深層学習を用いた声質変換の実装と実験的評価

An Experimental Study of Voice Conversion Using Deep Learning

情報工学専攻　髙橋 卓杜
Information and System Engineering　TAKAHASHI Takuto

あらまし：声質変換は, 入力話者による音声を, 音韻情報を保持したまま声質を対象とする話者に合わせて変換する技術である．本研究では, 多層ニューラルネットワークを用いた深層学習による声質変換を行い, 対象とする話者への音声の変換を実験的に評価する．音声を合成するには MLSA フィルタを利用する．さらに, 振幅スペクトログラムに対して位相を復元する手法を用いることで, 変換後の音声の話者らしさが改善されることを示す．

キーワード： 声質変換, 深層学習, メルケプストラム, MLSA フィルタ, RTISI-LA

1　はじめに

声質変換は, 入力話者による音声を, 音韻情報を保持したまま声の高さや声質など話者に関する情報を対象とする話者に合わせて変換する技術である．声質を変換するためには, 声道形状を表すメルケプストラムを用いることが多い．従来は, 声質の変換に混合正規分布を用いる手法が広く用いられてきた．近年では, 非線形的な形状を持つ声道の特徴を変換するために, 深層学習を利用する手法が提案されている．

Desai ら [1] は, 声質の変換に混合正規分布を用いる手法と多層ニューラルネットワークを用いる手法の比較実験を行い, 多層ニューラルネットワークによる手法の有用性を示した．多層ニューラルネットワークを用いた深層学習では, 中間層の数や各層のユニット数を与える必要がある．本研究では, Desai らの実験に基づき, 多層ニューラルネットワークを用いた深層学習による声質変換を行い, 対象とする話者への音声の変換を実験的に評価する．さらに, 振幅スペクトログラムに対して位相を復元し, 音声を合成する手法を用いることで, 変換後の音声の話者らしさが改善されることを示す．

2　音声特徴量の抽出

音声は口唇や歯, 舌などの調音器官の位置や動きによって生成される音響信号である．声帯で音源として生成された音声信号が, 調音器官により作られる音響的なフィルタを通過することで, 発話内容と対応する音韻の情報や話者固有の声質の情報が含まれた音声に変化する．音源として生成される音声信号は, 声帯の振動に伴う周期的な信号であり, 声の高さと対応する．一方フィルタは, 音韻や声質の情報を持つ．本研究では, 声の高さとして基本周波数を抽出し, フィルタを表現する特徴量にはメルケプストラム係数を用いる．

メルケプストラム係数とは, 人間の聴覚特性を考慮した周波数軸であるメル軸上の特徴量である．m 次のメルケプストラム係数 $\boldsymbol{mc}(m)$ は

$$\boldsymbol{mc}(m) = \frac{1}{2\pi j}\oint_C \log X(z)\tilde{z}^{m-1}d\tilde{z}$$
$$\log X(z) = \sum_{m=-\infty}^{\infty} \boldsymbol{mc}(m)\tilde{z}^{-m}$$
$$\tilde{z}^{-1} = \frac{z^{-1}-\alpha}{1-\alpha z^{-1}} \tag{1}$$

により定義され, 線形予測係数から求めることができる [4]．$X(z)$ は音声信号 $x(n)$ を z 変換したものであり, $\tilde{z}^{-1}$ は周波数軸をメル軸に変換するためのオールパスフィルタである．

3　声質変換

Desai ら [1] の声質変換の流れを図 1 に示す．まず音声信号から基本周波数とメルケプストラム係数を抽出する．次にこれらの特徴量を対象とする話者に合わせて変換し, 最後に音声信号を合成する．これをフレームごと

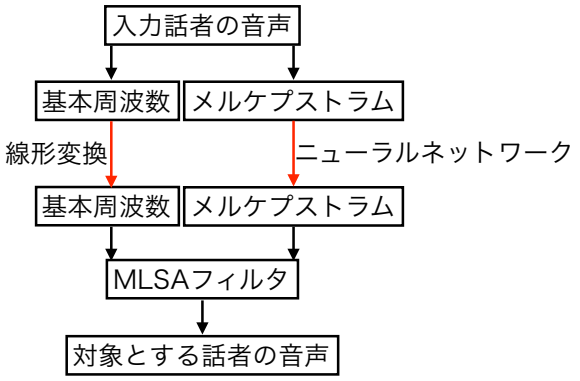

図 1. 声質変換の流れ

に繰り返し行うことで, 変換された対象とする話者の音声を得る.

