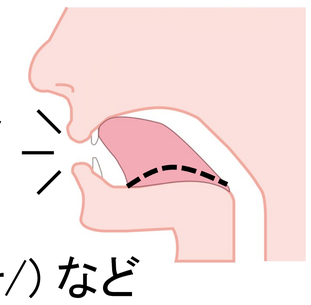
声質変換による舌亜全摘出者の音韻明瞭度改善のための補助情報の検討

村上 博紀（岡山大学大学院自然科学研究科 阿部研究室）

OKAYAMA UNIVERSITY

1. 研究背景・目的

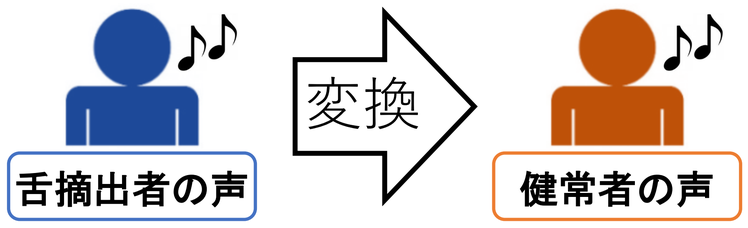
舌亜全摘出者



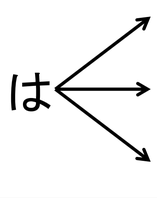
- 癌治療などにより，手術で舌を切除した人
- 舌が無いため聞き取りづらい声になる
- 舌を使って構音する代表的な音素
  - 摩擦音(/s/), 破裂音(/t/, /k/) , 流音(/r/) など

声質変換に基づく音韻明瞭度改善方式

- **声質変換**: ある話者の声を別の話者の声に変換する技術



一対多変換の問題



十分な音韻明瞭度改善は未だ実現できていない

⇒ 何らかの工夫が必要

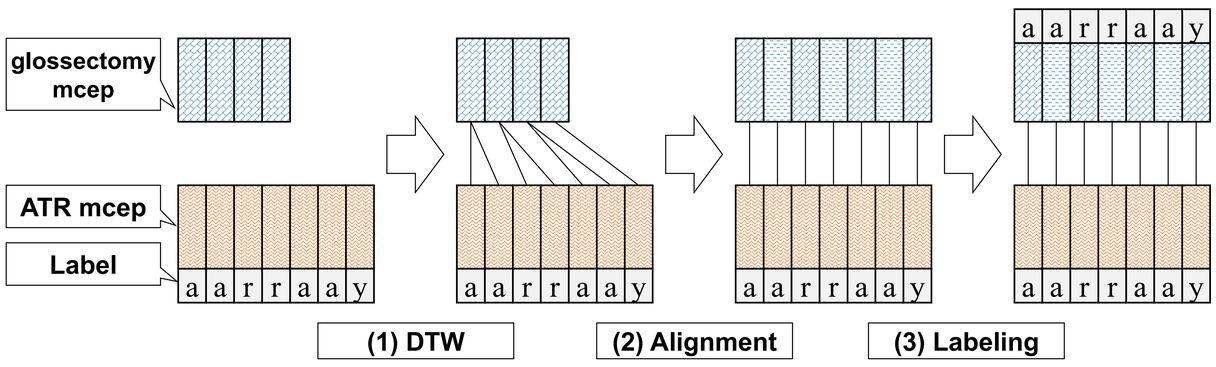
音声情報に加えて何らかの補助情報を用いることで解決
音声情報 + **音韻ラベル情報**

