差分スペクトル法に基づくDNN声質変換のための リフタ学習及びサブバンド処理

佐伯 高明[1,a)]　齋藤 佑樹[1,b)]　高道 慎之介[1,c)]　猿渡 洋[1,d)]

概要：本稿では，差分スペクトル法に基づく高品質かつ計算効率の高い声質変換のためのリフタ学習法及びサブバンド処理を提案する．最小位相フィルタを用いた差分スペクトル法では，高品質な反面フィルタのタップ長が長くなるという問題があるが，単にフィルタ打ち切りを行うだけでは計算量が削減される代わりに品質が劣化する．我々が提案するリフタ学習法では，フィルタの打ち切りタップ長を条件として，実ケプストラムを変換する deep neural network のパラメータだけでなく，実ケプストラムから位相を復元するためのリフタ係数をも学習することにより，フィルタ打ち切りが品質に与える影響を軽減する．また，差分スペクトル法を広帯域音声の変換に用いた場合，高域のランダムな成分を変換することにより，変換音声の品質が低下するという問題が生じる．本研究で提案するサブバンド処理では，低域のみをフィルタで変換することにより，この問題を回避する．実験的評価により，(1) フィルタ打ち切りによるリフタ学習法を用いることで，変換時のフィルタリングの計算量を削減できること，(2) サブバンド処理を広帯域音声の変換に用いることにより，変換音声の品質が大幅に向上することを示す．

キーワード：差分スペクトル法，DNN 声質変換，最小位相フィルタ，広帯域声質変換，サブバンド処理，フィルタ打ち切り，データドリブンな位相

Takaaki Saeki[1,a)]　Yuki Saito[1,b)]　Shinnosuke Takamichi[1,c)]　Hiroshi Saruwatari[1,d)]

1. はじめに

声質変換は，ある話者の音声を別の話者の音声に変換する技術であり，音声に含まれる言語情報を保持し，パラ・非言語情報のみを変換する [1]．これにより，人間の発声器官などの物理的制約を超えた音声コミュニケーションが可能になることが期待される [2]．最も一般的な声質変換は統計的声質変換であり，変換元話者の音響特徴量を変換先話者の音響特徴量に変換するための音響モデルを構築する．近年，Deep Neural Network (DNN) に基づく声質変換 [3], [4] が広く研究されており，高い音声品質を達成するモデルが提案されている．声質変換を実応用する上では，変換音声の品質だけでなく，リアルタイム性や省計算リソース性が求められる．これまで，Gaussian mixture model [5] や DNN [6] を用いた，CPU によるリアルタイム声質変換手法が提案されており，携帯用 PC の 1 つの CPU でサンプリング周波数 16 kHz の音声をリアルタイムに変換できることが確認されている．しかしながら，これらの手法は依然として計算コストが高く，携帯用端末への適用や広帯域化のためには，変換時の計算コストを削減する必要がある．

我々は，高品質かつ比較的計算効率の高い手法として，波形ドメインでの変換手法である差分スペクトル法 [7] を用いる．差分スペクトル法では，変換元話者と変換先話者のスペクトル包絡の差分を与えるフィルタを推定し，それを変換元話者の音声波形に直接適用することによって変換を行う．この方法は，ボコーダによるエラーを回避することによって高品質な変換を実現でき，ニューラルボコーダ [8], [9], [10] と比較すると計算効率が高い．差分スペクトル法に基づく声質変換では，Mel-log spectrum approximation (MLSA) フィルタ [11] を用いた場合よりも，最小位相フィルタを用いた場合の方が高品質であることが明らかになっている [12]．最小位相フィルタを用いる

[1] 東京大学 大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan.
a) takaaki_saeki@ipc.i.u-tokyo.ac.jp
b) yuuki_saito@ipc.i.u-tokyo.ac.jp
c) shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
d) hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

