畳込みニューラルネットワークを用いた音響特徴量変換とスペクトログラム高精細化による声質変換

廣芝 和之[1,a)]　能勢 隆[2]　宮本 颯[2]　伊藤 彰則[2]　小田桐 優理[1]

概要：本稿では，元話者の声質を目標話者の声質に変換する問題を扱う．ニューラルネットワークを用いた統計的声質変換では，自己回帰モデルである WaveNet を用いた手法が知られている．我々は今回，並列処理による高速化の難しい自己回帰を用いず，畳み込みのみで構成されたニューラルネットワークによって声質変換を行うことを目指す．提案するネットワークは 2 段階で構成されている．1 段階目は，元話者の声質を目標話者の声質に変換するネットワークである．ここでは，音響特徴量を時間方向に畳み込んで，時間変動を考慮した変換結果を得ることを期待する．音響特徴量としては低次のメルケプストラム系列を用いることで，少数のパラレルデータでも過学習しないことを期待する．2 段階目は，変換結果を更に高品質化するためのネットワークである．ここでは，1 段目で得られた音響特徴量をスペクトログラムに戻したあとに，改めて高品質化を行う．1 段階目と 2 段階目を独立に学習可能なため，高品質化を比較的入手が容易な目標話者のデータのみで学習できるようになる．

1. はじめに

本研究では入力話者の声質を別の目標話者の声質へと変換する声質変換技術の改善を目指す．機械学習モデルを用いた統計的声質変換は，学習と変換からなる．学習時は，入力話者と目標話者の同一文章を読み上げた音声で構成されるパラレルデータを用いて，両音声の音響特徴量間のマッピング関数を学習する．変換時は，学習されたマッピング関数を用いて入力話者の音響特徴量から目標話者の音響特徴量へ，音韻性を保ったまま変換する．本研究でもマッピング関数のモデルの提案を行う．従来から声質変換では，瞬時的特徴の変換を行うために音響特徴の動的特性が失われる問題や，変換結果が過剰に平滑化される問題があった．

音響特徴の動的特性が失われる問題については，入力話者の音響特徴の時間的な依存関係を考慮することで改善する手法が存在する．従来の声質変換では，瞬時的な音響特徴をフレームごとに変換する手法が取られており，マッピング関数のモデル化には混合ガウス分布 (GMM)[1] や，全結合な深層ニューラルネットワーク (DNN)[2] が用いられてきた．これに対し，長期間の依存関係を考慮するために，再帰的ニューラルネットワークである LSTM を用いてマッピング関数をモデル化する手法 [3] が提案されている．

変換結果が過剰に平滑化される問題については，学習時に抑制して改善する手法と，変換時に後処理して改善する手法が存在する．学習時に抑制する手法としては，敵対的生成ネットワーク (GAN) を用いて合成音声と自然音声の分布差を小さくする手法 [4][5] が提案されている．変換時に後処理する手法としては，系列内変動の減少を抑える手法 [6] や，マッピング関数とは別に，メルケプストラムやスペクトログラムといった音響特徴を高精細化する DNN モデルを学習し，マッピング関数によって生成された音響特徴に適用することで過剰な平滑化を抑制する手法 [7][8] が提案されている．

本研究では，別々に学習した 2 つの機械学習モデルを用いて，前述した問題をそれぞれ解決した声質変換手法を提案する．1 段目のモデルでは，時間方向に 1 次元畳み込み層を持つ畳込みニューラルネットワーク (CNN) を用いて，時間的な依存関係を考慮しつつ，入力話者の基本周波数と低次のメルケプストラム系列を目標話者のものに変換する．2 段目のモデルでは，GAN を用いて，過剰に平滑化された変換時の音響特徴を高精細化する．このモデルは，画像処理の分野で報告されている，変換と高精細化の 2 段階に分けて高品質な画像を生成する手法にならうものである [9][10]．この時，パラレルデータを必要とするのは，1 段目のモデルの学習時のみである．そこで，これらのモデルを独立に別のデータセットで学習することで，学習時に必要なパラレルデータ数の削減を試みる．実験結果から，従来手法と比べ，提案手法は同程度の自然性と高い個人性

