日本音響学会2020年 秋季講演発表会 1-2-11

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)

本発表の概要

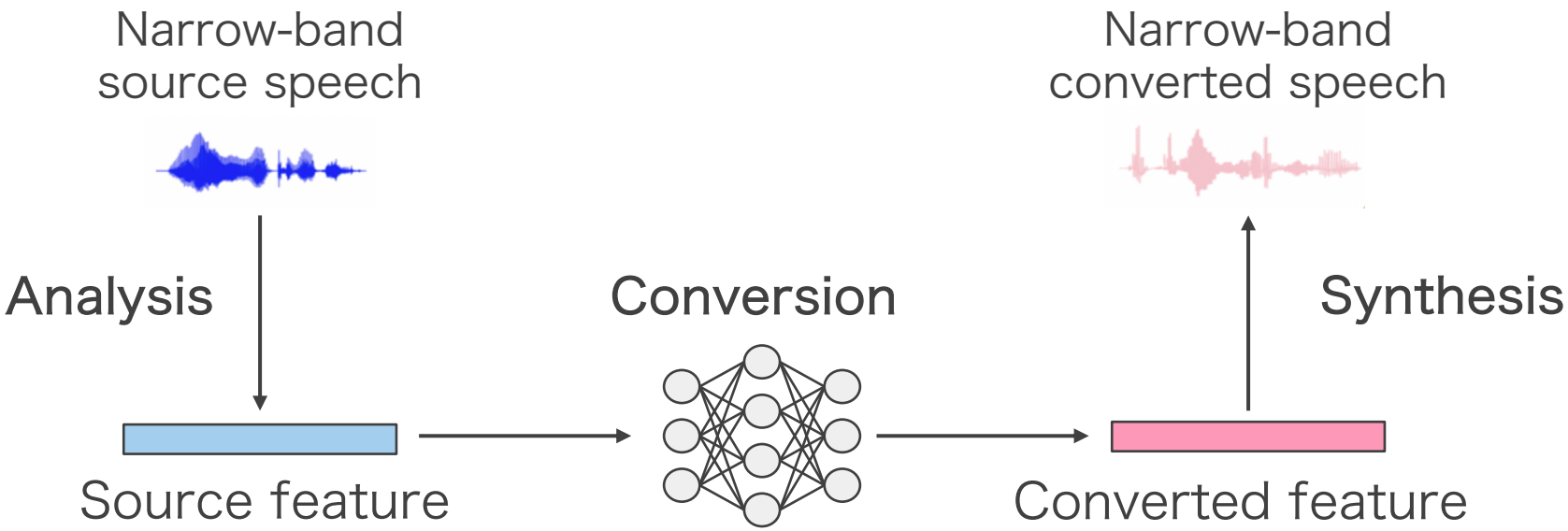
- **背景**
  - 声質変換では，話者再現度だけでなく**リアルタイム性・音質**が重要
  - 従来のリアルタイム声質変換では，狭帯域 (16 kHz) 音声のみ変換可能
  - 高音質なフルバンド (48 kHz) リアルタイム声質変換の実現が目的

- **手法**
  - サブバンド処理により低域 (0-8 kHz) のみをモデル化・変換 [佐伯20]
  - 低域変換の際にフィルタ打ち切りを行うことで計算量削減 [Saeki20]
  - **リアルタイム・オンライン処理のための実装 & F0変換機構**

- **評価結果**
  - 計算効率の評価の結果，1CPUでのリアルタイム動作を確認
  - 変換音声の品質評価の結果，Benchmark手法より有意に高い自然性

