サブバンドフィルタリングに基づくリアルタイム広帯域 DNN 声質変換の実装と評価 *

☆佐伯 高明, 齋藤 佑樹, 高道 慎之介, 猿渡 洋 (東大院・情報理工)

1 はじめに

近年，deep neural network (DNN) を用いて高精度かつ柔軟な変換を行う声質変換手法が盛んに研究されている．既存の DNN 声質変換の多くは，バッチ型の特徴量分析・変換・波形合成過程に基づくオフライン変換手法であるが，実応用上の観点から，CPU でリアルタイムかつオンラインに変換を行う DNN 声質変換手法が提案されている [1]．この手法は，単一 CPU を用いて遅延 50 ms 程度のリアルタイム変換を達成しているが，依然として計算コストが大きいため，人間の可聴域をカバーする広帯域音声の変換に拡張することは難しい．さらに，信号処理に基づくボコーダにより変換音声にアーティファクトが生じるという問題もある．

我々は，波形ドメインでのフィルタリングに基づく変換手法 [2] によってこれらの問題の解決を図る．これまで，我々は，差分スペクトル法に基づく，高品質かつ計算効率の高い広帯域声質変換手法を提案した [3]．この手法は，サブバンドマルチレート処理 [4] に基づく帯域分割を行い，低域のみを差分スペクトル法でモデル化・変換し，高域のモデル化を行わずに保持することによって計算コストの削減と音質の改善を実現する．これに加えて，低域の変換時にフィルタ打ち切りを行い，タップ長の短いフィルタを設計する [5, 6] ことで，変換にかかる計算量をさらに削減する．

本稿では，このサブバンドフィルタリングに基づくリアルタイム広帯域声質変換をオンライン変換の形で実装する手法を提示するとともに，その品質評価を行う．実験的評価では，まず 1) オンライン実装がオフライン実装の場合と同等品質の変換音声を出力できることを示す．さらに，2) 提案するシステムが単一 CPU でリアルタイムに動作することを，計算量の概算および real-time factor (RTF) の計測によって示し，3) 同性間変換・異性間変換の場合について，変換音声の品質を mean opinion score (MOS) により評価し，自然な変換音声を出力できることを示す．

2 サブバンドフィルタリングに基づく広帯域 DNN 声質変換

最小位相フィルタを用いた差分スペクトル法 [7] は高品質だが，この手法を広帯域声質変換にそのまま拡張した場合，1) 高域周波数スペクトルの変動により DNN のモデル化性能が低下し，2) 単位時間あたりのサンプル数の増大によってフィルタリングの計算コストが大きくなる．

これまで，我々はサブバンドフィルタリングと DNN 音響モデルに基づく高品質な広帯域声質変換を提案した [3]．この手法は，話者性に寄与する最低域のサブバンド信号のみをモデル化・変換し，高域を保持することによって，計算量を削減すると同時に変換音声の音質を改善する．まず，サブバンドマルチレート信号処理によって変換元話者の 48 kHz サンプリング音声信号を 3 つの帯域に分割し，0–8 kHz，8–16 kHz，16–24 kHz のそれぞれの帯域の信号を得る．その際，最低域 (0–8 kHz) の信号のみを差分スペクトル法でモデル化・変換し，高域 (8–24 kHz) の帯域は何も処理を加えない，またはモデル化を行わずに粗く変換する．変換された各々の帯域のサブバンド信号を再びサブバンドマルチレート処理によって合成することにより，最終的な広帯域変換音声が得られる．この手法は，既存の多くの DNN 声質変換と同様のオフライン変換手法であり，サブバンド処理や差分スペクトル法による低域の変換は発話単位で行う．