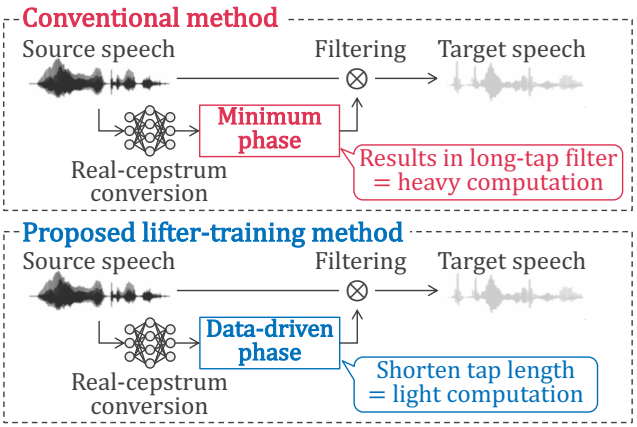

図 1　従来法とリフタ学習を用いた差分スペクトル法との比較．フィルタ打ち切りの影響を保証するようにリフタを動的に学習することで，品質を劣化させずにタップ長の短いフィルタを設計できる．

場合，DNN などの音響モデルは差分フィルタの実ケプストラムを出力し，定数値で与えられるリフタをかけてヒルベルト変換を行うことにより，時間領域の差分フィルタの位相を決定する．この方法では，フィルタ設計の計算コストは小さいものの，最小位相フィルタはフィルタのタップ長が短くなることが保証されないため，フィルタリングの計算量が増加するという問題がある．この計算量を削減する方法として，フィルタを時間領域で打ち切る [13] ことが常套手段であるが，単にフィルタ打ち切りを行った場合，計算量が削減される代わりに品質が劣化する．

そこで，本稿では，品質を劣化させずにフィルタリングの計算量を削減する方法として，フィルタ打ち切りを考慮したリフタ学習法を提案する．この方法では，フィルタの打ち切りタップ長を条件として，実ケプストラムを変換する DNN のパラメータだけでなく，実ケプストラムから位相を復元するためのリフタ係数をも同時に学習する．DNN のパラメータとリフタ係数は，フィルタ打ち切りの条件のもとで変換の精度を最大化するように学習されるため，変換音声の品質を劣化させずに計算量を削減できることが期待される．このフィルタ学習を用いた差分スペクトル法と，従来の最小位相フィルタを用いた差分スペクトル法の概念図を Fig. 1 に示す．最小位相フィルタを用いた差分スペクトル法では，ヒルベルト変換のリフタは定数であるのに対し，提案するリフタ学習法では，打ち切られたフィルタの形状変化の影響を補償するようにリフタを学習する．さらに，本研究では，差分スペクトル法を 16 kHz の狭帯域から 48 kHz の広帯域に拡張する．広帯域音声の変換に差分スペクトル法をそのまま適用した場合，広帯域の音声のランダムな変動をモデル化することが困難なため，変換音声の品質は比較的悪い．そこで，広帯域音声の変換に対し，サブバンド処理を提案する．この方法は，変換元話者の音声の低域成分のみをモデルによって変換し，高域はそのまま保持することによって，変換音声の品質向上を行う．以上の 2 つの提案法の有効性を検証するため，客観評価および主観評価を実施した．実験的評価により，(1) リフタ学習法をサンプリング周波数 16 kHz の音声の変換に適用し，変換音声の品質を劣化させずにフィルタのタップ長を 1/16 にまで削減できることを示し，(2) 提案するサブバンド処理を 48 kHz サンプリングの音声の変換に適用し，変換音声の品質を大幅に向上できることを示す．

2. 従来法：最小位相フィルタを用いた差分スペクトル法

差分スペクトル法に基づく声質変換では，最小位相フィルタを用いることによって，MLSA フィルタを用いた場合よりも高品質な変換音声が得られることが知られている [12]．本節では，最小位相フィルタを用いた差分スペクトル法の，学習時・変換時の処理について述べる．