[1] 株式会社ドワンゴ
[2] 東北大学
[a)] kazuyuki_hiroshiba@dwango.co.jp

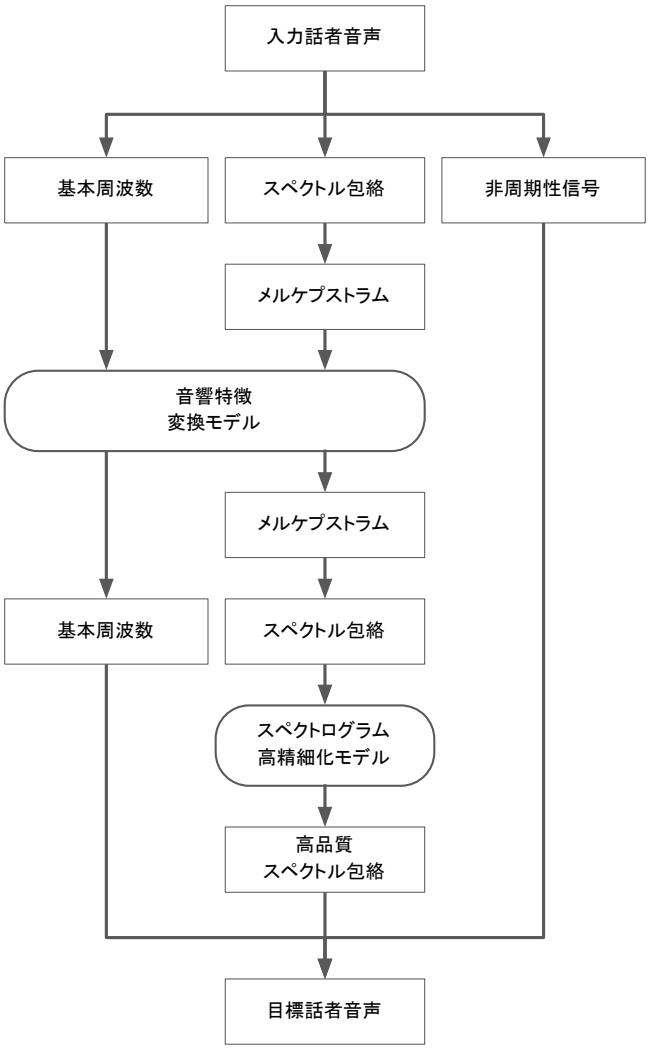

図 1 提案手法の変換時の処理

を持つ声質変換が可能であることを示す．

2. 2 段階の変換モデル

本稿では，畳込みニューラルネットワーク (CNN) を用いた音響特徴量変換モデルと，スペクトログラム高精細化モデルの 2 段階に分けた声質変換のモデルを用いる (図 1)．1 段目の音響特徴量変換モデルでは，入力話者の基本周波数とメルケプストラムを目標話者のものに変換する．このモデルでは低次のメルケプストラムを利用しているため，変換後のメルケプストラムから再構成したスペクトル包絡は荒いものとなる．そこで 2 段目のスペクトログラム高精細化モデルでは，変換後のメルケプストラム系列をスペクトログラムに戻し，このスペクトログラムを高品質なスペクトログラムに変換する．前者のモデルの学習には入力話者と目標話者のパラレルデータが必要である一方，後者のモデルは目標話者内での変換であるため，入力話者とのパラレルデータが無い目標話者の音声データも学習に用いることができる (図 2)．これらのモデルの詳細を順に説明する．

2.1 音響特徴量変換モデル

ここでは，時間的な依存関係を考慮しながら，入力話者の低次元音響特徴を，目標話者の低次元音響特徴に変換する音響特徴量変換モデルについて述べる．時間的な依存関係を考慮するために，1 次元 CNN を用いる [4]．この CNN では特徴量をチャンネルとみなし，時間方向に畳み込む．ネットワークは画像変換の分野で成功している U-Net[11] の構造を参考にし，1 次元の畳み込み層を重ねた構造を用いる．入力特徴を低次元に圧縮する Encoder と出力特徴まで伸長する Decoder を組み合わせ，Encoder と Decoder の中間層同士にスキップ接続が存在することで細部まで高精細な出力を生成できことが U-Net の特徴である．層の数や，各ネットワーク層のチャンネル数，活性化関数は文献 [11] と同じ値を用いる．