3 リアルタイム広帯域声質変換システムの実装

本節では，提案するリアルタイムオンライン広帯域声質変換システムの枠組みを説明する．システムの全体像を Fig. 1 に示す．本システムでのオンライン変換では，変換元話者の 5 ms 波形を入力して変換音声の 5 ms 波形を出力する．この際，逐次的に受け取った 5 ms 波形を 25 ms フレームの先頭に配置し，25 ms フレームを 5 ms 以内の処理時間で変換する．

3.1 基本的な構成

本節では，提案するシステムの基本的な構成を，分析部・変換部・合成部に分けて説明する．

3.1.1 分析部

分析部では，入力フレームのフルバンド音声波形から各々の帯域のサブバンド信号を取り出し，最低域の信号の声道特徴量となる低次実ケプストラムを抽出する．まずフルバンド音声のフレームに hanning 窓をかけ，フレーム長が 2 冪になるように 0 埋めしてサブバンドマルチレート処理を適用する．最低域のサブバンド信号に対し，高速フーリエ変換処理に基づくケプストラム分析を行うことにより，変換元話者の低次実ケプストラム $\boldsymbol{C}^{(\mathrm{X})}$ を抽出する．

3.1.2 変換部

低域の変換時は，DNN による特徴量変換を行い，時間領域の差分フィルタを構成して変換元話者のサブバンド音声に適用する．まず，変換元話者の低次実ケプストラム $\boldsymbol{C}^{(\mathrm{X})}$ を DNN の入力として差分フィルタの低次実ケプストラムを推定する．これに最小位

* Implementation and evaluation of real-time full-band DNN-based voice conversion based on sub-band filtering by Takaaki Saeki, Yuki Saito, Shinnosuke Takamichi, Hiroshi Saruwatari (The University of Tokyo)