2.1 学習時

まず，変換元話者の音声波形を短時間フーリエ変換することにより，複素スペクトル系列 $\boldsymbol{F}^{(\mathrm{X})} = [\boldsymbol{F}_1^{(\mathrm{X})\top}, ..., \boldsymbol{F}_t^{(\mathrm{X})\top}, ..., \boldsymbol{F}_T^{(\mathrm{X})\top}]^\top$ を得る．ただし，t はフレームインデックスで，T は総フレーム数である．以降，フレーム t のみに着目して議論する．$\boldsymbol{F}_t^{(\mathrm{X})}$ から低次実ケプストラム $\boldsymbol{C}_t^{(\mathrm{X})}$ を抽出し [14]，これを DNN の入力として，差分フィルタの低次実ケプストラム $\boldsymbol{C}_t^{(\mathrm{D})}$ を推定する．このとき，変換音声の低次実ケプストラム $\hat{\boldsymbol{C}}_t^{(\mathrm{Y})}$ は $\hat{\boldsymbol{C}}_t^{(\mathrm{Y})} = \boldsymbol{C}_t^{(\mathrm{X})} + \boldsymbol{C}_t^{(\mathrm{D})}$ と書け，フレーム t に関する損失関数 L_t は式 (1) のように求まる．

$$L_t = (\boldsymbol{C}_t^{(\mathrm{Y})} - \hat{\boldsymbol{C}}_t^{(\mathrm{Y})})^\top (\boldsymbol{C}_t^{(\mathrm{Y})} - \hat{\boldsymbol{C}}_t^{(\mathrm{Y})}) \qquad (1)$$

ただし，$\boldsymbol{C}_t^{(\mathrm{Y})}$ は変換先話者の自然音声の低次実ケプストラムである．このとき，式 (2) の損失関数 L を最小化するように DNN のパラメータを学習する．

$$L = \frac{1}{T}\sum_{t=1}^{T} L_t \qquad (2)$$

2.2 変換時

変換時は，まず学習済みの DNN によって差分フィルタの低次実ケプストラム $\boldsymbol{C}_t^{(\mathrm{D})}$ を推定する．さらに，$\boldsymbol{C}_t^{(\mathrm{D})}$ の高次の項を 0 埋めし，式 (3) によって表される最小位相化のためのリフタ係数 $\boldsymbol{u}_{\min}$ [15] を掛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 $\boldsymbol{F}_t^{(\mathrm{D})}$ を得る．

$$\boldsymbol{u}_{\min}(n) = \begin{cases} 1 & (n = 0, n = N/2) \\ 2 & (0 < n < N/2) \\ 0 & (n > N/2) \end{cases} \qquad (3)$$

ただし，N は周波数ビン数である．この $\boldsymbol{F}_t^{(\mathrm{D})}$ をフーリエ

変換することにより，時間領域での差分フィルタを得る．その後，この差分フィルタを変換元話者の音声波形に対して畳み込むことにより変換を行う．ここで，変換時に単純なフィルタ打ち切りを行う場合，計算量が削減できる代わりに音質が劣化する．

3. 提案法

ここでは，本研究で提案するフィルタ打ち切りを考慮したリフタ学習とサブバンド処理について述べる．

3.1 フィルタ打ち切りを考慮したリフタ学習法

フィルタ打ち切りを考慮したリフタ学習法では，フィルタの打ち切りを微分可能な形で学習過程に組み込み，DNN のパラメータのみならずヒルベルト変換のためのリフタ係数をも更新する．

