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．

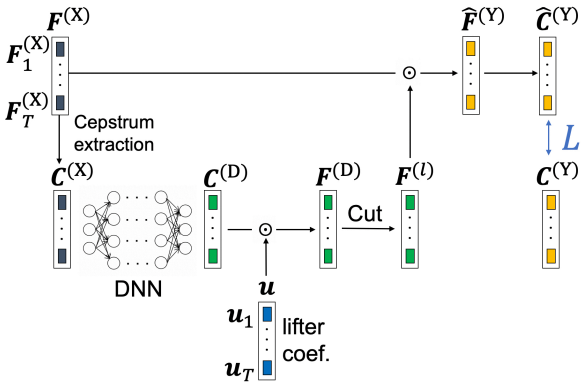

Fig. 1 提案法を用いた学習

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

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．

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)