差分スペクトル法に基づくDNN声質変換の計算量削減に向けたフィルタ推定*

☆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

統計的声質変換の一手法として，かねてより差分スペクトル法が提案されている [1] [2]．この手法は，変換元話者と変換先話者のスペクトル包絡成分の差分を与えるフィルタを推定し，それを元話者の音声波形に適用することによって話者を変換する手法であり，ボコーダによる音質劣化を回避できる．このフィルタ設計に関して，最小位相フィルタの使用は，従来の MLSA (Mel-Log Spectrum Approximation) フィルタ [1] よりも高い変換音声品質を達成できることが明らかになっている [2]．しかし，最小位相フィルタは，時刻 0 にパワーが集中するものの，MLSA フィルタと比較してタップ長の短さを保証しない．故に，最小位相フィルタは，変換時の畳み込みに必要な計算量を増大させるため，リアルタイム声質変換 [3] に不向きである．また，フィルタを短いタップ長で打ち切ることは演算量低減の常套手段だが，打ち切りによる変換音声品質の低下は免れない．

そこで本稿では，その演算量低減を目的としたフィルタの推定法を提案する．具体的には，フィルタが固定タップ長で打ち切られることを条件とし，その条件下で実ケプストラムの推定誤差が最小となるように，実ケプストラムに施すヒルベルト変換のリフタを音声データから学習する．実験的評価では，客観評価・主観評価の両軸から提案法の有効性を示す．

2　差分スペクトル法に基づく DNN 声質変換

本節では，DNN (Deep Neural Network) 声質変換における，最小位相フィルタを用いた差分スペクトル法を概説する．

2.1　学習時

変換元話者の複素スペクトル系列 $\boldsymbol{F}^{(\mathrm{X})} = [\boldsymbol{F}_1^{(\mathrm{X})\top}, ..., \boldsymbol{F}_T^{(\mathrm{X})\top}]$ から抽出した低次の実ケプストラム系列を $\boldsymbol{C}^{(\mathrm{X})} = [\boldsymbol{C}_1^{(\mathrm{X})\top}, ..., \boldsymbol{C}_T^{(\mathrm{X})\top}]$ とする．ただし，T は総時間フレーム数であり，ベクトルの各成分は各時間フレーム t での値である．$\boldsymbol{C}^{(\mathrm{X})}$ を入力として，DNN で差分フィルタの実ケプストラム系列 $\boldsymbol{C}^{(\mathrm{D})} = [\boldsymbol{C}_1^{(\mathrm{D})\top}, ..., \boldsymbol{C}_T^{(\mathrm{D})\top}]$ を推定する．DNN 学習時の損失関数は，$\hat{\boldsymbol{C}}^{(\mathrm{Y})} = \boldsymbol{C}^{(\mathrm{X})} + \boldsymbol{C}^{(\mathrm{D})}$ と変換先話者の低次の実ケプストラム系列 $\boldsymbol{C}^{(\mathrm{Y})} = [\boldsymbol{C}_1^{(\mathrm{Y})\top}, ..., \boldsymbol{C}_T^{(\mathrm{Y})\top}]$ の二乗誤差として式 (1) で与えられる．

$$L = \frac{1}{T}(\boldsymbol{C}^{(\mathrm{Y})} - \hat{\boldsymbol{C}}^{(\mathrm{Y})})^\top(\boldsymbol{C}^{(\mathrm{Y})} - \hat{\boldsymbol{C}}^{(\mathrm{Y})}) \qquad (1)$$

2.2　変換時

まず，差分フィルタの実ケプストラム系列 $\boldsymbol{C}^{(\mathrm{D})}$ を DNN によって推定する．その後，高次成分を 0 埋めして式 (2) で表される最小位相化のためのリフタを掛け，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系列 $\boldsymbol{F}^{(\mathrm{D})} = [\boldsymbol{F}_1^{(\mathrm{D})\top}, ..., \boldsymbol{F}_T^{(\mathrm{D})\top}]$ を得る．

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

ただし，N は周波数ビン数である．この $\boldsymbol{F}^{(\mathrm{D})}$ 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った場合，計算量が削減できる代わりに音質が劣化する．

3　リフタ学習に基づく計算量削減法

2.2 節に述べた音質劣化は，学習時にフィルタ打ち切りが考慮されていないためである．そのため提案法では，フィルタ打ち切りを計算過程に含め，DNN のパラメータのみならずリフタ係数 $\boldsymbol{u}(n)$ を学習する．