3.1.1 学習時

学習時の処理を Fig. 2 に示す．まず，2.1 節と同様に，DNN によって差分フィルタの低次実ケプストラム $\boldsymbol{C}_t^{(\mathrm{D})}$ を推定する．$\boldsymbol{C}_t^{(\mathrm{D})}$ の高次の項を 0 埋めし，学習によって更新するリフタ係数 $\boldsymbol{u}$ を掛ける．さらに逆フーリエ変換して exp を取ることにより，差分フィルタの複素スペクトル系列 $\boldsymbol{F}_t^{(\mathrm{D})}$ を得る．これを逆フーリエ変換して時間領域の差分フィルタ $\boldsymbol{f}_t^{(\mathrm{D})}$ とし，式 (4) のように窓関数 $\boldsymbol{w}$ を適用することによって打ち切りを行う．

$$\boldsymbol{f}_t^{(l)} = \boldsymbol{f}_t^{(\mathrm{D})} \cdot \boldsymbol{w} \tag{4}$$

$$\boldsymbol{w} = \left[\overset{0\text{th}}{1}, \cdots, \overset{(l-1)\text{th}}{1}, \overset{l\text{th}}{0}, \cdots, \overset{(N-1)\text{th}}{0}\right]^{\top} \tag{5}$$

時間領域で打ち切られたフィルタ $\boldsymbol{f}_t^{(l)}$ を再度フーリエ変換し，タップ長 l の差分フィルタの複素スペクトル系列 $\boldsymbol{F}_t^{(l)}$ を得る．変換音声の複素スペクトル系列 $\hat{\boldsymbol{F}}_t^{(\mathrm{Y})}$ は，$\boldsymbol{F}^{(\mathrm{X})}$ と $\boldsymbol{F}_t^{(l)}$ の要素積を取ることによって得られる．さらに，$\hat{\boldsymbol{F}}_t^{(\mathrm{Y})}$ から変換音声の実ケプストラム $\hat{\boldsymbol{C}}_t^{(\mathrm{Y})}$ を抽出する．学習時に用いる損失関数は式 (1) と同じだが，DNN のパラメータだけでなく，リフタ係数をも学習する．学習の全過程において微分可能であり，誤差逆伝播法によりパラメータ更新を行うことができる [16]．

3.1.2 変換時

変換時には，学習済みの DNN とリフタ係数で $\boldsymbol{F}_t^{(\mathrm{D})}$ を推定する．これをフーリエ変換して時間領域のフィルタ $\boldsymbol{f}_t^{(\mathrm{D})}$ とし，タップ長 l で打ち切ることによって $\boldsymbol{f}_t^{(l)}$ を得る．これを変換元話者の自然音声に直接適用することによって変換音声を得る．

3.2 サブバンド処理

ここでは，広帯域音声の変換におけるサブバンド処理について述べる．従来の最小位相フィルタに基づく差分スペクトル法を 48 kHz サンプリングに拡張した場合，高域のランダムに変動する成分を変換することにより，変換音声の品質が低下するという問題が生じる．そこで，低域成分と高域成分を別々に変換するサブバンド処理を提案する．具体的には，8 kHz 以下の周波数域のみに差分フィルタを適用し，8 kHz より大きな成分はフィルタを適用せずに変換音声を得る．この処理を，(1) $\boldsymbol{F}_t^{(\mathrm{D})}$ の絶対値から 1.0 を差し引き，(2) 8 kHz 周辺で 1 から 0 に遷移するシグモイド関数をかけ，(3) 1.0 を再度加算することによって実現する．