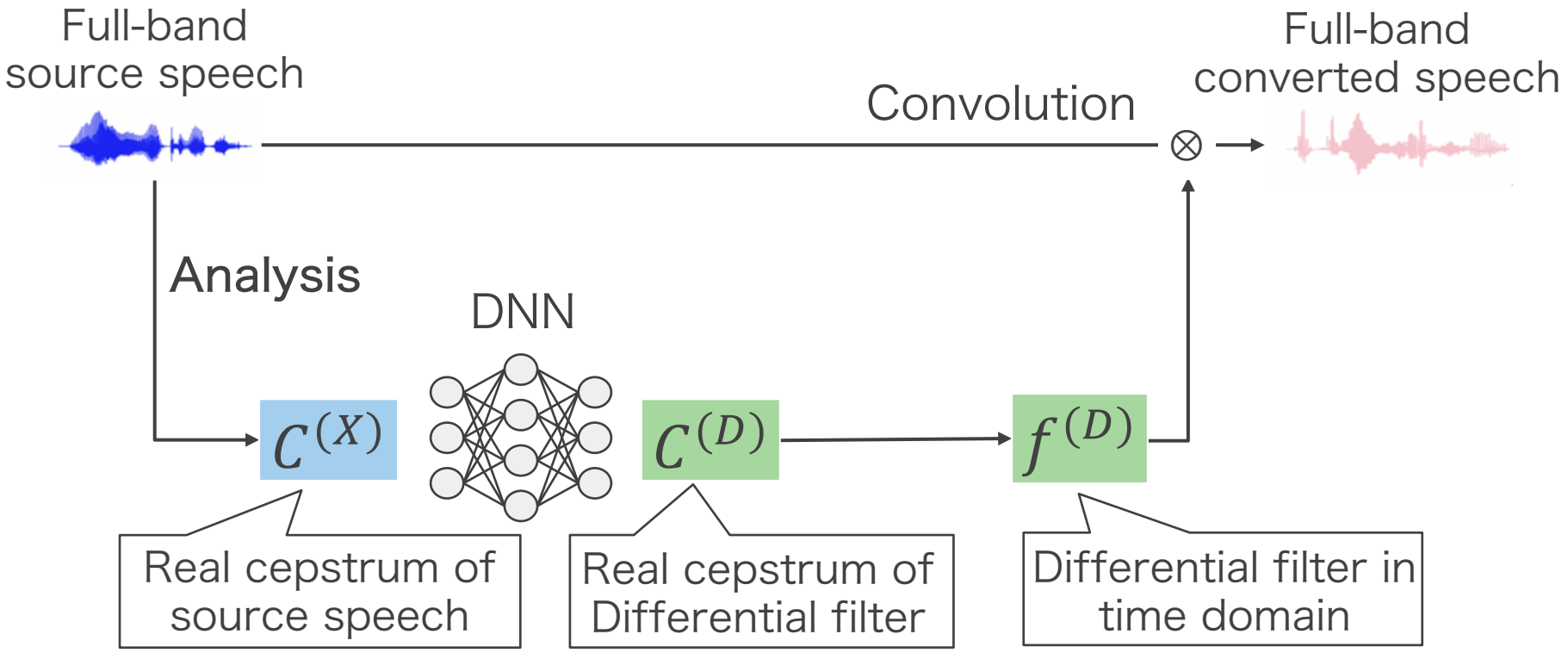
既存のリアルタイム声質変換の課題

- **従来のリアルタイム声質変換** [Toda12][Arakawa19]
  - 変換された音響特徴量から信号処理に基づくボコーダで波形生成
  - 狭帯域 (16 kHz) 音声を遅延50 ms程度で変換可能
  - 信号処理に基づくボコーダによるアーティファクト
  - 変換可能な帯域幅が狭く，変換音声が低音質



