日本音響学会 秋季研究発表会 2019

2-4-1
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

本発表の概要

研究目的:

– 高品質・低遅延な声質変換の実現

従来法: 最小位相フィルタを用いた差分スペクトル法 [Suda18]

– 高品質だが，変換時の畳み込みの計算コストが増大

– 単にフィルタ打ち切りで計算量を削減した場合，品質が劣化

提案法: フィルタ打ち切りを考慮したリフタ学習

– DNNのパラメータとヒルベルト変換のリフタ係数を同時に学習

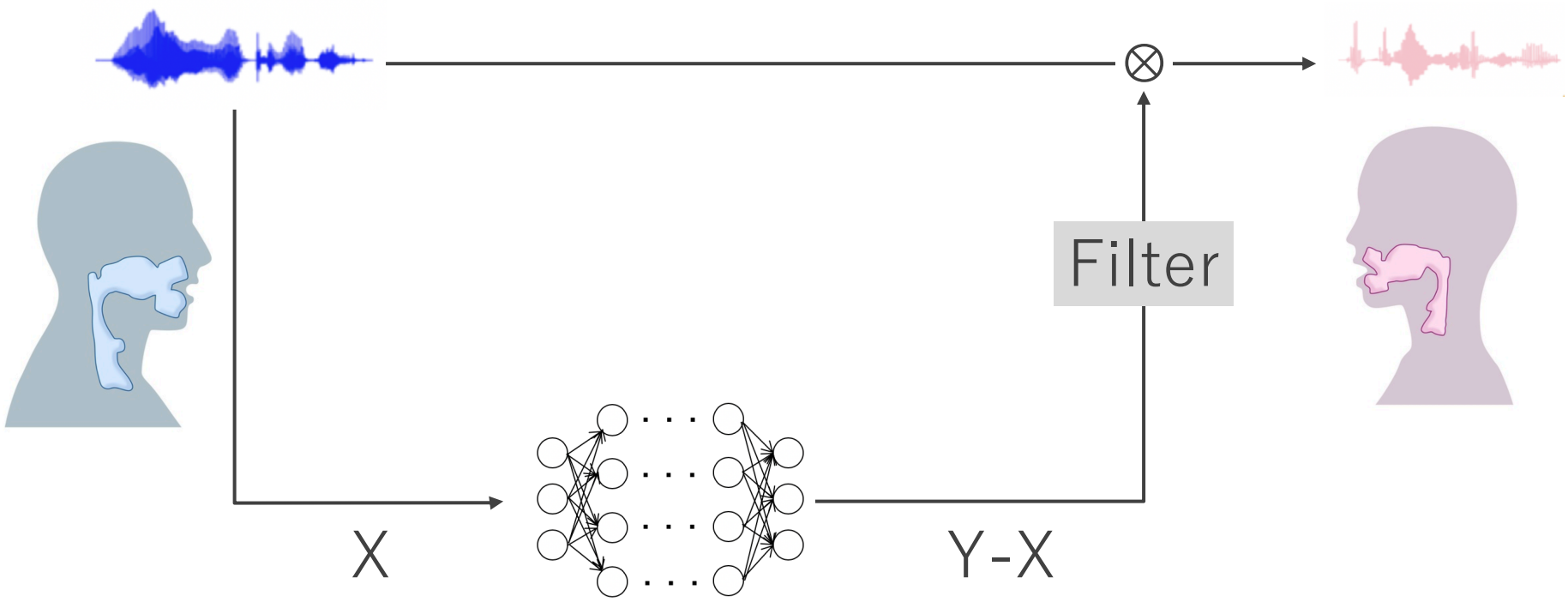
実験的評価:

– 提案法により，品質を劣化させずにタップ長を1/8まで短縮可能

差分スペクトル法による声質変換 [Kobayashi14][Suda18]