基本周波数を変換するには, 音声データから抽出した基本周波数の対数をとり, その平均と標準偏差を用いて線形変換を行う. 入力話者の平均を μ_s, 標準偏差を σ_s, 対象とする話者の平均を μ_t, 標準偏差を σ_t とする. 入力話者の基本周波数 $\mathrm{F0}_s$ は, 以下の式により対象とする話者に変換した基本周波数 $\mathrm{F0}_c$ に変換される.

$$\mathrm{F0}_c = \exp\left(\mu_t + \frac{\sigma_t}{\sigma_s}(\log(\mathrm{F0}_s) - \mu_s)\right) \qquad (2)$$

メルケプストラム係数の変換には, 混合正規分布や多層ニューラルネットワークによるモデルを用いる. Desai らは比較実験により, 多層ニューラルネットワークを用いる手法の有用性を示している. 本研究では, メルケプストラム係数を変換するために多層ニューラルネットワークを利用する.

変換した基本周波数とメルケプストラム係数から音声信号を合成するには, 対数振幅スペクトルを近似するように音声を合成する MLSA フィルタを用いる.

4 ニューラルネットワークによる声質変換

多層ニューラルネットワークとは, 入力層, 中間層, 出力層から構成されるネットワークである. 入力層に入力したベクトルが, 中間層を通ることで変化し, 出力層において変換後のベクトルを得ることができる. 各枝は重みを持ち, 各点はバイアス項と活性化関数を持つ.

メルケプストラム係数の変換には, 4 層構造のニューラルネットワークを用いる. メルケプストラム係数の 0 次成分は音声の大きさを表すため, 0 次成分を取り除き変換する. 中間層のユニット数は入力層のユニット数の 3 倍に設定し, 活性化関数 g には双曲線正接関数を用いる. 学習率は 0.01, 慣性項は 0.3, 学習を行う反復回数は 200 回とそれぞれ設定する. 学習する際にモデルを評価する損失関数には平均二乗誤差を採用する. また, メルケプストラム係数の変化を捉えるために前後 3 フレーム分のメルケプストラム係数を並べたものを学習データとして用いる.

学習データには, 入力話者と対象とする話者の同じ音韻に対応するメルケプストラム係数の組が必要となる. そのため, 二者が同じ内容を発話した音声データを利用する. しかし同じ内容を発話した音声データであったとしても, 発話スピードの差や間の取り方などにより, 同じ時刻のメルケプストラム係数同士は完全には対応しない. そこで動的伸縮法を適用し, 二者のメルケプストラム係数の対応関係を得る. このようにして求めた学習データを用いて, 深層学習を行う.

4 層構造のニューラルネットワークを用いて学習を行うと, 重み行列 $W^{(i)}$ とバイアス $\boldsymbol{b}^{(i)}$ $(i = 1, 2, 3)$ が求まる. 変換後のメルケプストラム係数 $\boldsymbol{mc}_c$ は, 変換前のメルケプストラム係数 $\boldsymbol{mc}_s$ と活性化関数 g を用いて以下のように計算できる.

$$\boldsymbol{mc}_c = W^{(3)}g(W^{(2)}g(W^{(1)}\boldsymbol{mc}_s + b^{(1)}) + b^{(2)}) + b^{(3)}$$

5 音声合成

音声を合成するためには, MLSA フィルタを利用する. MLSA (mel log spectrum approximation) フィルタとは, 対数振幅スペクトルを近似するように音声を合成するデジタルフィルタである. メルケプストラム係数によりフィルタが決定し, 音源に対応する波形を与えることで音声が合成される. M 次のメルケプストラム係数を用いた MLSA フィルタの伝達関数は以下のような式である.

$$H(z) = \exp\left(\sum_{m=0}^{M} \boldsymbol{mc}(m)\tilde{z}^{-m}\right)$$

本研究では, MLSA フィルタにより出力される音声信号に, さらに RTISI-LA [5] を適用する. RTISI-LA (real-time iterative spectrogram inversion look-ahead) とは, 振幅スペクトログラムに対して位相を復元し, フレーム間の位相の連続性を考慮した音声を合成する手法である. 隣り合う数フレームとの位相の連続性を考慮し, 時系列順に逐次位相を推定する.

RTISI-LA により位相を推定する手順は以下の通りである.