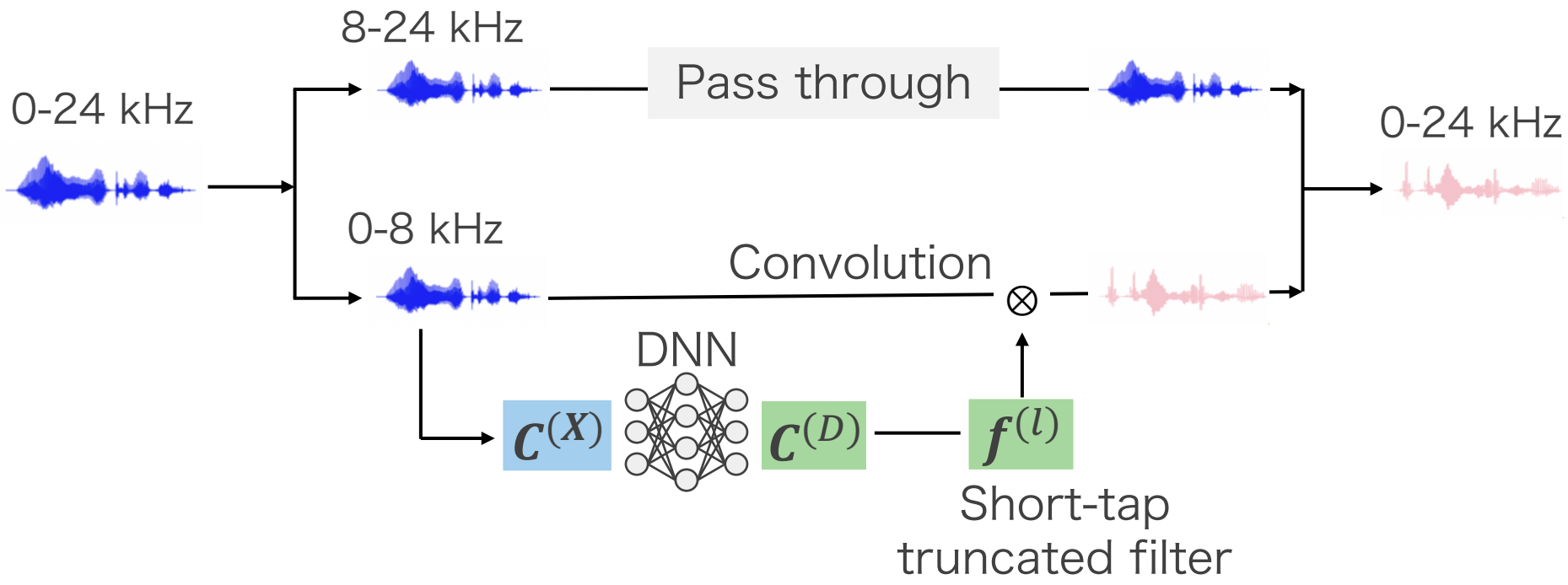
本研究でのアプローチ

- **差分スペクトル法に基づく声質変換** [Kobayashi18]
  - ボコーダを用いず，波形領域でのフィルタリングにより変換
  - 帯域拡張すると品質・計算量面で問題
    - サブバンドフィルタリングによる解決法を提案 [Saeki20]



サブバンドフィルタリングによる広帯域声質変換 [佐伯20]