差分フィルタを推定し，変換元話者の音声波形に直接適用

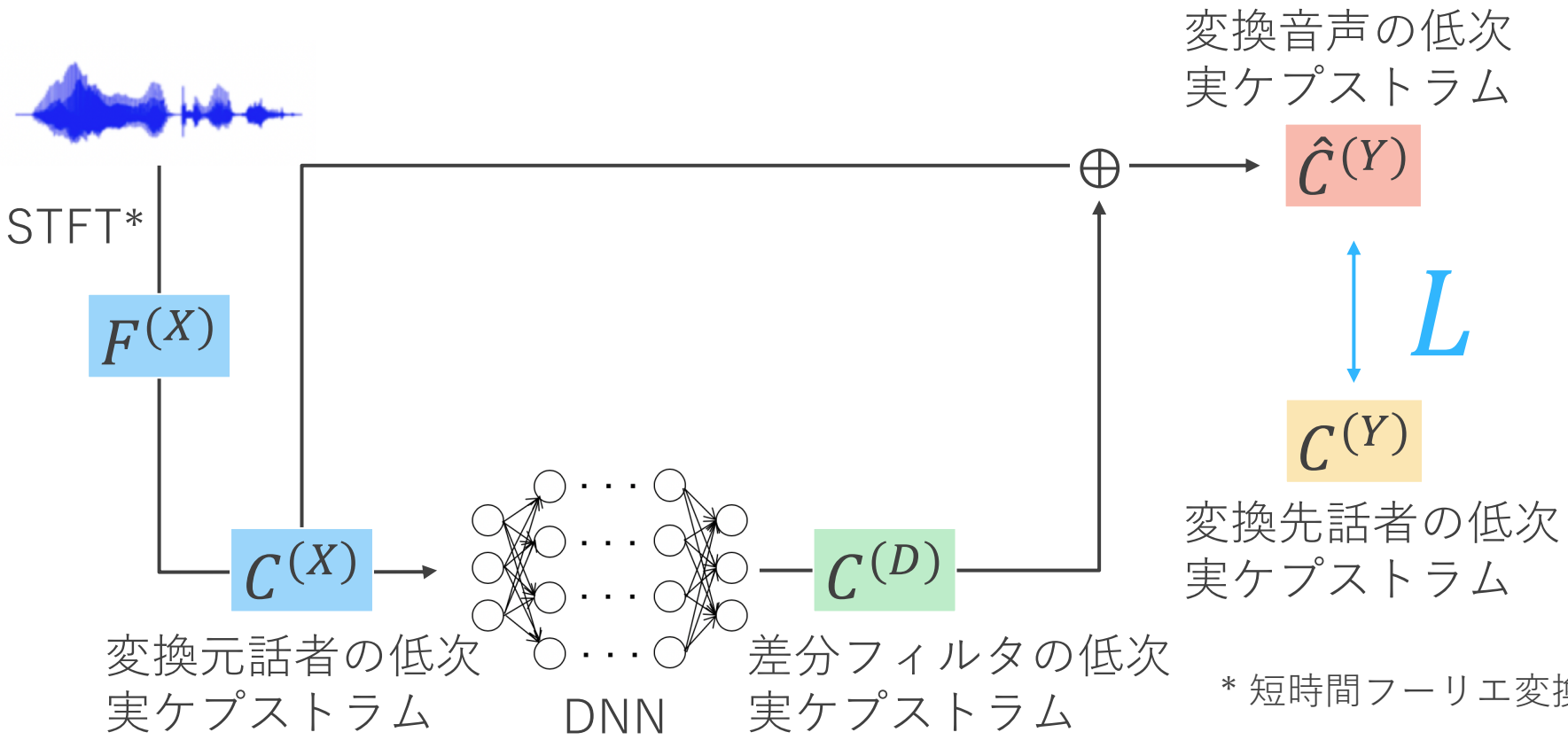
- ボコーダによる音質劣化を回避
- 最小位相フィルタによる手法は，MLSAフィルタより高品質



