習を用いているため，x-vector の性能を高めるためには異なる複数の話者が発声した多数の発話音声データから構成される x-vector 用の学習データが大量に必要となるが，それらを準備することは容易ではない．そこで，本研究では，x-vector モデルによる話者認証の性能向上を目的とし，音声の非言語情報を変換することができる声質変換を用いて，既存の音声から新しい音声データを生成し x-vector モデルの学習データを拡張することを行った．

3. 声質変換によるデータの拡張

3.1 声質変換の概要

音声には，文字で伝えられるような言語情報と文字で伝えることのできない非言語情報の両方が含まれている．声質変換とは，音声の中に含まれる音の高さや音色といった音響特徴を変換することによって言語情報を維持したままで，非言語情報を変換する技術である．声質変換では，話者性のように，身体的制約が大きく，発声者による制御が困難な情報でさえも，自在に制御できるとされる [3].

声質変換を行うためには，声質変換を行う変換元となる元話者と変換先となる目標話者の音声データを用いて，声質変換を行うためのモデルを事前に学習しておく必要がある．この時，声質変換の方法は，元話者と目標話者の音声データの関係性によって,「パラレル声質変換」と「ノンパラレル声質変換」に分けられる．パラレル声質変換とノンパラレル声質変換の違いについては以下の表 1 の通りである．

表 1 パラレル声質変換とノンパラレル声質変換の違い

	パラレル声質変換	ノンパラレル声質変換
声質変換に利用する音声データ	同じ発話内容の音声データ (パラレル音声データ)	同じ発話内容ではない音声データ (ノンパラレル音声データ)
利点	より高い精度で簡潔に声質変換することができる	音声データの準備が比較的容易
欠点	音声データの準備が比較的難しい	パラレルの場合よりも手順が複雑で精度が低くなる

表 1 のように，パラレル声質変換の場合では，元話者と目標話者が同じ発話内容の音声データ (パラレル音声データ) を利用し，ノンパラレル声質変換の場合では，同じ発話内容ではない音声データ (ノンパラレル音声データ) を利用して声質変換を行う．ノンパラレルによる方法では，パラレルよりもデータの準備が容易であり，簡単に構築できるという特徴があるがパラレルによる方法では，ノンパラレルと比べると制約がある分，より高い精度で声質変換することができる．

本研究では，声質変換によるデータの拡張が可能であることを実証するために，より高精度な手法であるパラレル声質変換の方法を用いて声質変換を行う．

3.2 本研究での声質変換の方法

本研究では，パラレル音声データを基にしたパラレル声質変換の中でも，混合正規分布モデル (Gaussian Mixture Model, GMM) を用いた最尤系列変換法によって行われる統計的声質変換と呼ばれる手法で音声データの声質変換を行った．この GMM を用いた統計的声質変換の流れについて図 3 に示す．

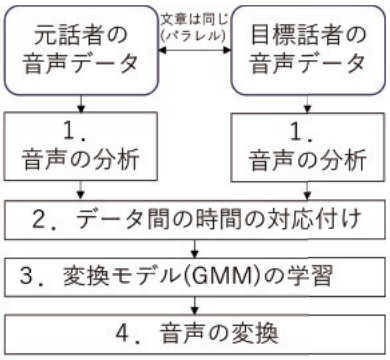

図 3 統計的声質変換の流れ

以下に，図 3 に示したこの手法による統計的声質変換の流れの詳細を示す．

1. 音声の分析
   用意したパラレル音声データに対して，それぞれ分析を行い，音の高低に対応する基本周波数と，音色に対応するメルケプストラムを抽出することを行う．そして，抽出された基本周波数とメルケプストラムから変換処理で必要になる話者依存統計量系列を推定し取得する．

2. データ間の時間の対応付け
   文章が共通しているパラレル音声であっても，時間的に完全に対応しているわけではないため，時間的に対応するように前工程で取得した特徴量系列を変換する．

3. 変換モデル (GMM) の学習
   時間的に対応するように変換した特徴量系列を学習データに用いて，元話者の特徴量と目標話者の特徴量の結合確率密度関数をモデル化する GMM を学習する．

4. 音声の変換
   学習した GMM を用いて，元話者の音声を目標話者の音声の声質への変換を行う．この時，変換を行うのは基本周波数とメルケプストラムだけである．基本周波数は自然対数を取り，線形変換を行う．メルケプストラムは学習した GMM を用いて，結合確率密度関数が最大化されるように変換を行う．これらの変換によって，変換元となった話者の音声が，言語情報はそのままに，目標話者の声質を持つように変換される．