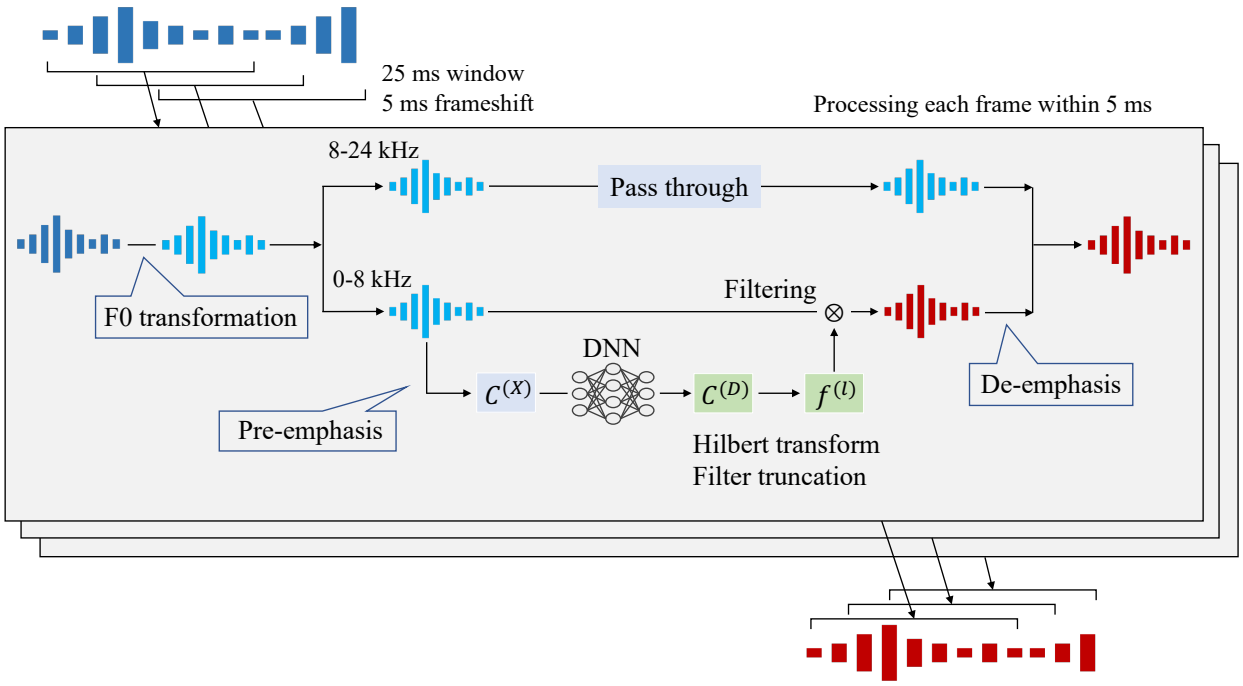

Fig. 1　提案するオンライン声質変換システムのパイプライン．フレーム長 25 ms，フレームシフト 5 ms で入力波形を受け取り，処理時間 5 ms 以内でリアルタイムに変換する．変換部では 0–8 kHz の低域のみに差分スペクトル法を適用し，8–24 kHz の高域は入力信号をそのまま通す．

相化のためのリフタ係数またはデータドリブンに得られるリフタ係数 [6] をかけてヒルベルト変換を行うことにより，差分フィルタの複素スペクトル $\boldsymbol{F}^{(\mathrm{D})}$ を得る．さらに，$\boldsymbol{F}^{(\mathrm{D})}$ を逆フーリエ変換し，フィルタ打ち切りを行ってタップ長の短いフィルタとすることにより，時間領域での差分フィルタ $\boldsymbol{f}^{(l)}$ を得る．この差分フィルタ $\boldsymbol{f}^{(l)}$ を変換元話者のサブバンド信号に対して畳み込むことにより変換を行う．

高域 (8–24 kHz) の信号はランダム性が強いため，低域と同様のモデル化を行うことは困難である．そのため，高域には何も処理を加えずに合成部へと通過させる．

3.1.3　合成部

3.1.2 節の手法によって変換された各帯域の信号を，再びサブバンドマルチレート処理により合成する．これにより，変換されたフルバンド音声のフレームが得られる．このフレームを前回までの計算結果に対して overlap-add させ，フレームの最初の 5 ms 波形を出力することにより，最終的な出力波形が得られる．

3.2　高精度化・広適用範囲化のための手法

本システムでは，サブバンドフィルタリングに基づく広帯域 DNN 声質変換を単にオンライン化するだけでなく，さらなる高精度化・広適用範囲化に向けた手法を導入する．

3.2.1　F0 変換

差分スペクトル法は声道特徴のみを変換するため，F0 の変換を行うことはできない．本システムでは，異性間変換にも対応したシステムを実装するため，PICOLA [8] による波形領域での F0 変換を導入する．まず，学習データに含まれる変換元話者の音声・変換先話者の音声を WORLD [9] で F0 分析し，それぞれの話者の平均 F0 値を算出する．F0 変換を行う際は，これらの比となる平均 F0 比だけ PICOLA を用いてピッチシフトする．学習時は，平均 F0 比だけ F0 変換した変換元話者のデータと，変換先話者のデータを用いて DNN を学習する．変換時は，Fig. 1 に示すように，入力波形を平均 F0 比だけピッチシフトして分析部に渡すことにより，F0 変換を行う．PICOLA に基づく F0 変換は，WORLD などのボコーダに基づく手法よりも著しく計算量が小さいため，リアルタイム広帯域声質変換に適した手法である．また，この手法は，平均 F0 比が大きくなると著しく音声品質が低下するものの，平均 F0 比が比較的小さい場合はボコーダに基づくピッチシフトよりも高品質な変換音声を出力できる．

3.2.2　プリエンファシスの適用

フィルタの推定に用いる声道特徴量を改良するため，低域の信号に対してプリエンファシスを導入する．Fig. 1 に示すように，0–8 kHz のサブバンド信号をケプストラム分析する前に，$E(z) = 1 - \alpha z^{-1}$ で表されるプリエンファシスフィルタを適用することにより高域強調を行い，スペクトル傾斜の影響を低減するさらに，差分フィルタの適用により 0–8 kHz の波形を変換した後，サブバンド合成を行う前に，逆フィルタ $D(z) = 1/(1 - \alpha z^{-1})$ を適用する．

4　実験的評価

4.1　実験条件

4.2 節では，オンライン変換での変換音声の品質について，男性話者から男性話者 (m2m)，女性話者か

Table 1　オンライン変換とオフライン変換での変換音声の品質の比較結果.