従来法の学習プロセス

DNNで差分フィルタのケプストラム系列を推定

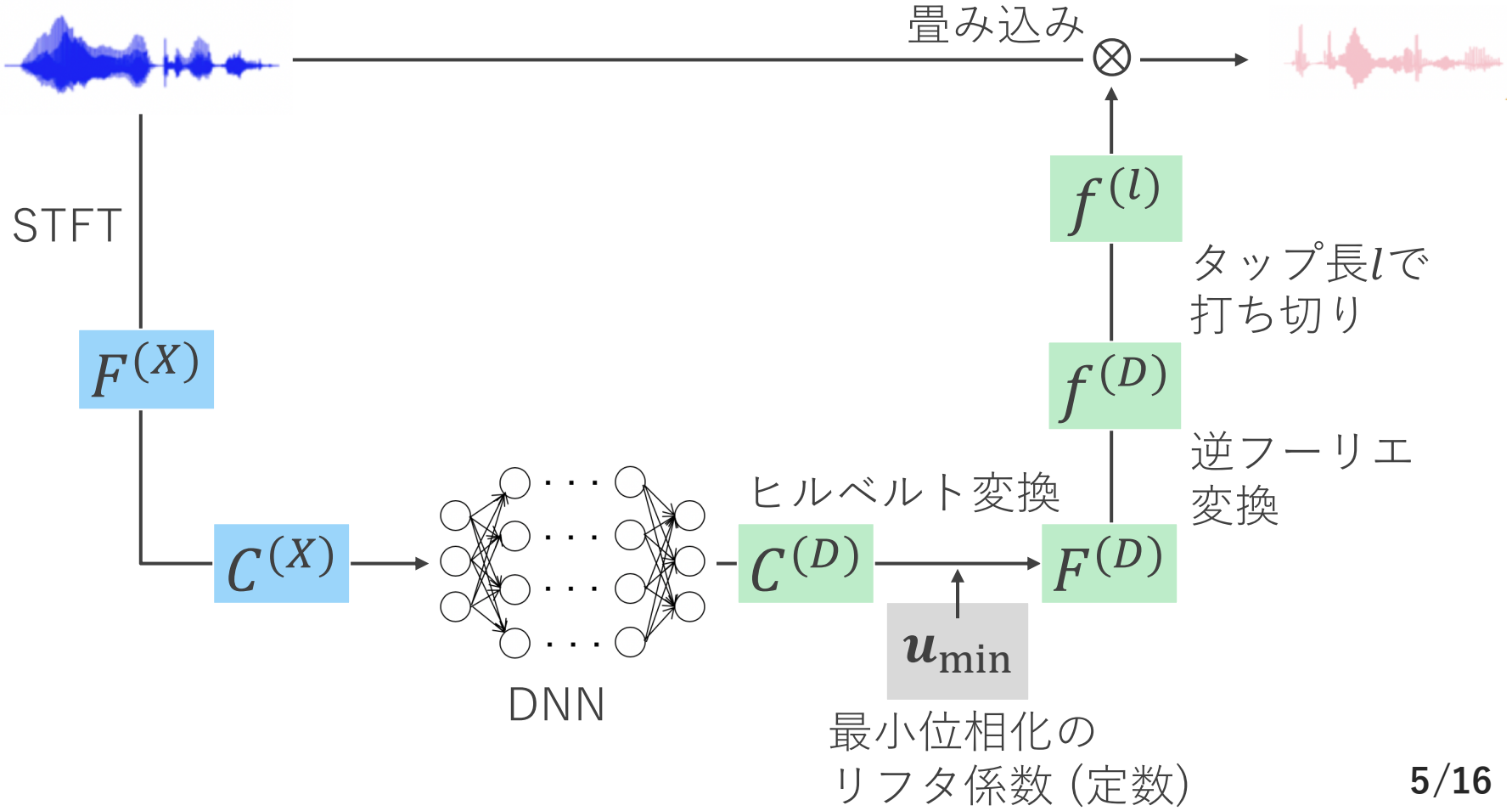
$\hat{C}^{(Y)}$ と $C^{(Y)}$ との二乗誤差を最小化するようにDNNを学習