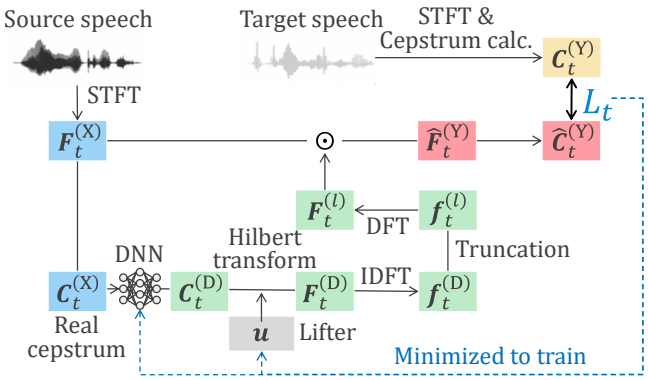

図 2 リフタ学習法での学習過程．

3.3 提案手法の分析及び他手法との関連

従来法では，ケプストラムにリフタ係数を乗じ，最小位相となるようにフィルタの形状を決定する．フィルタの形状は打ち切りによって変化するが，リフタを学習することにより，打ち切りの影響を補償するように位相が変換されると考えられる．結果的に，リフタ学習法により，変換音声の品質劣化を抑制しながらフィルタリングの計算量を削減できる．Fig. 3 に，従来法 ($l = 512$) とリフタ学習 ($l = 32$) による差分フィルタの累積パワー分布を示す．図より，差分フィルタのパワーが 0 から 100 までの領域に集中していることが確認できる．また，Fig. 4 に，従来法 (最小位相化) のリフタ係数と提案法で学習されたリフタ係数の値をそれぞれ示す．従来法のリフタが，ケプストラムの次元に対して一定値であるのに対し，リフタ学習法で得られたリフタは，次元ごとに異なることが分かる．

1 節で示したように，リフタリングに基づく位相推定は，変換時のフィルタ設計の計算量が小さくて済むという特長がある．リフタ学習法を用いた差分スペクトル法についても，位相推定の計算コストは従来法と同一である．

本研究では，リフタ学習法を声質変換に用いているが，この方法は音源分離や音声強調といった，フィルタリングを用いる他のタスクにも適用可能である．

さらに，サブバンド処理についても議論する．音声波形の特性は帯域ごとに大きく異なるため，帯域ごとに個別に波形を処理することが効果的である．サブバンド WaveNet [17] では，音声波形は複数の帯域に分割され，ダウンサンプリ

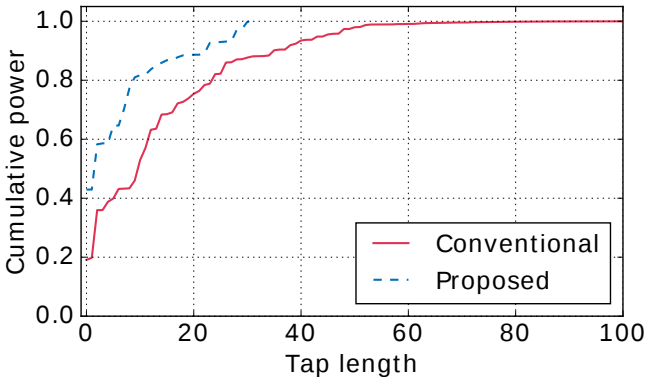

図 3　従来法 ($l = 512$) とリフタ学習法 ($l = 32$) による差分フィルタの累積パワー分布．縦軸の値は総和で正規化されている．

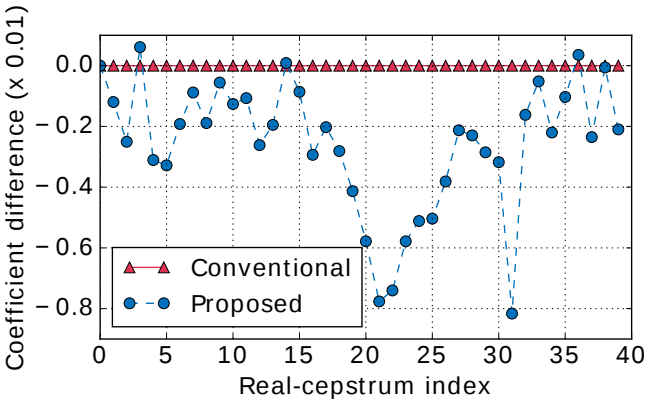

図 4　最小位相化のためのリフタ係数とリフタ学習法 ($l = 64$) で得られたリフタ係数の値．いずれも，最小位相化のためのリフタ係数との差分を取っている．

ングされて個別に処理される．本研究で提案するサブバンド処理では，フルバンド音声の各周波数帯域の波形はダウンサンプリングを行わない．リアルタイム変換を実装する場合は，3 つの周波数帯域に分割し，各帯域の波形をダウンサンプリングすることにより，フィルタリングの計算コストを約 1/3 に削減できる．