1. 次のフレームに対して, 初期位相を決定する.
2. 逆離散フーリエ変換と合成窓関数により, 一連の波形を作る.
3. 分析窓関数を掛け, 離散フーリエ変換を行う.
4. 各スペクトルに対して振幅を初期振幅に置換する.
5. ステップ 2 からステップ 4 を繰り返し行う.
6. 波形を出力し, 考慮するフレームを進める

本研究では, 分析窓関数 $w_a[n]$ と合成窓関数 $w_s[n]$ はハ

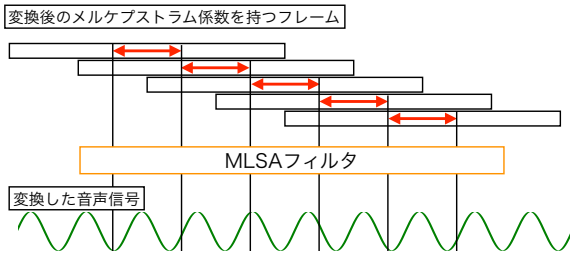

図 2. MLSA フィルタのフィルタ係数に用いるメルケプストラム係数の区間

ミング窓による以下の式を用いる.

$$w_a[n] = w_s[n] = \frac{2\sqrt{s}}{\sqrt{(4a^2+2b^2)N}}\left(a - b\cos\left(\frac{2\pi n}{N-1}\right)\right)$$

ただし $a = 0.54, b = 0.46$ であり, N はフレーム長, s はフレームシフトを表す.

MLSA フィルタのみを用いて音声を合成する手法では, フレーム同士の重なりを考慮し, 1 フレームのメルケプストラム係数を用いて出力する波形は, フレームシフト分の長さとする. 図 2 は各フレームが持つ変換後のメルケプストラム係数を, MLSA フィルタのフィルタ係数として用いる区間を表している.

RTISI-LA を適用する場合は, まず MLSA フィルタによりフレームと同じ長さの波形を出力する. 1 つのフレームに対して 1 つの MLSA フィルタを用いて波形を合成し, フレームごとに離散フーリエ変換を行い, 得られたスペクトログラムに対して RTISI-LA を適用する. そのため初期位相を設定する必要はない. また 6 節の実験ではフレームシフトをフレーム長の 1/4 と設定しているため, MLSA フィルタにより合成された波形は, 4 フレームごとに隣り合う. RTISI-LA を適用する手法では, 4 フレーム前の合成に用いた MLSA フィルタを引き継いで利用する.

RTISI-LA では, 位相の連続性を考慮する未来のフレーム数と反復回数を大きくすることで, より自然な音声を合成することができる. 一方で, 計算量が増加するため, リアルタイムに変換することは難しくなる. 本研究では, 考慮する未来のフレームを 3 フレームとし, 反復を 10 回まで変化させて実験を行う. RTISI-LA を用いることで, 数十ミリ秒程度の遅延で位相を推定し, MLSA フィルタによるフレームごとの波形が持つ振幅に近い音声信号を合成することができる.

6 計算機実験による評価

6.1 実験条件

実験では, CMU ARCTIC データベース [2] の音声データ (データ 1) と voiceactress100 [3] の音声データ (データ 2) を利用する. データ 1 を利用する学習では 50 文を用い, データ 2 を利用する学習では 15 文と 30 文を用いる 2 通りを行った. テストデータには各モデル一律で 50 文を利用する. データ 1 の 50 文は約 2 分 30 秒の音声データで, データ 2 の 15 文は約 2 分の長さである.

メルケプストラム係数には, 0 次から 25, 40, 60 次までの 3 パターンを抽出する. データ 1 では同性間, 異性間を考慮し 4 組の実験を行い, データ 2 を利用する実験では入力話者と対象とする話者を 1 組決定し学習データ数を変化させる. また, 式 (1) の α は, サンプリング周波数に依存する値であるためデータ 1 では 0.41, データ 2 では 0.554 と設定した.

6.2 メルケプストラム係数の変換

深層学習により構築したモデルを評価する. 評価指標には, 入力話者の音声を変換したメルケプストラム係数 $\boldsymbol{mc}_c$ と対象とする話者の音声から求めたメルケプストラム係数 $\boldsymbol{mc}_t$ を比較するメルケプストラム歪み (MCD) を用いる.

$$\mathrm{MCD} = \frac{10}{\log(10)}\sqrt{2\sum_{m=1}^{M}(\boldsymbol{mc}_c(m) - \boldsymbol{mc}_t(m))^2}$$

表 1 と表 2 では, メルケプストラム係数の次数を変化させて学習したモデルの MCD を比較する. 表 1 は, 同性間の変換と異性間の変換を行うモデルの評価を表す. (m) は男性話者を表し, (f) は女性話者を表す. 同性間と異性間には MCD の違いは見られず, 同性と異性の区別なく学習できていることが読み取れる.

表 2 は学習データ数を変化させたモデルの評価を表す. 学習データ数を増やしても MCD は改善されなかったことから学習データ数は約 2 分の長さで十分であることが確認できる.