* 短時間フーリエ変換

従来法の変換プロセス

最小位相化・ヒルベルト変換を行うことで差分フィルタを推定
計算量削減のためにタップ長lで打ち切り，畳み込み演算を行う



提案法

フィルタ打ち切りを考慮した
リフタ学習

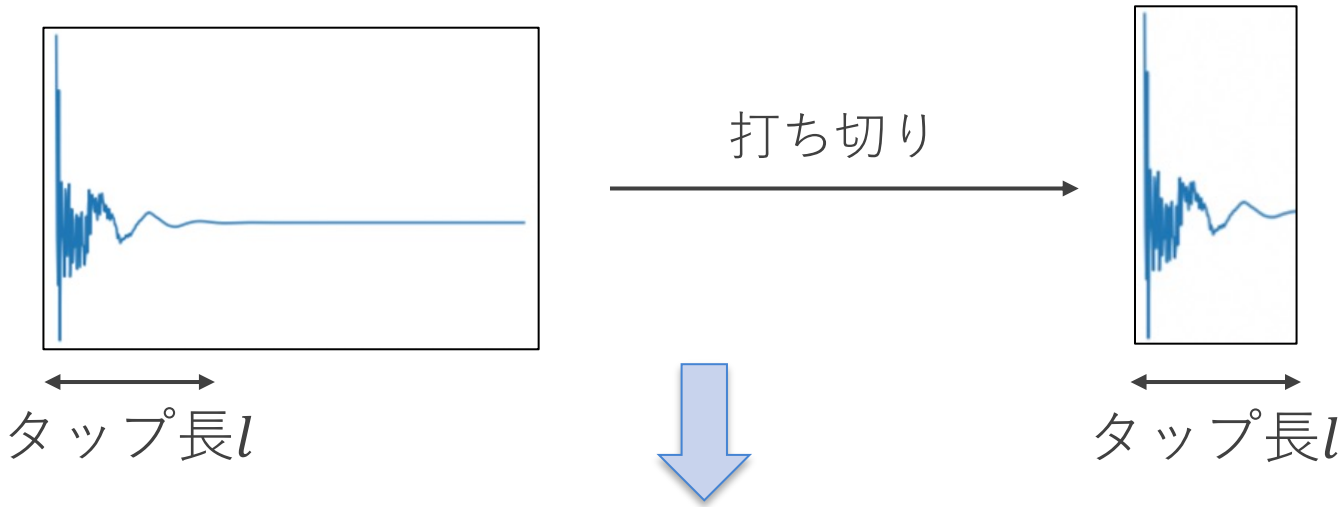
提案法: フィルタ打ち切りを考慮したリフタ学習

最小位相フィルタは，高品質な反面，計算量が大きい．

– MLSAフィルタと比較して必ずしもタップ長の短さが保証されない．

フィルタをあるタップ長で打ち切ることで，計算量削減