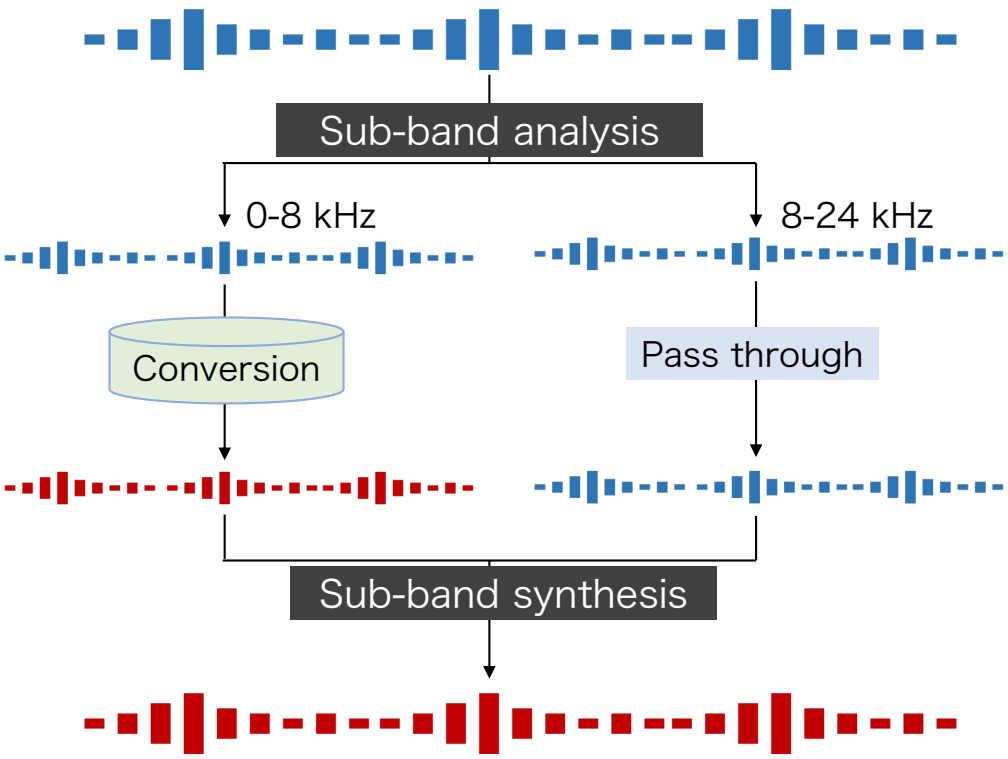
- **サブバンド処理により，0-8 kHzのみをモデル化・変換** [佐伯20]
  - 変動の大きい高域をそのまま使うことで，変換音声の音質を向上
  - フィルタをかける波形のサンプル数が減り，計算量を削減
- **フィルタのタップ長を短く打ち切ることで，さらに計算量削減** [Saeki20]



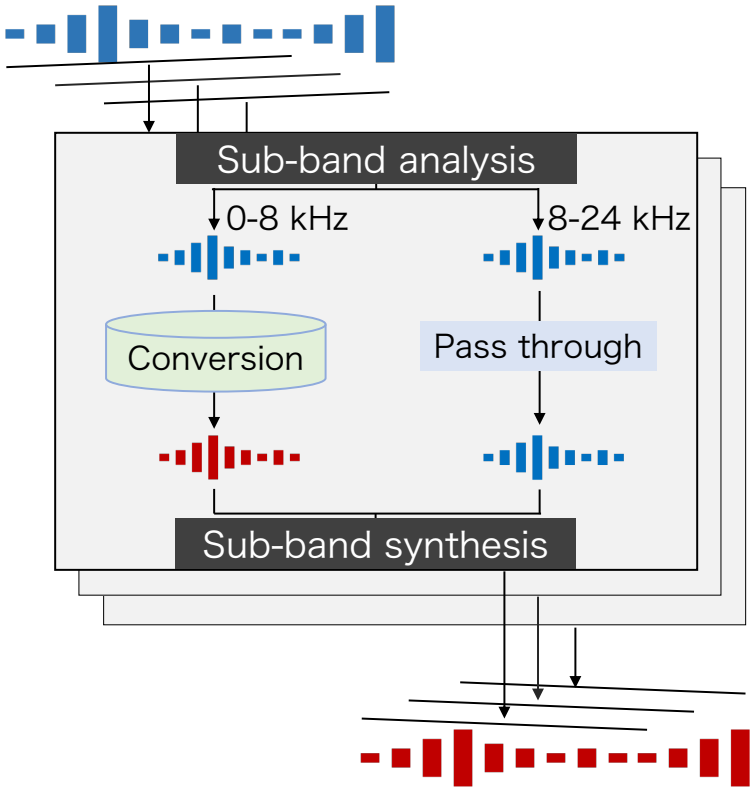
リアルタイム・オンライン広帯域声質変換

- サブバンドフィルタリングによる声質変換はオフライン変換を想定
- リアルタイムフルバンド声質変換のためのオンライン実装

Offline変換 (発話ごとに処理) [佐伯20]