2.2 スペクトログラム高精細化モデル

このモデルでは，第 1 段階のモデルによって得た目標話者のメルケプストラムをスペクトル包絡に戻し，このスペクトル包絡を高精細化する．学習には目標話者の音声から得たスペクトル包絡と，メルケプストラムを介して再構成したスペクトル包絡を用いる (図 2)．スペクトル包絡の予測誤差のみを最小化するように学習すると，推定されたスペクトル包絡が過剰に平滑化する．この現象を抑制するために，敵対的生成ネットワーク (GAN) を用いる [5]．本研究では，GAN を利用したモデルの中でも，pix2pix モデルを用いてスペクトログラムを高精細化する．画像変換の分野で有効な pix2pix の特徴により，時間・周波数方向に大局的な特徴を捉えつつ，変換結果を高精細化することを期待している．層の数や，各ネットワーク層のカーネルサイズ，チャンネル数，活性化関数，損失の比率は文献 [12] と同じ値を用いる．

3. 実験による評価

3.1 実験条件

本実験では，ATR 日本語音声データベースセット B の MHT 男声話者と FKN 女声話者の音声データを用いて，MHT 話者から FKN 話者に声質変換する．音声データの発話内容は同一の音素バランス文 503 文である．500 文の発話の内，50〜450 文を学習に，残り 53 文を評価に用いる．学習に用いるパラレルデータは，動的時間伸縮法でアライメントを行う．サンプリング周波数は 16kHz とする．本実験で用いる基本周波数，スペクトル包絡，非周期性信号は，WORLD 分析 [13] により得る．この時，スペクトル包絡は 513 次元，シフト長は 5ms とする．学習データのバッチサイズを 8 とし，Adam[14] を用いて最適化する．本研究で用いた学習・変換のソースコードは公開されてい

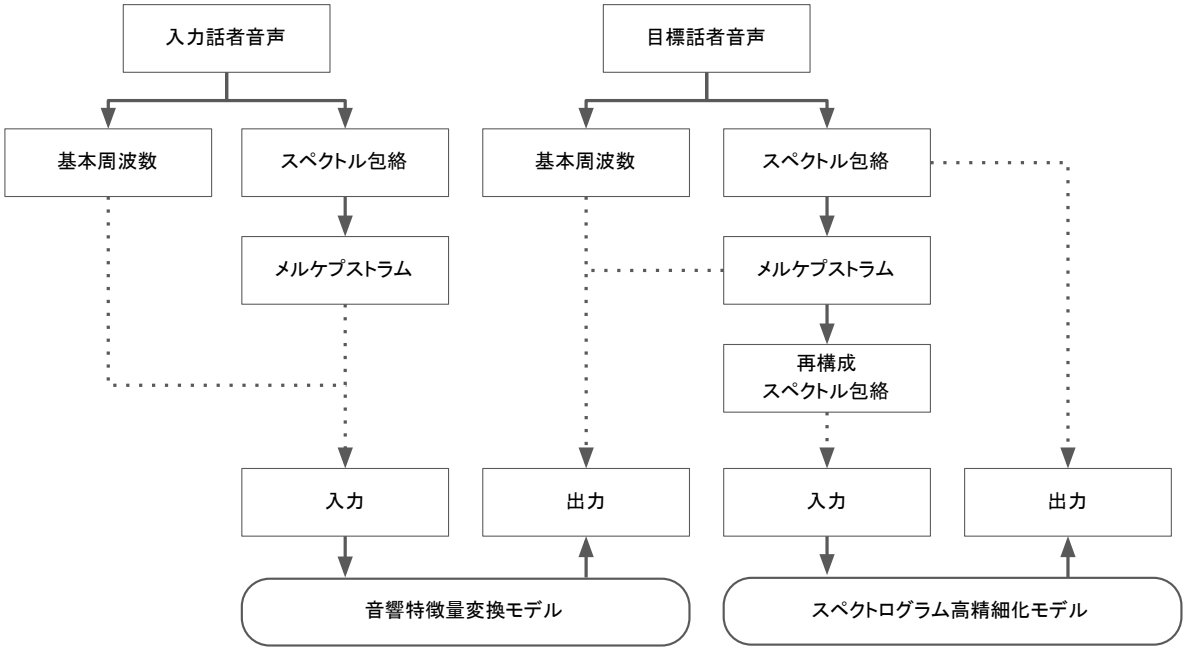

図 2 提案手法の学習時の処理

る[*1].