なお，本研究では，sprocket[4] という統計的声質変換のためのソフトウェアを利用して，この声質変換を行った．

3.3 声質変換によるデータ拡張

本研究では，声質変換技術を用いてデータ拡張することで，x-vector モデルに利用する学習データに含まれる音声データの総数と話者数を増やすことを目指した．本研究における声質変換を用いたデータ拡張の流れについて図 4 に示す．

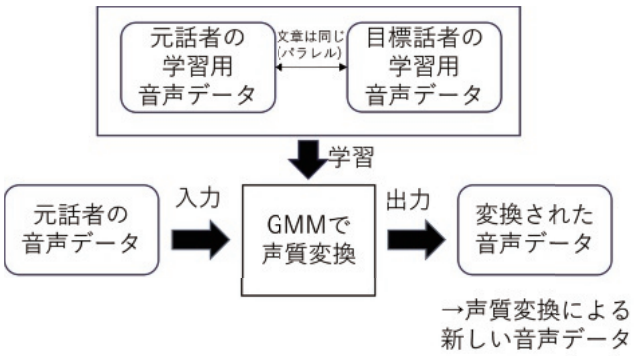

図 4 声質変換を用いたデータ拡張の流れ

図 4 のように，声質変換を行う際には，まず，変換元となる元音声と，変換先となる目標音声をパラレル音声データとして実際の人間が発声し録音した音声データのコーパスから用意した．その上で，目標音声の話者の声質に変換することを目指して，元音声の声質を目標音声話者の声質へと変換する GMM を元音声データと目標音声データのパラレル音声データから学習した．そして，学習した GMM を利用して元音声に対する声質変換を行い，変換された音声データを作成した．この声質変換された音声データは，元音声をベースにしたもので目標音声の話者の声質に近い声質を持っているものとなるが，本研究では，この音声データについては元音声の話者でも目標音声の話者でもない全く別の新たなる話者による音声データと考えることとした．この方法によって，学習データから新たなる話者の音声データを増やし，声質変換によるデータの拡張を行う．

4. 実験

本研究では，前章で記述した声質変換によって生成した新しい話者の音声データを x-vector の学習データに組み入れて学習を行った場合の有効性を調査するために，声質変換によって生成した音声データを学習データに加えて x-vector モデルの学習を行い，声質変換によって生成した音声データを学習データに加えずに学習したモデルとの間でその認証性能を比較するという実験を行った．

4.1 実験に用いるデータセット

実験では，日本語音声のデータセットである「JVS コーパス [5]」と「つくよみちゃんコーパス [6]」に含まれる音声データを x-vector モデルの学習及び検証を行うための音声データのベースとして用いた．

JVS コーパスは声優・俳優などのプロの日本語話者 100 人による合計 30 時間の音声データセットであり，それぞれの話者について，以下の表 1 に示す 4 つのサブコーパスから構成される 150 発話の音声データが含まれている．

表 2 JVS コーパスの詳細

サブコーパス名	発話数 (個)	説明
parallel100	100	パラレル音声
nonpara30	30	ノンパラレル音声
whisper10	10	ささやき音声
falset10	10	裏声音声

これらの音声は，wav 形式で提供されており，サンプリング周波数は 24kHz である．

また，つくよみちゃんコーパスには JVS コーパスに含まれる parallel100 と同一台本で発声されており JVS コーパス内に含まれない話者によって別環境で収録されたサンプリング周波数 96kHz，wav 形式での 100 個の音声データが含まれている．

本研究では，モデルの学習およびテストの両方をこれらのデータセットで全て行うために，学習用データとテスト用データで表 3 のように分割を行った．JVS コーパスに含まれる 100 人の話者のうち 90 人分 13500 発話を基本の学習用データとして利用し，JVS コーパスに含まれる残りの 10 人の話者のうち parallel100 だけの 100 発話およびつくよみちゃんコーパスの全発話を合わせた 11 人分 1100 発話をテスト用データとした．

表 3 データセットの分割

	話者数 (人)	発話数 (個)	説明
学習データ	90	13500	JVS コーパスの 90 人分の全音声
テストデータ	11	1100	JVS コーパスの 10 人分の Parallel100 + つくよみちゃんコーパス全音声 (1 人分)

4.2 声質変換による学習用データの拡張

モデルの学習データについては，学習用に分割した JVS コーパスの 90 人分のデータの中から声質変換を行うことで新しい話者の音声データを作成してデータ拡張を行った．