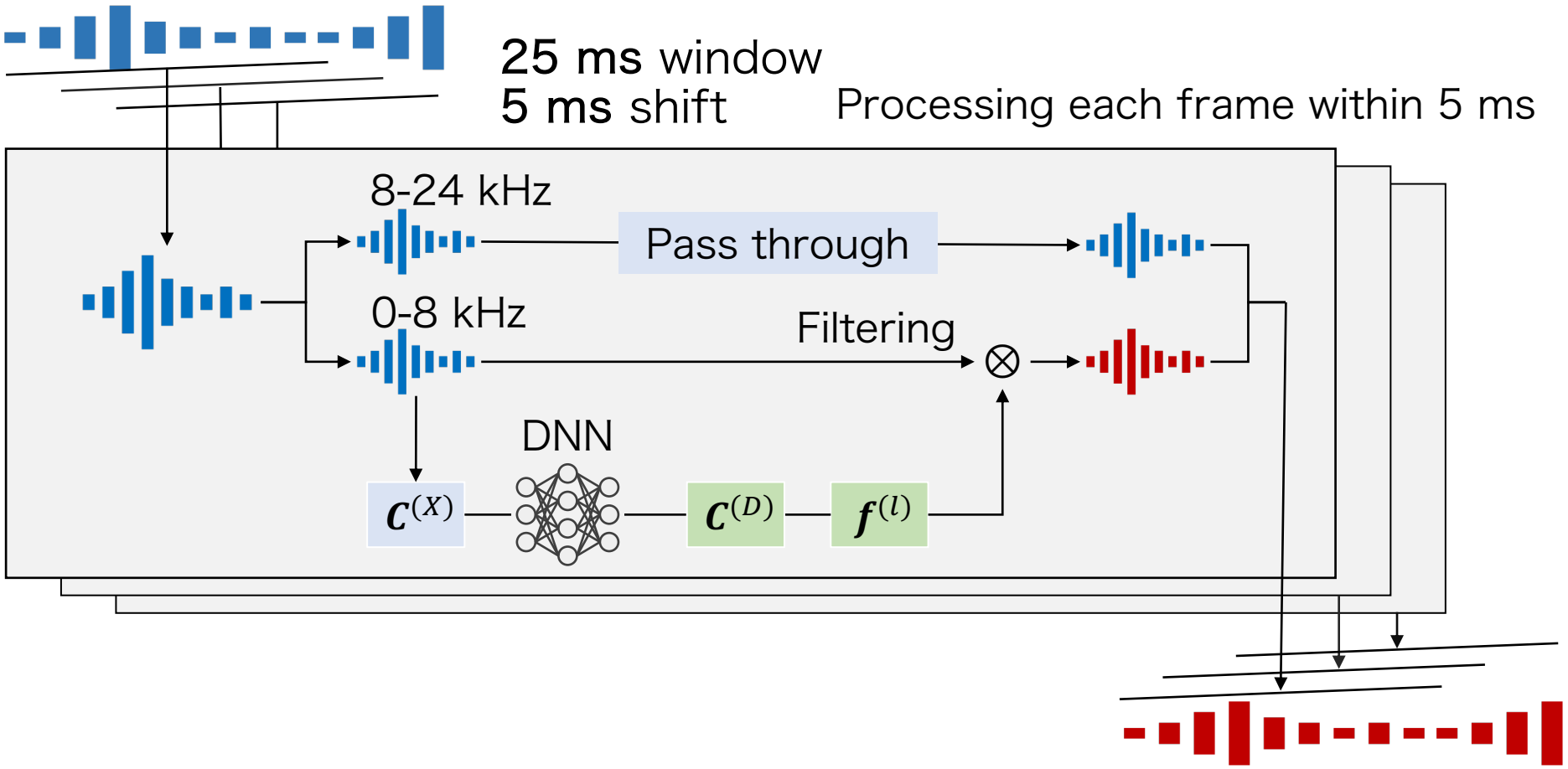


Online変換 (フレームごとに処理)