4. 実験的評価

4.1 実験条件

男性話者から男性話者 (m2m)，女性話者から女性話者 (f2f) の 2 種類の変換について実験を行った．男性の変換元話者・変換先話者にはいずれも JVS コーパス [18] の男性話者を用いた．女性の変換元話者には JSUT コーパス [19] の女性話者，変換先話者には声優統計コーパス [20] 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 16 kHz サンプリング音声と 48 kHz サンプリング音声を用いた．16 kHz サンプリング音声に短時間フーリエ変換を行う際は，窓長を 25 ms，フレームシフトを 5 ms，FFT 長を 512 点，低次ケプストラムの次元を 40 とした．48 kHz サンプリング音声に短時間フーリエ変換を行う際の窓長とフレームシフトは 16 kHz の場合と同一であり，FFT 長を 2048 点，低次ケプストラムの次元を 120 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．

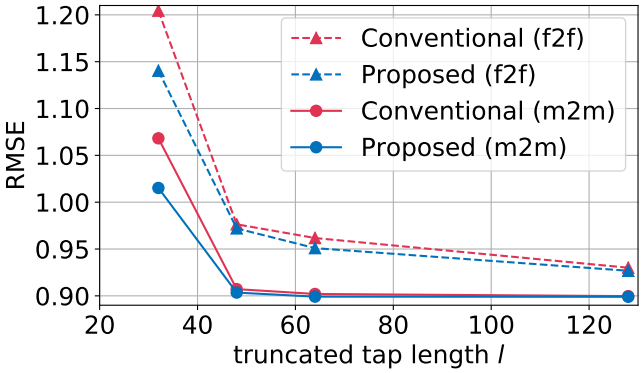

図 5　従来法とリフタ学習法の RMSE.

実験に用いた DNN アーキテクチャは，隠れ層 2 層の feedforward neural network とした．隠れユニット数は，16 kHz サンプリングの音声を変換する場合はそれぞれ 280，100 とし，48 kHz サンプリングの音声を変換する場合はそれぞれ 840，300 とした．隠れ層の活性化関数として，sigmoid 関数，tanh 関数からなる gated linear unit [21] を持ち，各々の活性化関数に通す前に batch normalization [22] を行った．また，最適化手法には Adam [23] 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 に正規化した．バッチサイズとエポック数はそれぞれ 1000，100 とし，リフタ学習法の DNN パラメータは，最小位相フィルタに基づく差分スペクトル法の学習後の値で初期化し，その際のリフタ係数は最小位相化のためのリフタ係数の値で初期化した．16 kHz サンプリングの音声に従来法とリフタ学習法を適用する際の学習率をそれぞれ 0.0005，0.00001 とし，48 kHz サンプリングの音声に従来法とリフタ学習法を適用する際の学習率はそれぞれ 0.0001，0.000005 とした．

提案法のうち，リフタ学習法は 16 kHz サンプリング音声と 48 kHz サンプリング音声の両ケースで評価した．サンプリング周波数 16 kHz の場合の打ち切りタップ長 l は 128，64，48，32 の 4 ケースとし，サンプリング周波数 48 kHz の場合の l は 224，192 の 2 ケースとした．サブバンド処理は，サンプリング周波数 48 kHz の音声のみに従来法を適用することによって評価した．

4.2 客観評価実験

打ち切りタップ長 l を変化させ，従来法と提案法を test データに対して用いた場合の Root Mean Squared Error (RMSE) を比較した．RMSE は式 (2) の平方根を取った