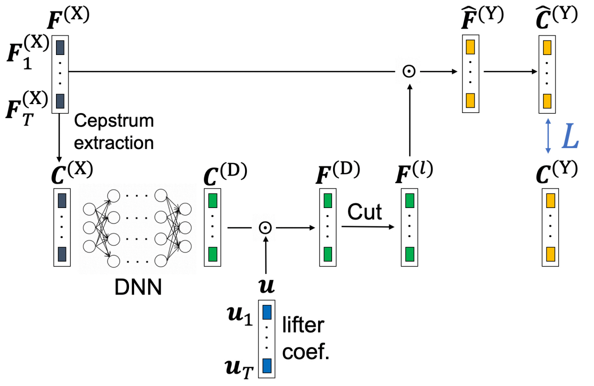

Fig. 1　提案法を用いた学習

3.1　学習時

学習時の処理を Fig. 1 に示す．まず差分フィルタの実ケプストラム系列 $\boldsymbol{C}^{(\mathrm{D})}$ を求め，これに学習により更新するリフタ係数 $\boldsymbol{u}(n)$ を掛け，フーリエ変換することによって差分フィルタの複素スペクトル系列 $\boldsymbol{F}^{(\mathrm{D})}$ を得る．これを逆フーリエ変換することにより，時間領域のフィルタに変換し，これに対し，時刻 l より前を 1，l 以降を 0 とした窓関数をかけることにより打ち切りを行う．これを再度フーリエ変換することにより，タップ長 l の差分フィルタの複素スペクトル系列 $\boldsymbol{F}^{(l)} = [\boldsymbol{F}_1^{(l)\top}, ..., \boldsymbol{F}_T^{(l)\top}]$ が得られる．これと変換元話者の複素スペクトル系列 $\boldsymbol{F}^{(\mathrm{X})}$ を掛ける事により変換後の複素スペクトル系列 $\hat{\boldsymbol{F}}^{(\mathrm{Y})}$ が得られ，さらに $\hat{\boldsymbol{F}}^{(\mathrm{Y})}$ から低次の実ケプストラム系列 $\hat{\boldsymbol{C}}^{(\mathrm{Y})} = [\hat{\boldsymbol{C}}_1^{(\mathrm{Y})\top}, ..., \hat{\boldsymbol{C}}_T^{(\mathrm{Y})\top}]$ を抽出する．学習における損失関数は，これと変換先話者の低次の実ケプストラム係数 $\boldsymbol{C}^{(\mathrm{Y})}$ の二乗誤差として式 (3) で与えられる．

$$L = \frac{1}{T}(\boldsymbol{C}^{(\mathrm{Y})} - \hat{\boldsymbol{C}}^{(\mathrm{Y})})^\top(\boldsymbol{C}^{(\mathrm{Y})} - \hat{\boldsymbol{C}}^{(\mathrm{Y})}) \qquad (3)$$

---

* Filter Estimation for Computational Complexity Reduction of DNN-based Voice Conversion Using Spectral Differentials by Takaaki Saeki, Yuki Saito, Shinnosuke Takamichi, Hiroshi Saruwatari (The University of Tokyo)

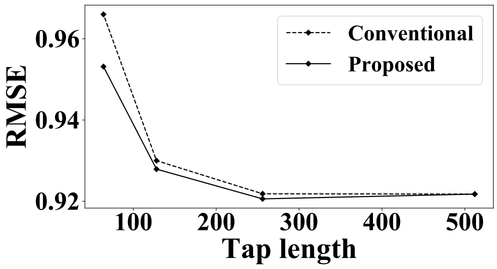

Fig. 2　客観評価結果

この損失関数を最小化するように，DNN と $\boldsymbol{u}(n)$ を学習する．$\boldsymbol{u}(n)$ の学習により，フィルタ打ち切りが実ケプストラムに与える影響を抑えられ，より短いタップ長のフィルタでも高品質な変換を可能にすると期待される．なお，提案法の計算過程は全て微分可能であるため，Back Propagation により学習できる．

3.2　変換時

学習した DNN とリフタ係数により，差分フィルタの複素スペクトル系列 $\boldsymbol{F}^{(\mathrm{D})}$ を推定し，これを逆フーリエ変換して時間領域に移し，ある固定長 l で打ち切る．これ以降の処理は 2.2 節と同様である．

4　実験的評価

4.1　実験条件

本実験の変換元話者には JSUT コーパス [4] の女性話者，変換先話者には声優統計コーパスの女性話者を用いた．サンプリング周波数を 16 kHz とし，変換元話者，変換先話者それぞれの全発話の 8 割を train データ，1 割を validation データ，1 割を test データに用いた．