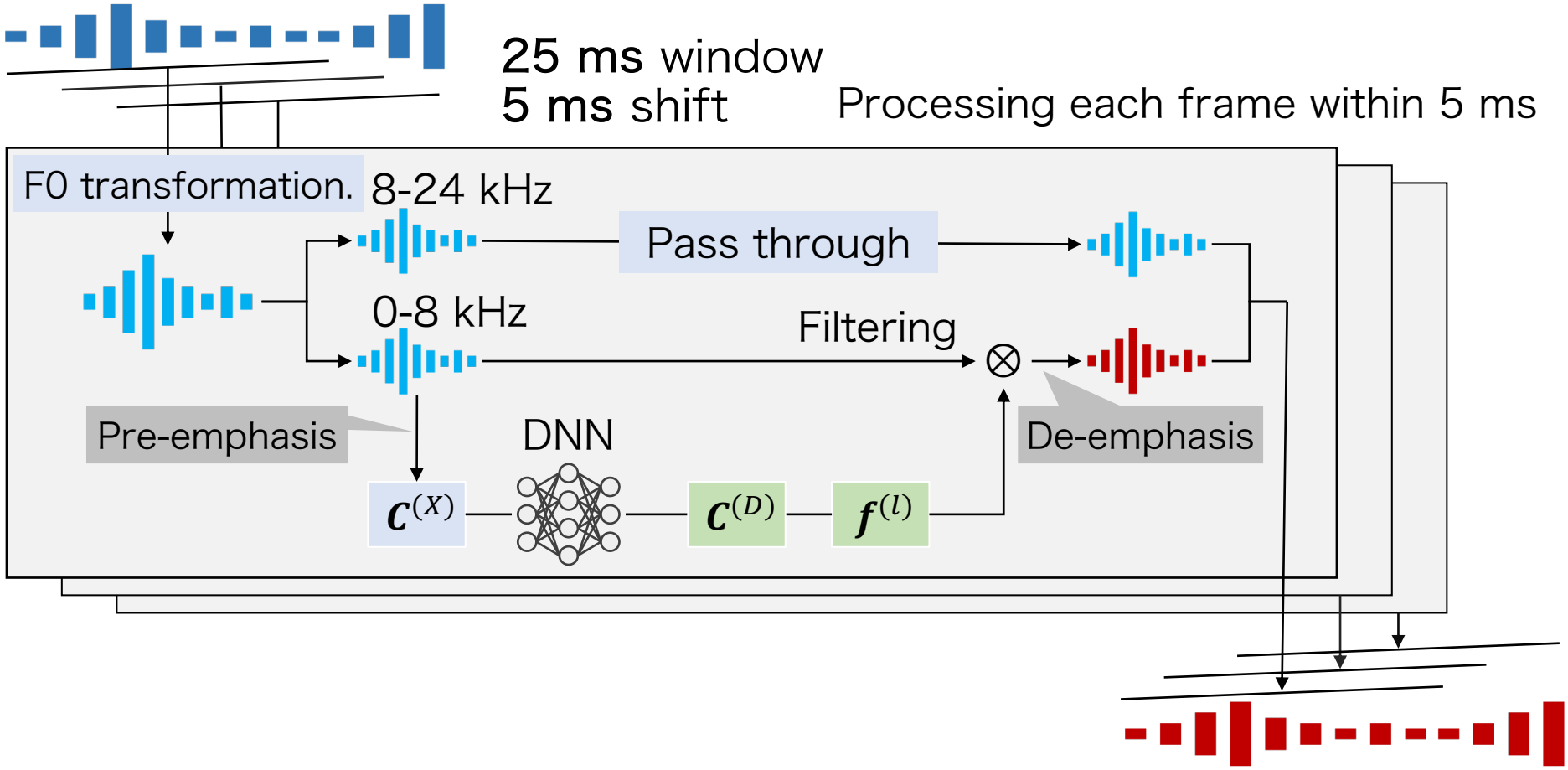
提案するシステムの主要部

- **分析部**: 変換元話者の音声をサブバンド処理し，低域から特徴量抽出
- **変換部**: DNNでケプストラム変換し，差分フィルタを推定
- **合成部**: 変換先話者の音声波形を得る