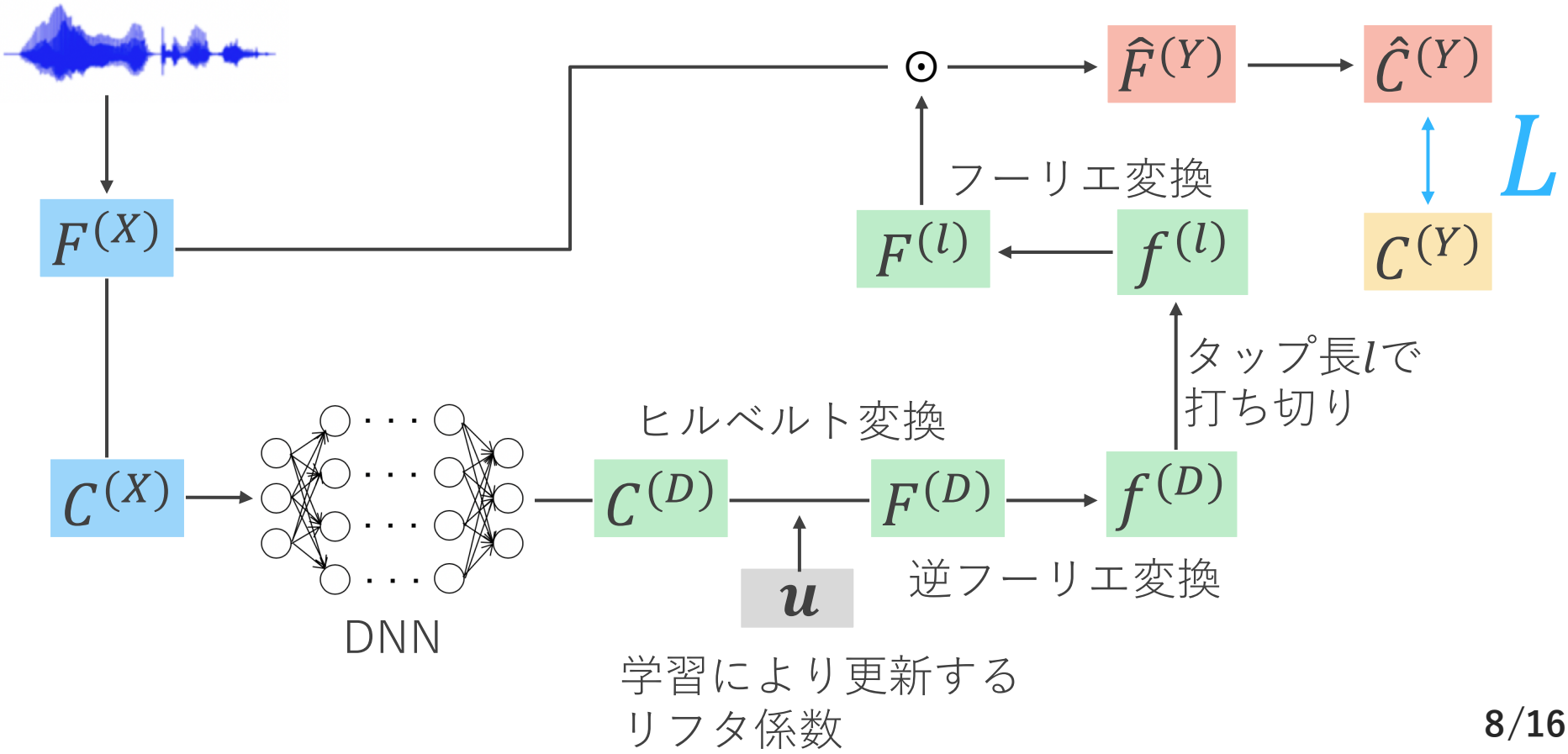
– 単に打ち切っただけでは，品質が劣化してしまう．



提案法:
フィルタの打ち切りタップ長を条件として，DNNのパラメータとヒルベルト変換のリフタを同時に学習

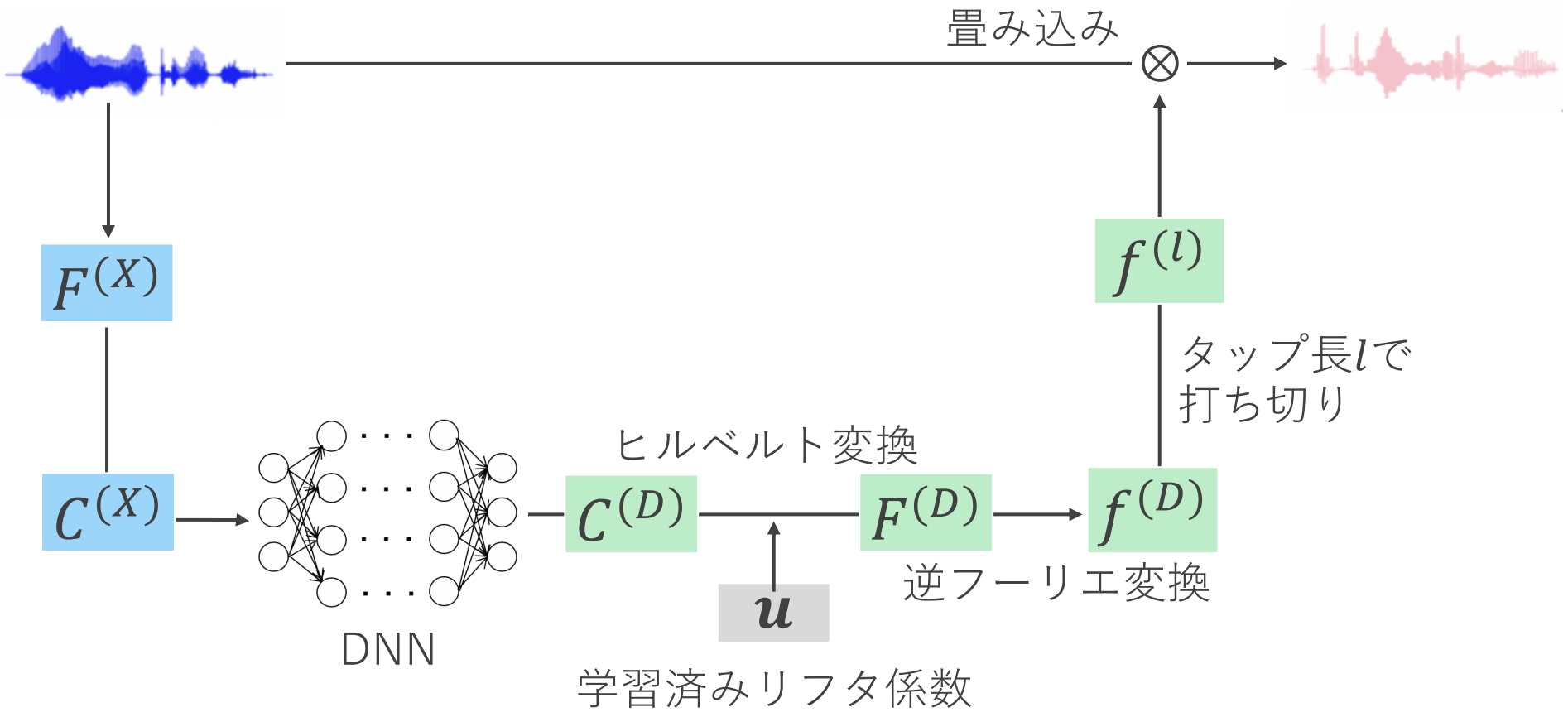
提案法の学習プロセス

フィルタ打ち切りを学習に含め，DNNとリフタ係数を同時に学習
全過程で微分可能であり，誤差逆伝播によりパラメータ更新可能



提案法の変換プロセス

学習したDNNとリフタ係数により差分フィルタを求める.
フィルタを時間領域で畳み込むことにより変換.