表 1　サンプリング周波数 16 kHz の音声に従来法とリフタ学習法を用いた場合ののプリファレンススコア

(a) 話者類似性

Lifter training	Score	p-value	Conventional
l : 32 (m2m)	**0.587** vs. 0.413	1.3×10^{-5}	l : 32 (m2m)
l : 32 (m2m)	0.463 vs. 0.537	7.3×10^{-2}	l : 512 (m2m)
l : 32 (f2f)	**0.642** vs. 0.358	$< 10^{-10}$	l : 32 (f2f)
l : 32 (f2f)	**0.543** vs. 0.457	3.4×10^{-2}	l : 512 (f2f)
l : 48 (m2m)	0.533 vs. 0.467	1.0×10^{-1}	l : 48 (m2m)
l : 48 (m2m)	**0.550** vs. 0.450	1.4×10^{-2}	l : 512 (m2m)
l : 48 (f2f)	**0.613** vs. 0.387	1.3×10^{-8}	l : 48 (f2f)
l : 48 (f2f)	**0.548** vs. 0.452	2.0×10^{-2}	l : 512 (f2f)

(b) 音質

Lifter training	Score	p-value	Conventional
l : 32 (m2m)	**0.687** vs. 0.313	$< 10^{-10}$	l : 32 (m2m)
l : 32 (m2m)	0.529 vs. 0.471	2.3×10^{-1}	l : 512 (m2m)
l : 32 (f2f)	**0.807** vs. 0.193	$< 10^{-10}$	l : 32 (f2f)
l : 32 (f2f)	**0.742** vs. 0.258	$< 10^{-10}$	l : 512 (f2f)
l : 48 (m2m)	**0.606** vs. 0.394	8.7×10^{-8}	l : 48 (m2m)
l : 48 (m2m)	0.523 vs. 0.477	2.6×10^{-1}	l : 512 (m2m)
l : 48 (f2f)	**0.581** vs. 0.419	5.5×10^{-5}	l : 48 (f2f)
l : 48 (f2f)	0.513 vs. 0.487	5.1×10^{-1}	l : 512 (f2f)

値である．Fig. 5 は，サンプリング周波数 16 kHz の音声に対し，男性から男性，女性から女性の変換を行った場合の RMSE を示している．l の全ケースについて，リフタ学習法は従来法よりも高精度な変換を実現していることが確認できる．従来法とリフタ学習法の RMSE の差は，l が小さくなるとより顕著になる傾向が見て取れる．この結果から，リフタ学習法はフィルタ打ち切りの影響を軽減できていることがわかる．

4.3　主観評価実験

4.3.1　リフタ学習法の評価

クラウドソーシングを用いて音質に関する AB テストおよび話者類似性に関する XAB テストを行った．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 テストの参照音声 X には変換先話者の自然音声を用いた．

表 1 は，サンプリング周波数 16 kHz の音声を用いた場合の，従来法とリフタ学習法の比較を示している．m2m は男性から男性への変換を，f2f は女性から女性への変換を表している．従来法でフィルタを打ち切った場合 (“Conventional ($l = 32, 48$)”) と比較すると，リフタ学習法でフィルタを打ち切った場合 (“Proposed ($l = 32, 48$)”) は話者類似性・音質のいずれについても高いスコアを示していることが確認できる．また，従来法でフィルタを打ち切らない場合 (“Conventional ($l = 512$)”) と比較しても，リフタ学習法 (“Proposed ($l = 32, 48$)”) は同等かより高いスコアを示していることが見て取れる．この結果から，従来法でフィルタを打ち切った場合は変換音声の品質が大きく劣化する一方，リフタ学習法により，品質を劣化させずにタップ長を 1/16 まで削減できることが分かる．表 2 にサンプリング周波数 48 kHz の場合の結果を示す．この場合も，サンプリング周波数 16 kHz の場合と同様の傾向が見られ，リフタ学習法で品質を落とさずにタップ長を $l = 224$ に削減できるが，$l = 192$ では話者類似性・音質が劣化する場合がある．したがって，48 kHz サンプリングの場合も，16 kHz サンプリングの場合ほどではないが，フィルタのタップ長を大幅に短くできることが分かる．