その他の部分

- **F0変換**: 入力波形を一定F0比だけPICOLA [森田1986] でピッチシフト
- 特徴量分析時に**プリエンファシス**を適用



実験的評価

評価の概要

- Proposed methodとBenchmarkを比較
  - Benchmark: 差分スペクトル法を広帯域音声に帯域分割なしで適用
- 評価内容
  - **計算効率**: FLOPSによる計算量概算 & Real-time factor (RTF) 計測
  - **変換音声品質**: 主観評価実験

Evaluation cases	male-to-male (**m2m**), female-to-female (**f2f**) female-to-male (**f2m**), male-to-female (**m2f**)
Dataset	**f2f**: JSUTcorpus [Sonobe17] Voice Actress corpus [y_benjo17] **m2m**, **f2m**, **m2f**: JVS corpus [Takamichi19]
Train / Valid / Test	80 sentences / 10 sentences / 10 sentences
DFT length	Proposed: 512 samples Benchmark: 2048 samples
DNN architecture	Multi layer perceptron with 2 hidden layers
CPU	Intel (R) core (TM) i7-6850K CPU @ 3.60 GHz

計算量の評価

- FLOPS: 1秒間に浮動小数点演算が何回行えるかという指標
- 1秒間のフルバンド音声を処理するのにかかる計算量をFLOPS単位で概算

単位はGFLOPS

	Analysis	Conversion	Synthesis	Other	**Total**
Benchmark	0.21	3.04	16.80	0.30	**20.4**
	↓ 350 %	↓ 12 %	↓ 6%		↓ 12 %
Proposed	0.74	0.37	1.05	0.30	**2.5**

- 理論上はモバイル端末でもリアルタイム変換が可能
  - Iphone6 single CPU: **2.8** GFLOPS (理論値)
  - Intel core i7-6850K single CPU: **14.4** GFLOPS (理論値)

処理時間の計測

- RTFによる評価
  - 1フレームごとの処理時間を算出し，波形の長さ (5 ms) で除算
  - 全フレームでのRTFの平均値を算出

	Analysis	Conversion	Synthesis	Other	Total
Benchmark	0.02	0.33	2.82	0.06	3.23
	↓ 800 %	↓ 42 %	↓ 8%		↓ 18 %
Proposed	0.16	0.14	0.22	0.06	0.58

1CPUでRTF < 1 を達成し，リアルタイム動作を確認

オンライン・オフラインの品質比較

- Online実装とOffline実装 [佐伯20]の品質を主観評価実験により比較
  - 各ケース30人の評価者
  - 話者性についてのXABテスト，音質についてのABテスト