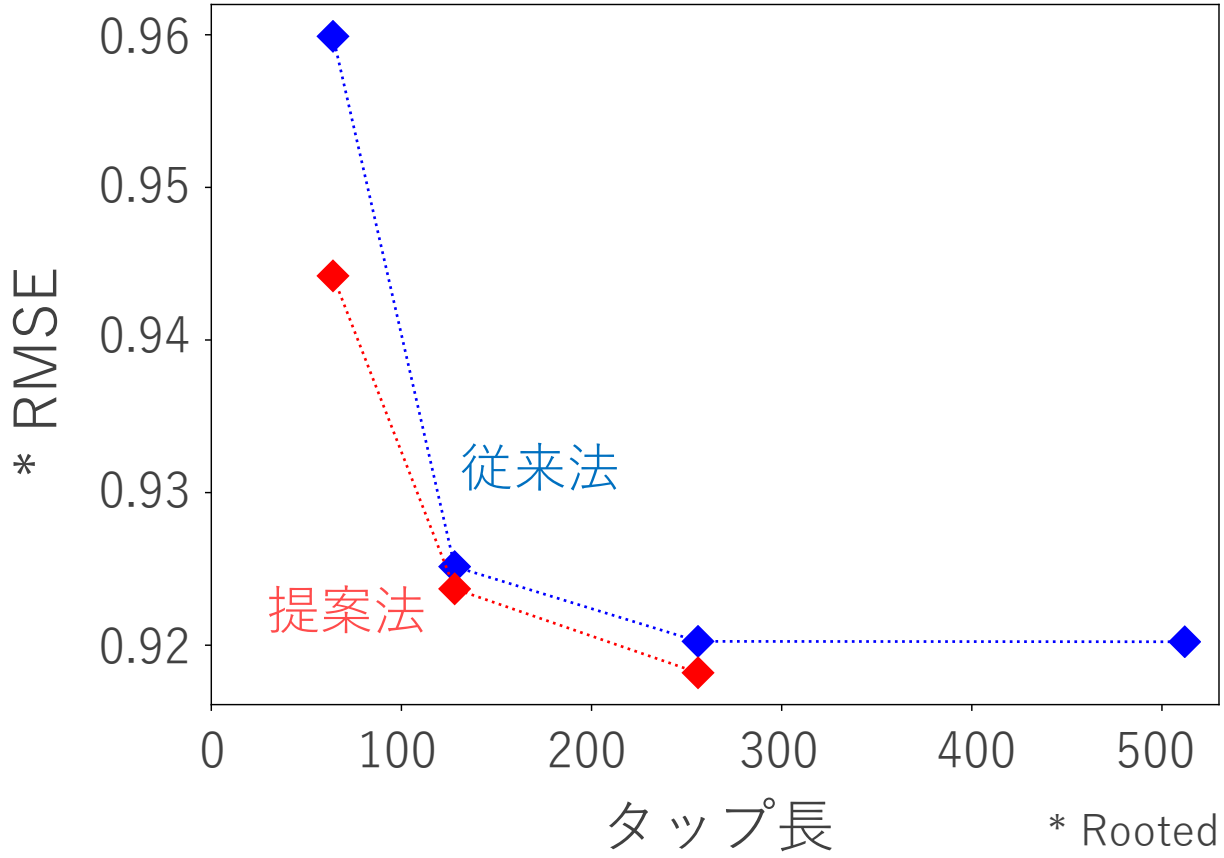
実験的評価

実験条件

提案法と従来法でフィルタ打ち切りを行い，品質を評価
タップ長が512(打ち切りなし)，256，128，64の4ケースを比較

データセット	変換元話者: JSUT corpus [Sonobe17] 変換先話者: 声優統計 corpus [y_benjo17] (100文，約12分からなるパラレルデータ)
Train/Valid/Test	80文/10文/10文
サンプリングレート	16 kHz
DNN	隠れ層2層のMulti Layer Perceptron
FFT長	512
窓長	25 ms
低次ケプストラム次数	40次