表 2　サンプリング周波数 48 kHz の音声に従来法とリフタ学習法を用いた場合ののプリファレンススコア

(a) 話者類似性

Lifter training	Score	p-value	Conventional
l : 192 (m2m)	0.461 vs. **0.569**	4.9×10^{-4}	l : 2048 (m2m)
l : 192 (f2f)	0.519 vs. 0.481	3.4×10^{-1}	l : 2048 (f2f)
l : 224 (m2m)	0.474 vs. 0.526	2.0×10^{-1}	l : 2048 (m2m)
l : 224 (f2f)	0.519 vs. 0.481	3.4×10^{-1}	l : 2048 (f2f)

(b) 音質

Lifter training	Score	p-value	Conventional
l : 192 (m2m)	0.529 vs. 0.471	2.3×10^{-1}	l : 2048 (m2m)
l : 192 (f2f)	0.447 vs. **0.553**	8.9×10^{-3}	l : 2048 (f2f)
l : 224 (m2m)	0.513 vs. 0.487	5.2×10^{-1}	l : 2048 (m2m)
l : 224 (f2f)	0.517 vs. 0.483	4.2×10^{-1}	l : 2048 (f2f)

表 3　サンプリング周波数 48 kHz の音声に従来法とサブバンド処理を用いた場合ののプリファレンススコア

(a) 話者類似性

Sub-band	Score	p-value	Conventional
m2m	0.519 vs. 0.481	3.4×10^{-1}	m2m
f2f	**0.603** vs. 0.397	5.0×10^{-7}	f2f

(b) 音質

Sub-band	Score	p-value	Conventional
m2m	**0.721** vs. 0.279	$< 10^{-10}$	m2m
f2f	**0.700** vs. 0.300	$< 10^{-10}$	f2f

4.3.2　サブバンド処理の評価

4.3.1 節と同様の方法で，サブバンド処理についても主観評価を行った．表 3 に結果を示す．従来法にサブバンド処理を適用した場合を評価しており，フィルタ打ち切りもリフタ学習法も用いていない．結果から，サブバンド処理により，話者類似性・音質のいずれについても大幅に改善されていることが分かる．

5.　おわりに

本稿では，差分スペクトル法に基づく声質変換に対し，リフタ学習法およびサブバンド処理を提案した．リフタ学習法により，品質を劣化させずにフィルタリングの計算量を大きく削減でき，サブバンド処理により，広帯域の変換音声の品質を改善できることを示した．

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] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. ICASSP*, New York, U.S.A., Apr. 1988, pp. 655–658.

[2] T. Toda, "Augmented speech production based on real-time statistical voice conversion," in *Proc. GlobalSIP*, Atlanta, U.S.A, Dec. 2014, pp. 592–596.

[3] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 3893–3896.

[4] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4869–4873.

[5] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion," in *Proc. INTERSPEECH*, Portland, U.S.A., Sep. 2012, pp. 94–97.

[6] 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.

[7] 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, 2018.

[8] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 1118–1122.

[9] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv*, vol. abs/1609.03499, 2018.

[10] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 5916–5920.

[11] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan*, vol. 66, no. 2, pp. 10–18, 1983.

[12] 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.

[13] M. Sunohara, C. Haruta, and N. Ono, "Low-latency real-time blind source separation with binaural directional hearing aids," in *1st International Workshop on Challenges in Hearing Assistive Technology*, Stockholm, Sweden, Aug. 2017, pp. 9–13.

[14] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, San Francisco, U.S.A., Mar 1992, pp. 137–140.

[15] S.-C. Pei and H.-S. Lin, "Minimum-phase FIR filter design using real cepstrum," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 10, pp. 1113–1117, 2006.

[16] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[17] T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "Subband WaveNet with overlapped single-sideband filterbanks," in *Proc. ASRU*, Okinawa, Japan, Dec. 2017, pp. 698–704.

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

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

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

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

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv*, vol. abs/1502.03167, 2015.

[23] D. Kingma and B. Jimmy, "Adam: A method for stochastic optimization," *arXiv*, vol. abs/1412.6980, 2014.