6.3 合成した音声の評価

最後に, 合成した音声の評価を行う. 学習データをデータ 2 の 15 文, メルケプストラム係数を 40 次まで,

表 1. 同性, 異性間のモデルの MCD(データ 1 を利用)

入力話者 → 対象話者	26 次元	41 次元	61 次元
bdl (m) → rms (m)	8.009	8.799	9.762
bdl (m) → slt (f)	7.998	8.795	9.602
slt (f) → bdl (m)	7.803	8.643	9.450
slt (f) → clb (f)	7.484	7.851	9.135

表 2. 学習データ数を変化させたモデルの MCD (データ 2 を利用)

学習データ数	26 次元	41 次元	61 次元
15 文	6.467	7.915	8.534
30 文	6.588	7.632	8.596

60 次までとしたモデルに注目する. 図 3 では, RTISI-LA を適用せず 1 つの MLSA フィルタのみを用いる手法 (a) と, 4 つの MLSA フィルタを用いてフレームごとに合成した音声を足し合わす手法 (b), RTISI-LA を適用する手法 (c) の MCD を比較する. 手法 (c) では, RTISI-LA の反復回数を 1 回から 10 回まで変化させた. 手法 (a) での変換後のメルケプストラム係数には, MLSA フィルタを用いて合成した音声波形から再度計算したメルケプストラム係数を用いる.

手法 (a) と比較すると, フレームごとに合成する手法 (b) と RTISI-LA を適用する手法 (c) では MCD が小さくなり, 対象とする話者らしさが改善されることが確認できた. これは, MLSA フィルタを用いて合成した際に生じるノイズが, 前後のフレームと足し合わされることで平滑化されたためであると推測される. 一方で, RTISI-LA の反復を数回行った以降は, MCD にはほとんど変化は見られなかった. RTISI-LA では, MLSA フィルタにより出力された音声波形に対して離散フーリエ変換を行うため, 初期位相を推定する必要がない. 振幅のみをもつスペクトログラムから位相を推定する場合と異なり, すでに良い位相を保持しており, 速い段階で収束したためであると考えられる.

7 結論

本研究では, 多層ニューラルネットワークを用いた深層学習による声質変換を実装し, 実験による評価を行った. 音声の合成には, MLSA フィルタのみを用いる手法と MLSA フィルタに加えて RTISI-LA を用いる手法を実験した. RTISI-LA の適用により, 対象とする話者

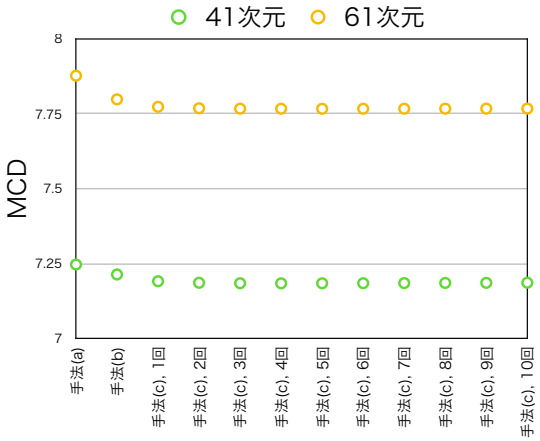

図 3. 音声合成手法と MCD

の特徴をより強く持つ音声を合成できることを示した.

実験では, 入力話者と対象とする話者の組を設定し, 同じ発話内容を記録した音声データを用いて行った. 任意の話者の音声を別の任意の話者のものに変換する手法や, 発話内容の異なる音声データを利用し声質を変換する手法に対して, RTISI-LA を利用する実験を行うことが今後の課題として挙げられる.

参考文献

[1] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad: Spectral mapping using artificial neural networks for voice conversion, *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 954–964, 2010.

[2] J. Kominek, and A. W. Black: CMU ARCTIC databases for speech synthesis, CMU-LTI-03-177, Carnegie Mellon University, 2003.

[3] R. Sonobe, S. Takamichi, and H. Saruwatari: JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis, arXiv preprint, 1711.00354, 2017.

[4] K. Tokuda, T. Kobayashi, and S. Imai: Recursive calculation of mel-cepstrum from LP coefficients, 入手先 <https://www.sp.nitech.ac.jp/~tokuda/tips/mgceptr_sa2.pdf> (参照 2018-02-10).

[5] X. Zhu, G. T. Beauregard, and L. L. Wyse: Real-time signal estimation from modified short-time Fourier transform magnitude spectra, *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 1645–1653, 2007.