2. 音声データへの音韻ラベル付与

音韻ラベルの概要

- ATR音声DBに付属する47種類の音韻ラベルの音声記号層
- **3種類の情報**： 音韻の種類，開始時間，終了時間

フレーム単位での音韻ラベルの付与方式

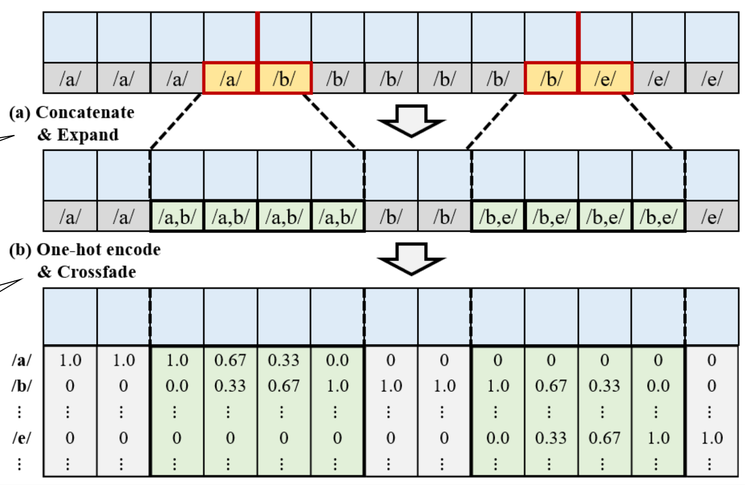


時間遷移を考慮した事後処理

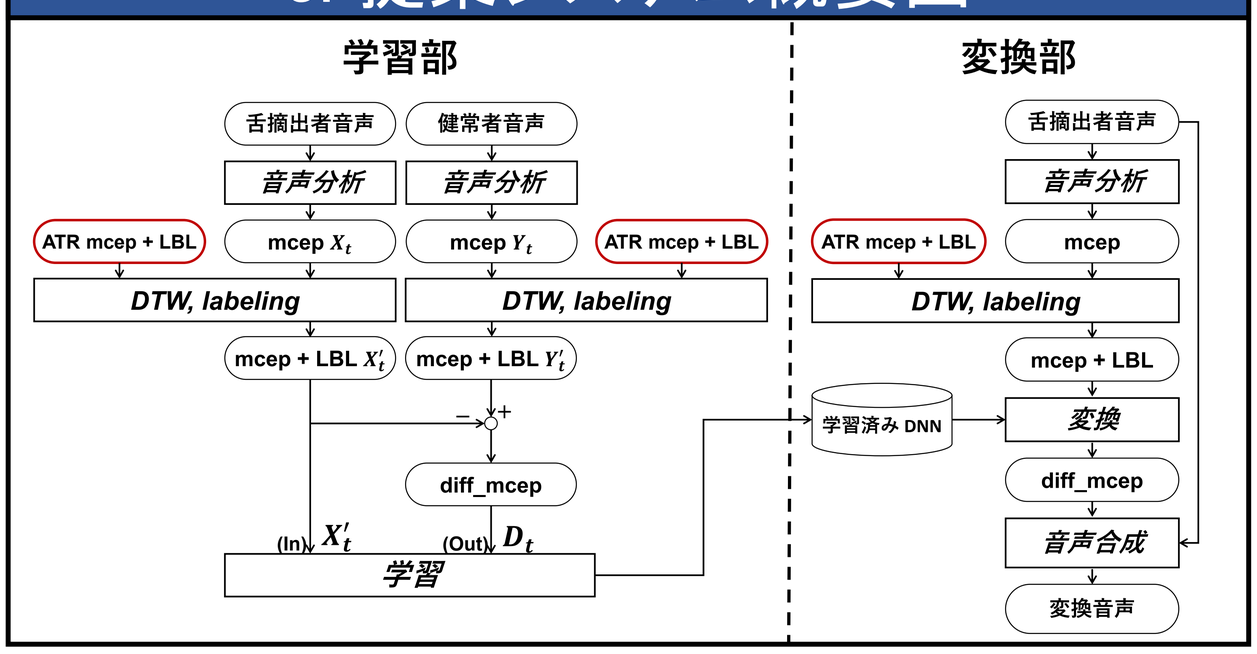
【目的】
フレーム境界での時間的遷移の不自然さを緩和する

(a) フレーム境界から前後Nフレームのラベルを**結合音韻ラベル**とする

(b) 音韻ラベルはone-hotエンコーディングをおこなう
結合音韻ラベルはクロスフェード処理



3. 提案システム概要図

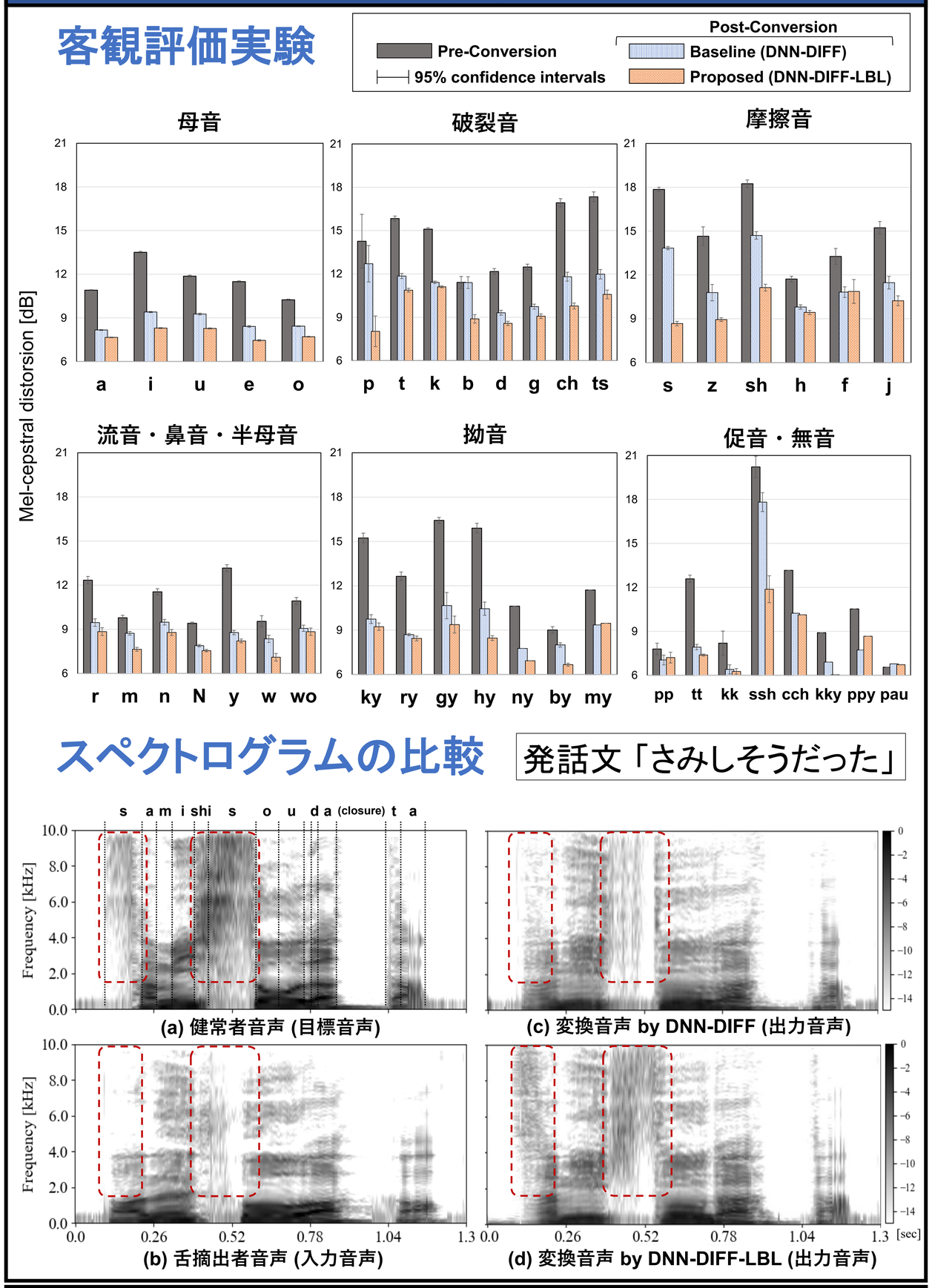


4. 実験条件

学習データ数	ATR音素バランス文 A～H セット 400 文
検証データ数	ATR音素バランス文 I セット 50 文
評価データ数	ATR音素バランス文 J セット 53 文
サンプリング周波数	20000 Hz
フレームシフト長	5 ms
音声分析	WORLD
入力特徴量	0～25次mcep + Δmcep + LBL (99次元)
出力特徴量	0～25次mcep + Δmcep (52次元)
NN モデル	全結合多層パーセプトロン (MLP)
中間層・ユニット数	3層 1024ユニット
ネットワークの形	[99,1024,1024,1024,52]
損失関数	平均二乗誤差 (MSE: Mean Squared Error)
活性化関数	中間: ReLU, 出力: 線形関数
最適化手法	Adam (学習率 α = 0.001)
特徴量の正規化	各次元ごとに平均0, 分散1