実験では，sprocket を用いて JVS コーパスの parallel100 に含まれるパラレル音声データを声質変換する

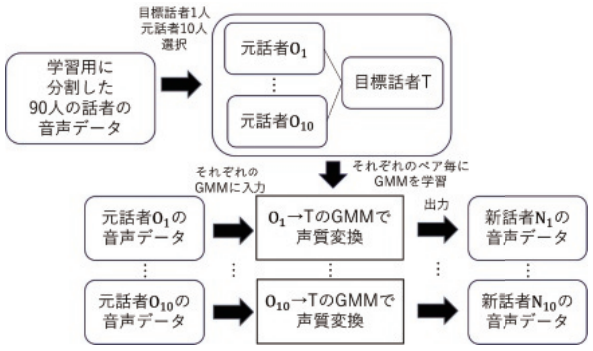

図 5 実験における声質変換によるデータ拡張の流れ

ことで，声質変換による新しい話者の音声データを作成した．この声質変換によるデータ拡張の流れについて図 5 に示す.

図 5 のように，まず，学習データとして利用するために分割した 90 人分の JVS コーパスの中から 1 人を選択しそれを声質変換先となる目標音声の話者とした．次に，90 人のうち，目標音声の話者として選択した話者以外の残りの 89 人の中から 10 人を選択しそれを声質変換元となる元音声の話者とした．これらによって選ばれた元音声話者 10 人と目標音声話者 1 人から (元音声話者，目標音声話者) として 10 組のペアを作りそれぞれの parallel100 の音声から GMM を 10 組分学習することを行った．そして，学習した GMM を用いて，元音声話者の音声データに対して，声質変換を行うことで声質変換によって生成された新しい音声データを作成した.

この時，sprocket の仕様に合わせるため，声質変換のために利用する元音声話者と目標話者の両方の音声データについては，サンプリング周波数を 24kHz から 48kHz へと音声処理ライブラリである librosa[7] を用いてアップサンプリングした上で，学習と声質変換を行った．また，声質変換をするための GMM の学習データには parallel100 に含まれる 100 個の発話の中から 50 発話分だけを利用した．声質変換に関しては，GMM の学習データに利用したものも利用しなかったものも含む parallel100 に含まれる全ての音声について，変換を行った．これによって 10 人分 1000 発話の新しい音声データを生成し x-vector モデルの学習データに組み入れることでデータの拡張を行った.

4.3 x-vector モデルの学習

本研究では，学習用として分割した 90 人 13500 発話の音声及び声質変換で生成した 10 人 1000 発話を利用して 2 つの x-vector モデルの学習を行った．このうち，声質変換で生成した音声 10 人 1000 発話についてはサンプリング周波数が 48kHz であったため，librosa を用いて，48kHz から 24kHz にダウンサンプリングを行い，学習データの全ての音声のサンプリング周波数は 24kHz に統一している．モデルの学習には，音声関連の機械学習ライブラリである Speechbrain[8] を用いた．また，一般的なデータ拡張の一種として部屋の残響ノイズなどが含まれている RIRNOISE[9] も利用した．学習回数は 15Epoch とし，学習データ以外については，全モデルで統一して学習を行った．以下に各モデル毎に異なる部分について示す.

- モデル 1
  90 人 13500 発話の音声だけを用いて学習を行った.
- モデル 2
  90 人 13500 発話の音声に加えて，声質変換で作成した 10 人分 1000 発話も加えた 100 人 14500 発話で学習を行った.

4.4 x-vector モデルの評価方法

学習したモデルに対して，事前に分割したテスト用データを用いて，性能を評価するための実験を行った．評価実験では，11 人 1100 個のテスト用データから作成される全パターンである 1100 × 1100=1210000 ペアを，各モデルに入力し，各モデル毎に，それぞれの音声の x-vector のコサイン類似度の計算を行った．この時，つくよみちゃんコーパスのデータに関しては，各モデルに入力する前に，librosa を用いてダウンサンプリングを行い，テストデータに関してもサンプリング周波数を 24kHz に統一している．その後，各モデル毎に得られた 1210000 個の類似度をペアの音声が両方とも同一話者である場合と，同一話者ではない場合の 2 つのパターンにデータを分割した.

ここで，ペアの音声が両方とも同一話者である場合のデータは，システムに登録されていた本人が認証を試みたデータとし，同一話者ではない場合のデータは，システムに登録されていない他人が認証を試みたデータとした．そして，それぞれのデータに対して閾値を与え，類似度が閾値よりも低い場合には認証が拒否され，閾値よりも高い場合には認証が受理されたと考えた.