短時間フーリエ変換を行う際には，窓長 25 ms，フレームシフト 5 ms，FFT 長 512 点とし，低次のケプストラムの打ち切りは 40 次元とした．

学習した DNN は 280 次元，100 次元の 2 層の中間層からなる Multi Layer Perceptron である．また，隠れ層の活性化関数として Gated Linear Unit [6] を持ち，各々の活性化関数に通す前に Batch Normalization [5] を適用した．最適化手法には Adam [7] を用いた．学習時には，元話者と目標話者のケプストラムを平均 0，分散 1 に正規化した．従来法・提案法ともにバッチサイズ 1000，epoch 数は 100 である．提案法では，従来法の学習後の DNN パラメータとリフタ (式 (2)) の値を初期値として学習を行なった．

4.2　客観評価

タップ長が 512 (すなわち，打ち切りなし)，256，128，64 の場合の 4 ケースについて，提案法を用いた場合と従来法を用いた場合の test データに対する RMSE (Rooted Mean Squared Error) の収束値を比較した．Fig. 2 に結果を示す．ただし，RMSE は式 (3) の平方根を取った値である．

評価結果より，全てのタップ長について，提案法が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．また，この差はタップ長が小さい場合により顕著になることも確認できる．この結果より，提案法のリフタ学習により，フィルタ打ち切りによる影響を低減できることが分かる．

4.3　主観評価

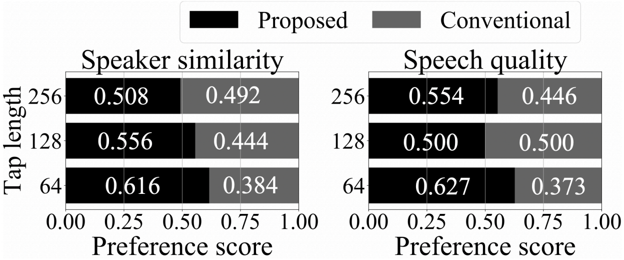

Fig. 3　提案法と従来法の比較結果

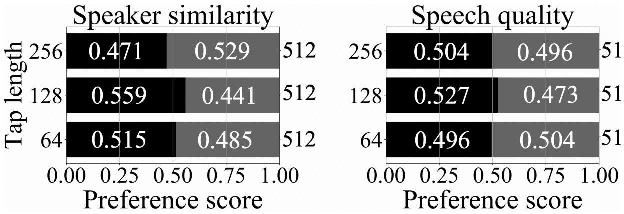

Fig. 4　提案法でのタップ長ごとの比較結果

変換音声の品質を評価するために，クラウドソーシングを用いて主観評価を実施した．まず，フィルタのタップ長が 256，128，64 となるそれぞれの場合について提案法と従来法の品質を比較した．さらに，タップ長が 512 の場合と，提案法でタップ長を 256，128，64 とした場合とを比較した．

変換音声の品質は，話者類似性・音質の 2 軸によって評価し，各評価につき 25 人の参加者がそれぞれ 10 文の test データについて評価を行なった．従来法と提案法との比較結果を Fig. 3 に示す．特にタップ長が 64 の場合，話者類似性での p 値が 1.55×10^{-7}，音質での p 値が 4.33×10^{-9} であり，話者類似性・音質の両面で，提案法の preference score が有意に高いことが確認できる．

また，タップ長が 512 の場合と，提案法でタップ長を 256，128，64 とした場合とを比較した結果を Fig. 4 に示す．全てのケースについて p 値が 0.05 より大きくなり，タップ長が 512 の場合と，提案法でタップ長を 256，128，64 とした場合との間には話者類似性・音質の両面で有意差が無いことが確認できる．従って，提案法により，品質を保持しながらタップ長を 1/8 まで短縮できることが分かる．

5　おわりに

本稿では，差分スペクトル法に基づく DNN 声質変換に対して，その計算量削減のためのフィルタ推定法を提案した．さらに，実験的評価により，提案法を用いることで品質を保持しながら計算量を削減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] K. Kobayashi *et al.*, *Speech Communication,* vol. 99, pp. 211–220, May 2018.
[2] H. Suda *et al.*, *Proc. APSIPA,* pp. 816–822, Hawaii, U.S.A., Nov. 2018.
[3] 荒川 他, 音講論 (春), 1-10-4, 2019.
[4] R. Sonobe, *et al.*, *arXiv preprint arXiv:1711.00354,* Oct. 2017.
[5] Sergey Ioffe and Christian Szegedy, *arXiv preprint arXiv:1502.03167,* Feb. 2015.
[6] Y. N. Dauphin *et al.*, *arXiv preprint arXiv:1612.08083,* Dec. 2016.
[7] Diederik P. Kingma, Jimmy Ba, *arXiv preprint arXiv:1412.6980,* Dec. 2014.