客観評価結果: TestデータでのRMSE比較

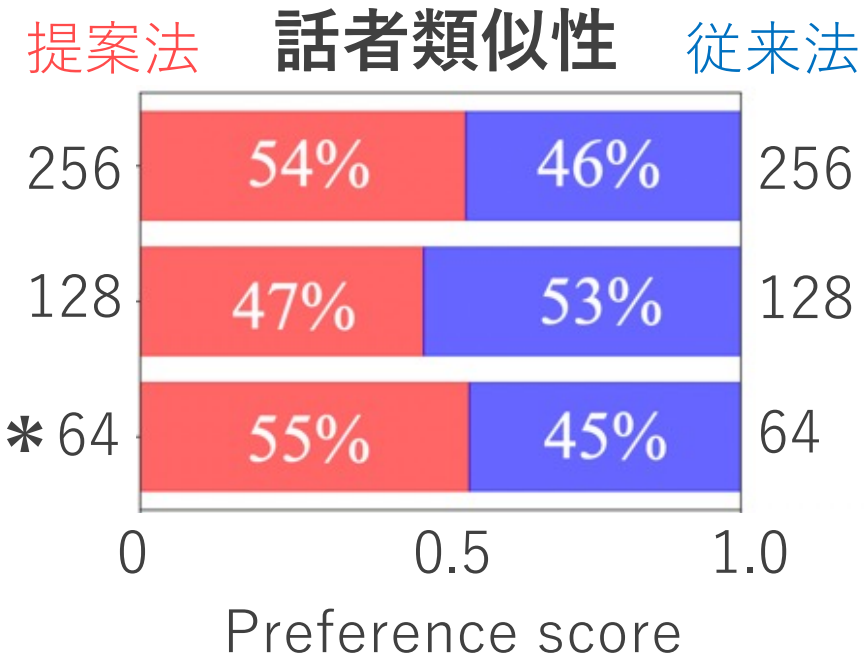


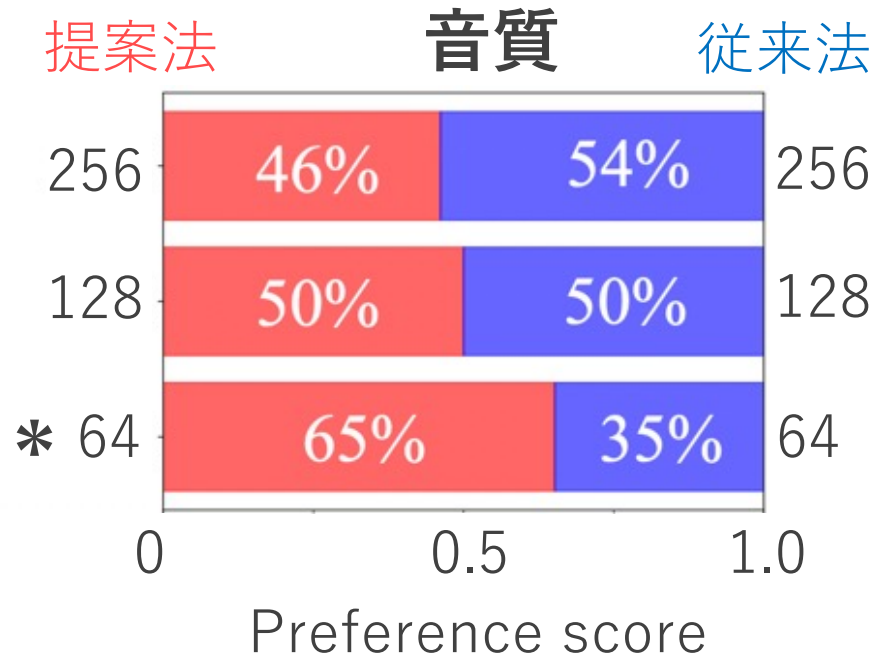
* Rooted Mean Squared Error

全ケースで提案法のRMSE < 従来法のRMSE

– 提案法は従来法よりも，フィルタ打ち切りの影響を低減

主観評価結果: 提案法と従来法の比較

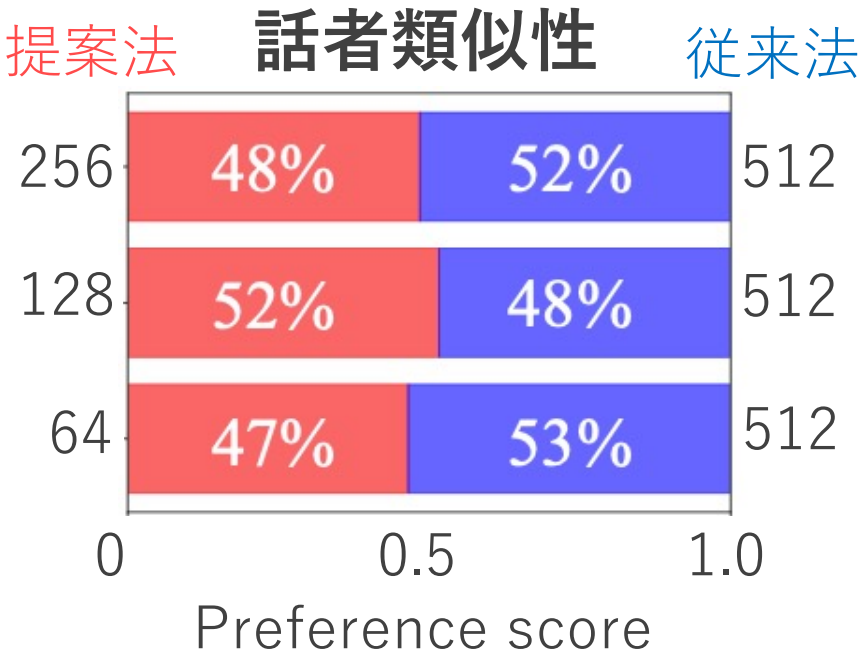


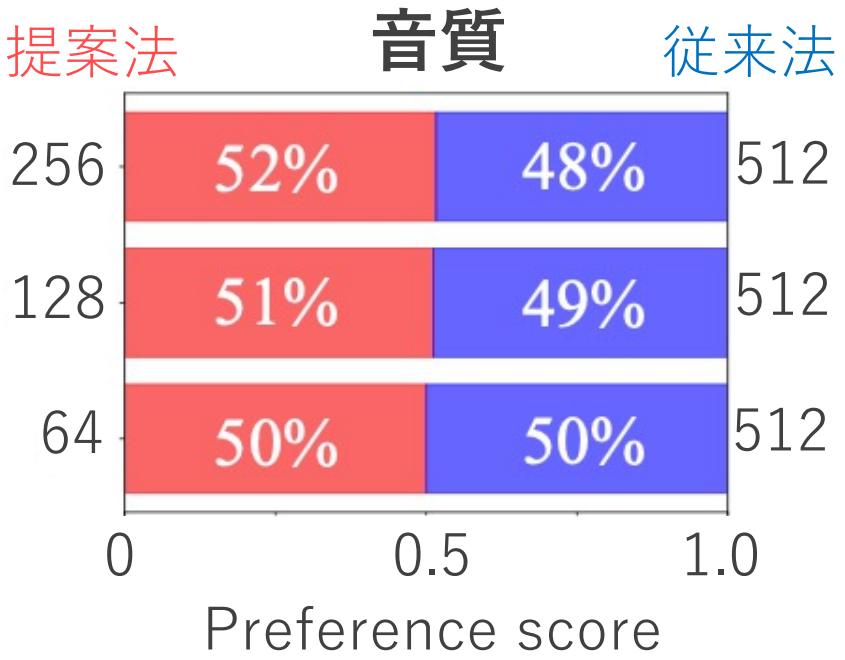


タップ長64の場合で，有意に提案法 > 従来法

– 提案法により，フィルタを打ち切ったときの品質の劣化を低減

主観評価結果: 提案法と打ち切りなしの場合との比較



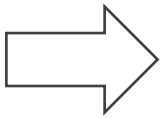


打ち切りなし(512)・打ち切った場合との間に有意差なし

– 話者類似性・音質を劣化させずにタップ長を1/8まで短縮可能

サンプル音源

変換元話者



変換先話者

	タップ長: 512	タップ長: 64
従来法		
提案法		

まとめ

研究目的:

- 高品質・低遅延な声質変換の実現

提案:

- 差分スペクトル法に基づくDNN声質変換の計算量削減法
- フィルタの打ち切りタップ長を条件として，DNNのパラメータとヒルベルト変換のリフタを同時に学習

実験結果:

- 提案法により，品質を劣化させずにタップ長を1/8まで短縮

今後の課題:

- 様々な話者データでの実験的評価の実施
- 提案法に基づく，リアルタイム・広帯域な声質変換の実現

付録

実験条件の詳細

サンプリングパラメータ	サンプリングレート	16kHz
STFTパラメータ	窓長	25ms
	FFT長	512 点
	低次ケプストラム次数	40次元
	フレームシフト	5ms
DNNパラメータ	DNNの種類	Multi Layer Perceptron
	loss	Mean squared error
	optimizer	Adam
	バッチサイズ	1000
	活性化関数	Gated Linear Unit (sigmoid, tanh)
	隠れ層の数	2層 (280次元，100次元)
	BatchNorm	隠れ層全てに適用