音響特徴量変換モデルで変換する音響特徴量として，基本周波数と 0 次から 8 次のメルケプストラム係数を用いる．本実験において非周期性信号の変換は行わず，変換時は入力話者の非周期性信号を用いる．それぞれの特徴量は平均を 0，分散を 1 に正規化して学習に用いる．スペクトログラム高精細化モデルで変換するスペクトル特徴量として，0 次から 511 次のスペクトル包絡を用いる．512 次のスペクトル包絡は変換に用いず，変換後に 0 埋めする．スペクトル特徴量は対数スケールに変換して学習に用いる．

評価実験により，従来手法と提案手法を比較する．従来手法には，sprocket[*2]の DIFFVC 手法 [15] を用いる．従来手法のモデル (図 3 および図 4 の Conventional) の学習には 50 文のパラレルデータを用いる．提案手法には，高精細化モデルの学習データ数を変えて，2 種類のモデルを用いる．提案手法の 1 段目のモデルの学習には 50 文のパラレルデータを用いる．2 段目は，1 段目の学習に用いた 50 文のデータのみを用いて学習するモデル (図 3 および図 4 の Proposed 50-50) と，その 50 文に 400 文を加えたデータで学習するモデル (図 3 図 4 の Proposed 50-450) を用いる．この時，従来手法と提案手法の学習で用いるパラレルデータ数は全て同じである．

3.2 実験方法

主観評価実験により，変換音声の自然性と個人性をそれぞれ比較する．変換音声の自然性を AB テストにより評価する．従来手法および提案手法で変換された音声をランダムな順序で再生し，高い自然性を持つ変換音声を被験者に選択させる．また，変換音声の個人性を XAB テストにより評価する．目標の音声を再生した後に，従来手法と提案手法により変換された音声をランダムな順序で再生し，目標音声と似ている変換音声を被験者に選択させる．被験者 9 名は 15 対の音声に対してそれぞれ評価を行う．なお，被験者の内 1 名は本報告の実験に携わった者である．

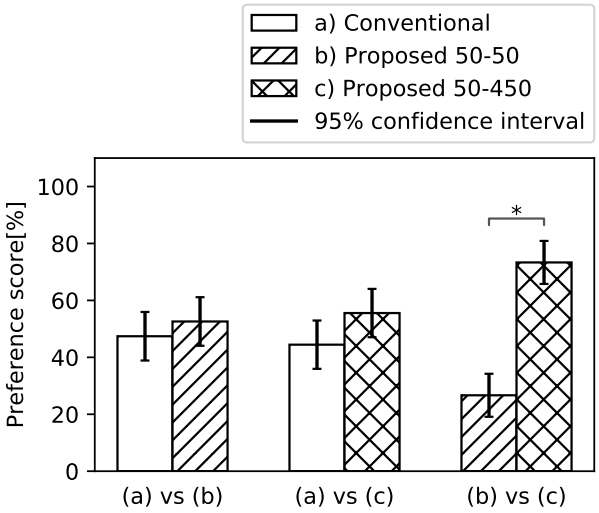

図 3 自然性に関する評価結果．多重比較:Bonferroni 法, * p <0.001

3.3 実験結果

図 3 に変換音声の自然性評価結果を示す．また，図 4 に変換音声の個人性評価結果を示す．提案手法の変換音声は，従来手法の変換結果に比べて，同程度の自然性と，高い個人性を持つことが分かる．この理由として，既存手法

[*1] https://github.com/Hiroshiba/become-yukarin
[*2] https://github.com/k2kobayashi/sprocket