その上で，同一人物である場合のデータからは，閾値よりも低い値の割合を計算することによって，本人であるのに認証されない確率である本人拒否率 (False Rejectance Rate , FRR) を導出した．また，同一人物ではない場合のデータからは，閾値よりも高い値のデータの割合を計算して他人であるのに認証される確率である他人受入率 (False Acceptance Rate , FAR) を導出した．これらの計算を与える閾値を変動させながら繰り

返して，閾値毎にそれぞれのモデルの FAR と FRR の結果を得た．

計算によって得られた結果については横軸に FRR を，縦軸に FAR を置き，閾値を変動させることによって描写して作られる両対数グラフである Detection Error Tradeoff(DET) 曲線 [10] を図示してそれぞれのモデルの性能を比較するということを行った．また，FRR と FAR が一致する際のエラー率である等価エラー率 (Equal Error Rate , EER) についてもそれぞれのモデルにおいて計算して比較するということを行った．

5. 実験結果

2 つのモデルに対して，テストデータ全てによる 1210000 ペアを入力して類似度を計算し閾値を変動させて FAR と FRR を計算した場合における DET 曲線について図 6 に示す．ここで，青色の点線は声質変換で作成した音声データを用いずに学習した「モデル 1」を，赤色の線は声質変換で作成した音声データも学習データに組み入れて学習した「モデル 2」を示している．また，線上にあるそれぞれの色の●印は，それぞれの曲線上で FAR と FRR が一致している箇所であり，EER となるエラー率を示している．EER についてはモデル 1 では 1.9%，モデル 2 では 1.8%となった．

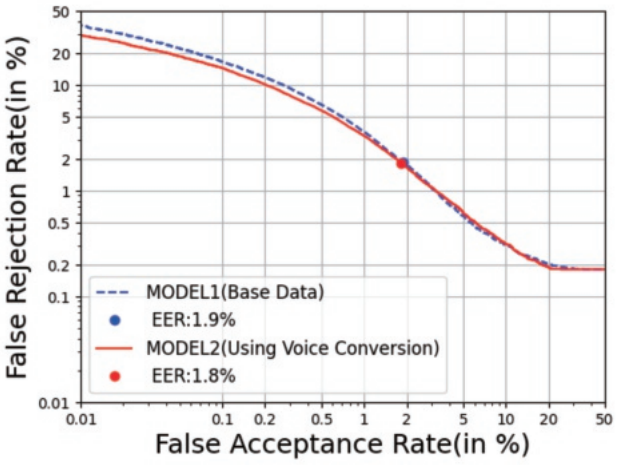

図 6 全てのデータにおける DET 曲線

図 6 の DET 曲線では，FAR の 3%程度の部分を境界にして，3%程度よりも低い部分ではモデル 2 がモデル 1 よりも図中の下側にあり，3%程度よりも高い部分では殆ど差が見られないという結果になった．また，EER の方ではモデル 1 が 1.9%に対して，モデル 2 が 1.8%であったため，差は大きくないが，モデル 2 の方が低いという結果になった．

6. 考察

図 6 では，声質変換によるデータ拡張を行ったモデル 2 が，FAR が 3%程度よりも低い部分では，声質変換によるデータ拡張を行っていないモデル 1 と比べて低い位置にあるが，FAR が 3%程度よりも高い部分ではあまり差が見られず，部分的にはモデル 2 の方が低いところもあるという結果になっている．DET 曲線では，FAR と FRR を図示しているが，これらは誤りの割合であるため両方ともより低い方が理想的であり，グラフでは下部に位置している方が望ましい．他方で，実際の認証システムにおいては，本人の認証を誤って拒絶することよりも，他人の認証を誤って受理してしまう方が，認証システムの役割上で大きな問題となるため FAR と FRR であれば，他人を誤って受理してしまう確率である FAR の方が重要視されやすいものである．よって，FAR が低い部分で，より高い性能を発揮しているモデル 2 の方が，DET 曲線による比較の上では，優れているのではないかと考えられる．

また，EER では，モデル 1 が 1.9%，モデル 2 が 1.8%となったが，EER も誤り率であり，前述の通り FAR も FRR もなるべく低い方が良いため，FAR と FRR が一致する場合の誤り率である EER も低い方が優れていると考えられるものである．したがって，EER による比較の上でもこの結果からはモデル 2 の方が優れているのではないかと考えられる．