(a) **Speaker similarity**

Online	Score	p-value	Offline
m2m	0.493 vs. 0.506	7.4×10^{-1}	m2m
f2f	0.486 vs. 0.513	5.1×10^{-1}	f2f

(b) **Speech quality**

Online	Score	p-value	Offline
m2m	0.517 vs. 0.483	4.2×10^{-1}	m2m
f2f	0.490 vs. 0.510	6.2×10^{-1}	f2f

ら女性話者 (f2f) の 2 種類の変換について評価を行う．4.4 節では，実装したリアルタイム広帯域声質変換システムによる変換音声の品質について，m2m，f2f のケースに加えて，男性話者から女性話者 (m2f)，女性話者から男性話者 (f2m) についても評価を行う．男性の変換元話者・変換先話者にはいずれも JVS コーパス [10] の男性話者を用いた．女性の変換元話者には JVS コーパス [10] と JSUT コーパス [11] の女性話者，変換先話者には JVS コーパス [10] と声優統計コーパス [12] の女性話者を用いた．それぞれの話者データについて 100 発話 (約 12 分) を使用し，80 文を training データ，10 文を validation データ，10 文を test データとした．

評価には 48 kHz サンプリング音声を用いた．分析の際は，窓長を 25 ms，フレームシフトを 5 ms，FFT 長を 2048 点とした．0–8 kHz の音声に短時間フーリエ変換を行う際は，窓長とフレームシフトには 48 kHz の場合と同じものを用い，FFT 長を 512 点，低次実ケプストラムの次元を 40 とした．前処理として，training データと validation データの無音区間を除去し，変換元話者の音声と変換先話者の音声のデータ長を dynamic time warping により揃えた．差分フィルタを適用する際，フィルタのタップ長を 1/4 に打ち切った．プリエンファシスの際の係数 α の値として $\alpha = 0.97$ を用いた．

実験に用いた DNN アーキテクチャは，隠れ層 2 層の Feedforward Neural Network であり，隠れユニット数はそれぞれ 280，100 とした．隠れ層の活性化関数は，sigmoid 関数，tanh 関数からなる Gated Linear Unit [13] であり，各々の活性化関数に通す前に Batch Normalization [14] を行った．また，最適化手法には Adam [15] を用いた．学習時に変換元話者と変換先話者のケプストラムを平均 0・分散 1 に正規化し，学習率を 0.0001 として 100 epoch 学習を行った．

本システムが単一 CPU でリアルタイムに動作することを評価するため，処理時間の評価では Intel (R) Core i7-6850K CPU @ 3.60 GHz を用いた．

4.2　オフライン変換とオンライン変換の品質比較

本研究でのオンライン実装の品質を評価するため，オンライン変換とオフライン変換による変換音声の品質に関する主観評価実験を実施した．オフライン変換は 2 節に示す手法である．クラウドソーシングを用いて音質に関する AB テスト・話者類似性に関する XAB テストを行い，各条件につき 30 人の聴取者が 10 文の音声サンプルを評価した．XAB テストにおけるリファレンス音声 X として，変換先話者の自然音声を使用した．公平な条件での比較となるよう，この評価でのオンライン変換に対しては 3.2.2 節のプリエンファシスを適用しなかった．

Table 1 の評価結果より，m2m，f2f の両ケースについて，話者類似性・音質ともに有意差は確認できず，オンライン実装がオフラインの場合と同等品質の変換音声を出力できることが分かる．