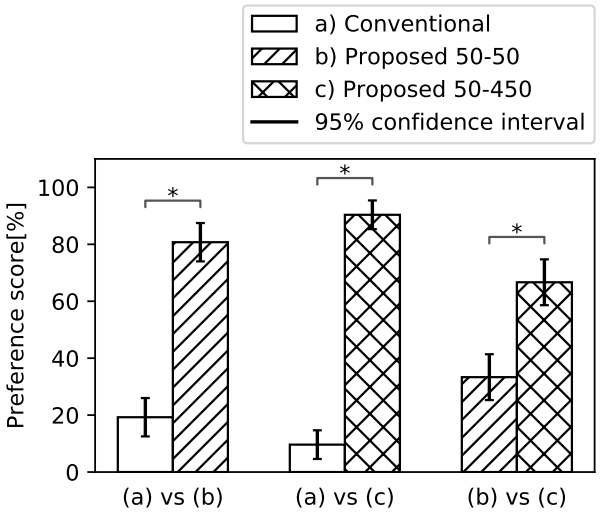

図 4 個人性に関する評価結果．多重比較:Bonferroni 法, * p <0.001

に比べて提案手法のモデルの表現能力が大きかったことや，音響特性の時間的な依存関係を考慮して基本周波数を変換し，目標話者の特徴を上手く捉えたことが考えられる．また，学習に用いる目標話者の音声データ数を増やすことで，より自然性・個人性の高い変換音声が得られることが分かる．これは，より多くのデータを用いることで，目標話者の音響特徴を詳細に捉えることができるためである．

4. まとめ

機械学習モデルを用いた統計的声質変換において，音響特徴の変換と高精細化の 2 段階に分けたモデルを提案した．既存手法と提案手法の比較実験により，提案手法は同等の自然性と保ちつつ，個人性の高い音声に変換できることを示した．また，目標話者の音声を増やして学習することで，自然性・個人性共に高い変換音声が得られることを確認した．今後の研究として，変換精度向上や，任意話者からの変換，リアルタイム声質変換への応用に取り組む．

謝辞

本研究を進めるにあたり，株式会社ドワンゴ大垣慶介氏からご指導を賜りました．ここに感謝の意を表します．

参考文献

[1] Stylianou, Y., Cappé, O. and Moulines, E.: Continuous probabilistic transform for voice conversion, *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 2, pp. 131–142 (1998).
[2] Ling, Z.-H., Kang, S.-Y., Zen, H., Senior, A., Schuster, M., Qian, X.-J., Meng, H. M. and Deng, L.: Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends, *IEEE Signal Processing Magazine*, Vol. 32, No. 3, pp. 35–52 (2015).
[3] Sun, L., Kang, S., Li, K. and Meng, H.: Voice conversion using deep bidirectional long short-term memory based recurrent neural networks, *ICASSP*, pp. 4869–4873 (2015).
[4] Kaneko, T., Kameoka, H., Hiramatsu, K. and Kashino, K.: Sequence-to-Sequence Voice Conversion with Similarity Metric Learned Using Generative Adversarial Networks, *Interspeech*, pp. 1283–1287 (2017).
[5] Saito, Y., Takamichi, S. and Saruwatari, H.: Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks, *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 26, No. 1, pp. 84–96 (2018).
[6] Toda, T., Black, A. W. and Tokuda, K.: Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory, *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2222–2235 (2007).
[7] Kaneko, T., Kameoka, H., Hojo, N., Ijima, Y., Hiramatsu, K. and Kashino, K.: Generative adversarial network-based postfilter for statistical parametric speech synthesis, *ICASSP*, pp. 4910–4914 (2017).
[8] Kaneko, T., Takaki, S., Kameoka, H. and Yamagishi, J.: Generative Adversarial Network-Based Postfilter for STFT Spectrograms, *Interspeech*, pp. 3389–3393 (2017).
[9] Furusawa, C., Hiroshiba, K., Ogaki, K. and Odagiri, Y.: Comicolorization: semi-automatic manga colorization, *SIGGRAPH Asia Technical Briefs*, p. 12 (2017).
[10] Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X. and Metaxas, D.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, *ICCV*, pp. 5907–5915 (2017).
[11] Ronneberger, O., Fischer, P. and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, *MICCAI*, pp. 234–241 (2015).
[12] Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A. A.: Image-to-image translation with conditional adversarial networks, *CVPR* (2017).
[13] Morise, M., Yokomori, F. and Ozawa, K.: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, *IEICE TRANSACTIONS on Information and Systems*, Vol. 99, No. 7, pp. 1877–1884 (2016).
[14] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *ICLR* (2015).
[15] Kobayashi, K., Toda, T. and Nakamura, S.: F0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential, *SLT*, pp. 693–700 (2016).