Speaker similarity

Online	Score		Offline [佐伯20]
m2m	0.493	0.507	m2m
f2f	0.487	0.513	f2f

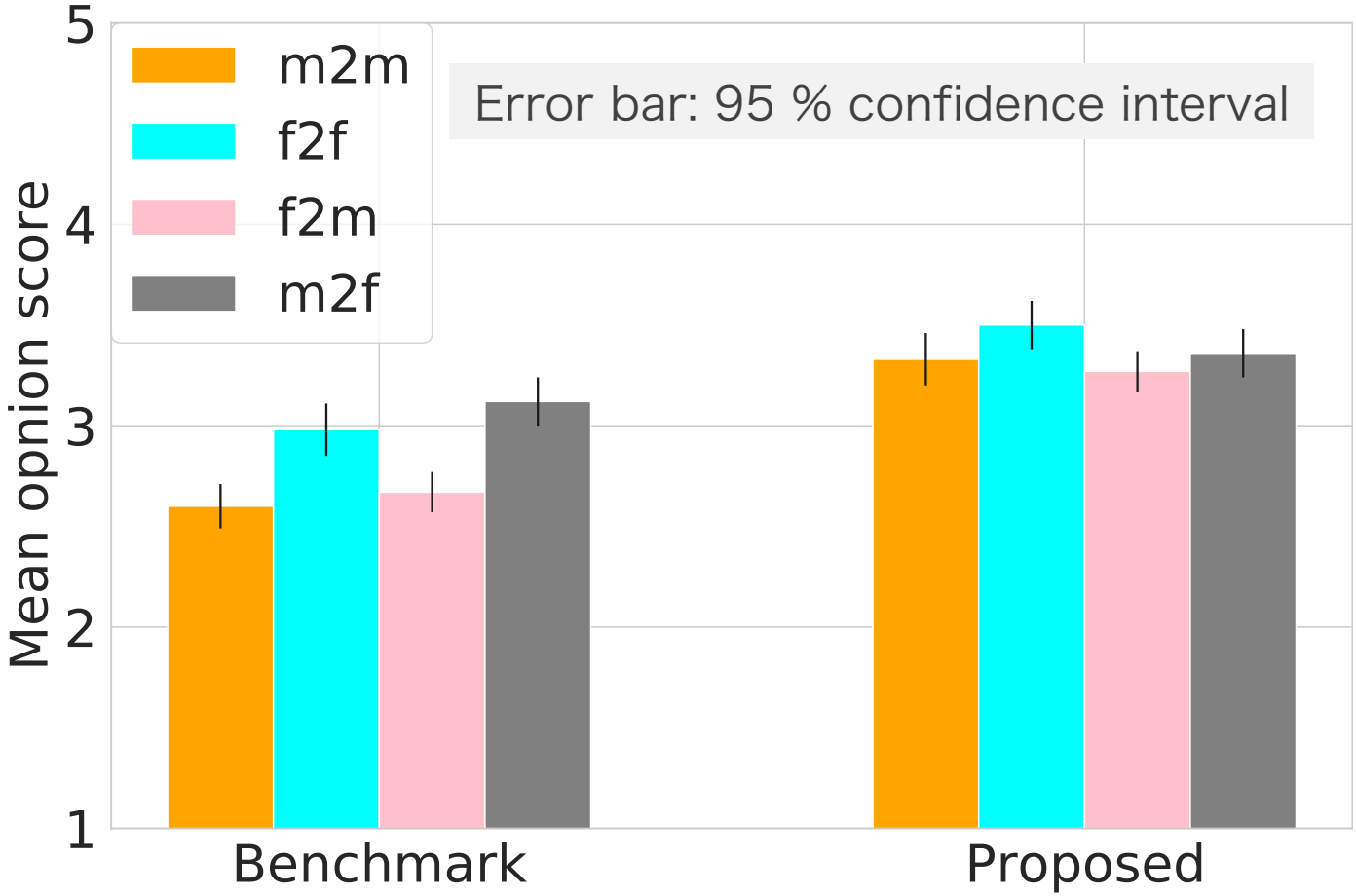
Speech quality

Online	Score		Offline [佐伯20]
m2m	0.483	0.517	m2m
f2f	0.510	0.490	f2f

オンライン変換 vs. オフライン変換に有意差なく，変換音声の品質は同等

変換音声品質の絶対評価

- **自然性**をmean opinion score (MOS) により評価
  - 同性間・異性間変換の各ケースにつき40人の聴取者が評価



提案するreal-time VCのMOSはbenchmarkより有意に高く，3.5程度

デモ

	Source	Target	Narrow band	Benchmark	Proposed
Intra-gender (f2f)					
Cross-gender (m2f)					

まとめ

- **研究目的**
  - 高音質なリアルタイム広帯域DNN声質変換の実現
- **手法**
  - サブバンド処理により低域 (0-8 kHz) のみをモデル化・変換 [佐伯20]
  - 低域変換の際にフィルタ打ち切りを行うことで計算量削減 [Saeki20]
  - **リアルタイム・オンライン処理のための実装 & F0変換機構**
- **評価結果**
  - 計算量の概算の結果，2.5 GFLOPS程度で広帯域音声を変換可能
  - RTF計測の結果，1CPUでのリアルタイム動作を確認
  - Online変換でも，offline変換 [佐伯20]と同等品質の変換音声を出力可能
  - そのまま帯域拡張したケース (Benchmark) より有意に高い自然性
- **今後の課題**
  - 実環境での頑健性の検討
  - さらなる変換音声品質の改善