4.3　オンライン変換システムの計算効率に関する評価

計算効率に関する評価を，サブバンド処理 (Subband)，ケプストラム分析 (Cepstrum)，DNN による推論 (Inference)，ヒルベルト変換 (Hilbert trans.)，フィルタリング (Filtering)，F0 変換やプリエンファシスなどのその他の処理 (Other) ごとに行った．

まず，システムの各処理での計算量を floating-point operations per second (FLOPS) 単位で概算した．前節に示した実験条件から計算量のオーダーを見積もり，和算・乗算を同時に行えるとして 2 を乗じる．さらに，提案するシステムは 5 ms の音声波形を入出力することから，1 秒単位での値に変換することにより，各処理での FLOPS の値を算出した．また，F0 変換やプリエンファシスなどの処理は 0.3 GFLOPS として見積もった．Table 2 に結果を示す．システム全体の計算量は 2.5 GFLOPS となっており，理論的には Apple iPhone 6 などのモバイル端末の単一 CPU でも広帯域音声をリアルタイムに変換可能であることが確認できる．また，軽量なニューラルボコーダとして知られる LPCNet [16] よりも小さな計算量で，48 kHz サンプリング音声の変換を実現できることが分かる．

次に，提案するシステムが実際にリアルタイム動作可能であるかについて，4.1 節に示した計算機の単一 CPU を用いて実験を行うことにより評価した．F0 変換も含めた処理時間を計測するため，異性間変換の場合について計測を行った．まず，フレームシフトの音声波形の変換ごとに各プロセスでの処理時間を計測し，平均値を算出した．さらに，処理時間の平均値から，処理時間と入力の長さの比である RTF を算出した．Table 2 に結果を示す．結果的に，全プロセス合計での RTF は 0.58 であり，本システムがリアルタイム動作可能であることが確認できる．

参考値として，従来の差分スペクトル法 [7] を広帯域声質変換にそのまま適用した場合の計算量および処理時間について記す．先程と同様の方法で計算量を FLOPS 単位で概算すると，全体で 20.4 GFLOPS となった．また，従来の差分スペクトル法 [7] をオンライン変換の形で実装し，処理時間を計測した結果，全体での処理時間の平均値は 15.09 ms，RTF は 3.02 となった．以上の結果より，本システムは，既存の差分スペクトル法を用いた場合と比較すると大幅に計算効率が高いことが分かる．

4.4　オンライン変換システムの品質評価

本システムを用いて出力した変換音声の自然性に関して，MOS による主観評価実験を実施した．同性

Table 2　計算効率に関する評価結果．GFLOPS 単位での計算量 (Complexity)，1 フレームの処理時間 (Processing time)，RTF を各プロセスについて算出した．

	Sub-band	Cepstrum	Inference	Hilbert trans.	Filtering	Other	Total
Complexity (GFLOPS)	1.40	0.043	0.33	0.041	0.35	0.30	**2.5**
Processing time (ms)	1.54	0.025	0.66	0.042	0.35	0.29	**2.9**
RTF	0.31	0.0050	0.13	0.0084	0.070	0.058	**0.58**

Table 3　変換音声の自然性に関する MOS 評価結果．

(a) **Intra-gender conversion**

	m2m	f2f
Conventional [7]	2.60 ± 0.11	2.98 ± 0.13
Proposed	$\mathbf{3.33} \pm 0.13$	$\mathbf{3.50} \pm 0.12$

(b) **Cross-gender conversion**

	f2m	m2f
Conventional [7]	2.67 ± 0.10	3.12 ± 0.12
Proposed	$\mathbf{3.27} \pm 0.10$	$\mathbf{3.36} \pm 0.12$