5. 評価実験

客観評価実験



スペクトログラムの比較

発話文「さみしそうだった」

6. まとめと今後の課題

まとめ

- 音韻ラベルを補助情報とした舌摘出者の音韻明瞭度改善のための声質変換システムを提案
- DTWによるフレーム同期によって音韻ラベルを付与
- 評価実験から提案手法の有効性が示された

今後の課題

- 他の情報から音韻系列を推定する手法を考案
- 音声からの書き起こしによる主観評価実験

舌亜全摘出者の音韻明瞭度改善のための推定音素ラベルを用いた声質変換の検討

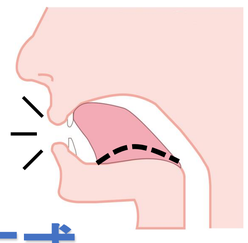
荻野 聖也（岡山大学大学院ヘルスシステム統合科学研究科 阿部研究室）

OKAYAMA UNIVERSITY

1. 研究背景・目的

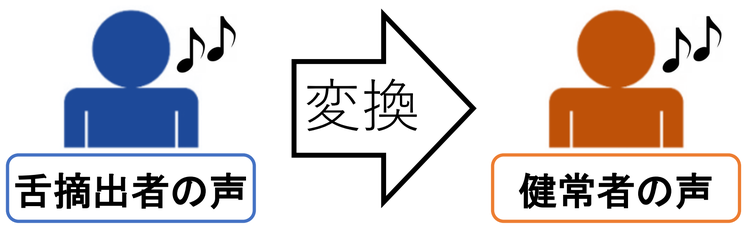
舌亜全摘出者

- 癌治療などにより，手術で舌を切除した人
- 舌が無いため聞き取りづらい声になる



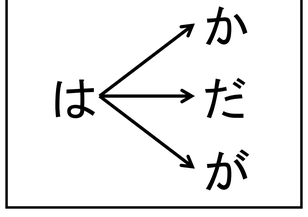
声質変換に基づく音韻明瞭性改善方式

- **声質変換**: ある話者の声を別の話者の声に変換する技術



「おはやまはいはふ」　「おかやまだいがく」

一対多変換の問題



過去の研究成果

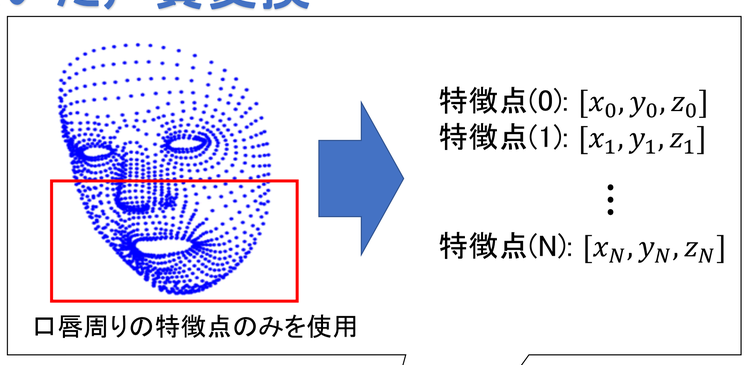
- **DNNに基づく声質変換**
  - ✓ 舌摘出者の音声を健常者の音声に変換
  - ✓ 一対多変換問題により，未だ改善が不十分
  - ✓ 音声以外の補助情報が必要
- **音素ラベルを用いた声質変換**
  - ✓ 摩擦音等の子音の音韻明瞭度が大きく改善
  - ✓ 発話に応じた音素ラベルの用意が必要