これらの結果の分析から，声質変換によるデータを学習データに組み入れてデータ拡張を行ったモデル 2 の方が声質変換によるデータを学習データに組み入れていないモデル 1 よりも優れていると考えられる．よって，本研究の手法で声質変換を用いて生成したデータを学習データに組み入れてデータ拡張することによっても x-vector による話者認証モデルの性能が向上することが示されたと考えられる．

しかし，一方で，本研究における実験での各モデルの性能差はそれほど大きくないものとなった．このことに関しては，声質変換によるデータを利用していないものでも性能が高いことや声質変換による学習データの拡張においてその量や手法が限られていることが影響しているのではないかと考えられる．したがって，発話長が短いものや声質が近い話者でのより難しいデータにおける実験の検討や，他の声質変換手法も交えて学習データ内における声質変換によって生成されたデータをより多様化したりその割合をさらに増やした場合での検討が必要になると考えられる．

7. 結言

本研究では，深層学習による話者認証モデルは高い性能を発揮するがその性能を発揮するために必要な学習データが膨大になるという問題点に注目し，実際の音声データから話者特徴も変換可能な声質変換技術を用いて音声データの声質を変換することによって新しい音声データを生成することで，深層学習を用いた話者認証モデルの一つである「x-vector」の学習データの拡張を行い，その話者認証モデルの認証性能を向上させるということを試みた．

声質変換による音声データの拡張の上では，より高精度な声質変換を行うことができるパラレル音声を利用した GMM による統計的声質変換を利用して，変換元となる元話者の音声データを，変換先となる目標話者の音声データの声質へと変換することを行い，その変換された音声データを新しい話者の音声データとすることで，話者数と音声データ数の両方を増加させるということを行った．

学習したモデルの性能を確かめるために行った実験より，声質変換によって生成した音声データを学習データとして用いてデータ拡張をしたモデルと声質変換によって生成した音声データを学習データとして用いてデータ拡張をしていないモデルとの間での比較から，声質変換によって生成した音声データを学習データとして用いてデータ拡張をしたモデルの方が限定的ではあるがその認証性能が向上することが示された．このことから，声質変換によって生成した音声データを深層学習による話者認証モデルの学習データとして用いることができるということが示唆された．

一方で，本研究での声質変換によって生成した学習データはその生成方法が限られており，データ増量の範囲も限定的である．そのため，声質変換によって生成したデータをより多く学習データに組み入れた場合や声質変換技術の中でもノンパラレルな音声を使った場合など別の方法を使って生成したデータを用いた場合にはどのように性能が変化していくのかということなどについてはより詳しく検討する必要があると考えられる．また，本研究での手法は実際の音声データからの声質変換による新しいデータの準備とモデルの学習が分離されているが，モデルの学習をより容易にしていくためにはこれらを一挙に行うことができる新しいモデルを考案する必要性もあると考えられる．

参考文献

[1] Najim Dehak, Patrick J Kenny, Reda Dehak, et al., ”Frond-end factor analysis for speaker verification”, IEEE Transactions on Audio, Speech, and Language Processing, Vol.19, No.4, pp.788-798, 2010

[2] David Snyder, Daniel Garcia-Romero, Gregory Sell, et al., ”X-vectors: Robust DNN Embeddings for Speaker Recognition”, 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.5329-5333, 2018

[3] 戸田智基, ”確率モデルに基づく声質変換技術”, 日本音響学会誌, Vol.24, No.1, pp.34-39, 2010

[4] Kazuhiro Kobayashi, Tomoki Toda, ”sprocket: Open-Source Voice Conversion Software”, Odyssey, pp.203-210, 2018

[5] Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito, et al., ”JVS corpus: free Japanese multi-speaker voice corpus”, arXiv preprint, arxiv:1908.06248, 2019

[6] つくよみちゃんコーパス (CV. 夢前黎), https://tyc.rei-yumesaki.net/material/corpus/

[7] Brian McFee, Colin Raffel, Dawen Liang, et al., ”librosa: Audio and Music Signal Analysis in Python”, Proceedings of the 14th python in science Conference 8, pp18-25, 2015

[8] Mirco Ravanelli Titouan Parcollet, Peter Plantinga, et al., ”SpeechBrain: A General-Purpose Speech Toolkit”, arXiv preprint, arxiv:2106.04624, 2021

[9] Tom Ko, Vijayaditya Peddinti, Daniel Povey, et al., ”A study on data augumentation of reverberant speech for robust speech recognition”, 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.5220-5224, 2017

[10] Alvin F Martin, George R Doddington, Terri Kamm, et al., ”The DET curve in assessment of detection task performance”, Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), pp.1895-1898, 1997