間変換 (m2m，f2f) および異性間変換 (m2f，f2m) の計 4 ケースについて，本システムで出力した変換音声と，従来の差分スペクトル法 [7] で出力した変換音声とを比較して MOS 評価を行った．各条件につき，40 人の聴取者が計 12 文の音声サンプルを評価した．評価結果となる MOS の値を，95%信頼区間とともに Table 3 に示す．全ケースについて，本システムで出力した変換音声の自然性は，従来の差分スペクトル法を用いた場合を上回っていることが確認できる．また，同性間変換だけでなく，異性間変換においても比較的自然性の高い広帯域変換音声を出力できることが分かる．

5　おわりに

本稿では，サブバンドフィルタリングに基づくリアルタイム広帯域 DNN 声質変換の実装法を提示し，さらに高精度化・適用範囲の拡張に向けた手法を導入した．実験的評価により，まずオンライン実装が，既存のオフライン変換の場合と同等品質の変換音声を出力できることを示した．また，実装したシステムの計算効率を計算量の概算および処理時間の計測によって評価し，単一 CPU でリアルタイムに動作することを示した．さらに，変換音声の自然性を MOS により評価し，同性間変換だけでなく，異性間変換の場合についても比較的自然な変換音声を合成できることを示した．今後は，特徴量変換を行う DNN をより詳細に検討し，話者類似性の改善を行う．また，変換音声の自然性や音質のさらなる向上に向けた信号処理的手法の検討を行う．**謝辞:** 本研究開発は総務省 SCOPE(受付番号 182103104) の委託を受けたものです．

参考文献

[1] R. Arakawa, S. Takamichi, and H. Saruwatari, "Implementation of DNN-based real-time voice conversion and its improvements by audio data augmentation and mask-shaped device," in *Proc. SSW10*, Vienna, Austria, Sep. 2019, pp. 93–98.

[2] K. Kobayashi, T. Toda, and S. Nakamura, "Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential," *Speech Communication*, vol. 99, pp. 211–220, May. 2018.

[3] 佐伯高明, 齋藤佑樹, 高道慎之介, and 猿渡洋, "差分スペクトル法に基づく広帯域声質変換のためのサブバンドリフタ学習," in **音講論 （春）**, no. 2-2-5, 埼玉, Mar. 2020.

[4] R. Crochiere and L. Rabiner, *Multirate digital signal processing*. Englewood Cliffs, N.J. : Prentice-Hall, 1983.

[5] M. Sunohara, C. Haruta, and N. Ono, "Low-latency real-time blind source separation with binaural directional hearing aids," in *Proc. CHAT*, Stockholm, Sweden, Aug. 2017, pp. 9–13.

[6] T. Saeki, Y. Saito, S. Takamichi, and H. Saruwatari, "Lifter training and sub-band modeling for computationally efficient and high-quality voice conversion using spectral differentials," in *Proc. ICASSP*, Barcelona, Spain, May 2020, pp. 7784–7788.

[7] H. Suda, G. Kotani, S. Takamichi, and D. Saito, "A revisit to feature handling for high-quality voice conversion," in *Proc. APSIPA ASC*, Hawaii, U.S.A., Nov. 2018, pp. 816–822.

[8] 森田直孝 and 板倉文忠, "ポインター移動量制御による重複加算法 (PICOLA) を用いた音声の時間軸での伸長圧縮とその評価," in **音講論 （秋）**, no. 1-4-14, Oct. 1986.

[9] M. Morise, F. Yokomori, and K. 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, Jul. 2016.

[10] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free japanese multi-speaker voice corpus," *arXiv*, vol. abs/1908.06248, 2019.

[11] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," vol. abs/1711.00354, 2017.

[12] y_benjo and MagnesiumRibbon, "Voice-actress corpus," http://voice-statistics.github.io/.

[13] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," *arXiv*, vol. abs/1612.08083, 2016.

[14] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[15] D. Kingma and B. Jimmy, "Adam: a method for stochastic optimization," in *arXiv preprint arXiv:1412.6980*, 2014.

[16] J. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, Brighton, U.K., May 2019, pp. 5891–5895.