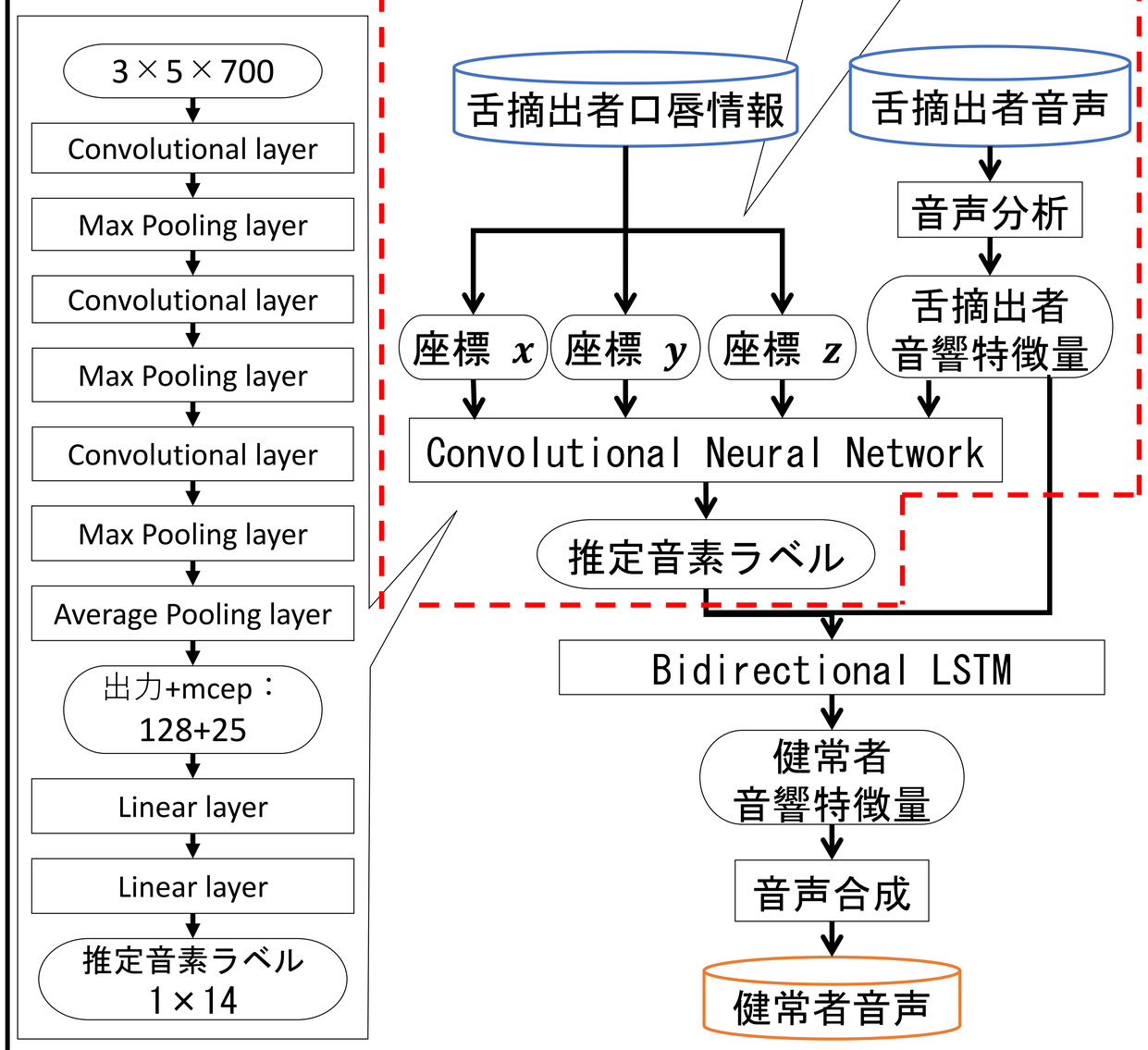
・音素ラベルの推定
・推定音素ラベルを用いた声質変換の検討

- 過去の研究から，音声情報と口唇画像情報を用いることで音素ラベルの推定精度の向上が期待できる

2. 提案方式

推定音素ラベルを用いた声質変換





3. 評価実験

実験条件

□声質変換方式

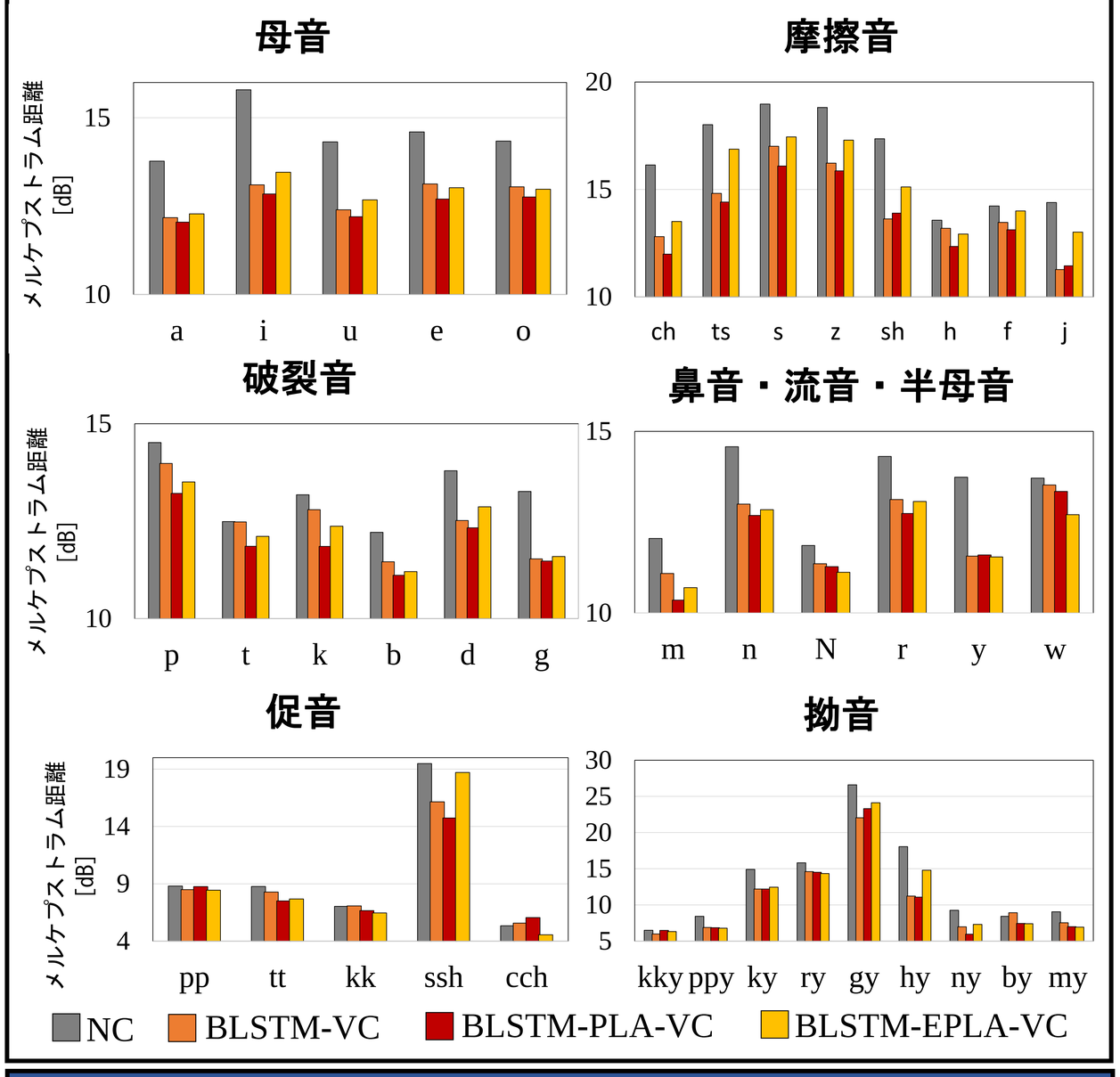
1. NC：変換なし
2. BLSTM-VC：音素ラベルを用いない方式
3. BLSTM-PLA-VC：事前に用意された音素ラベルを用いる方式
4. BLSTM-EPLA-VC：推定音素ラベルを用いる方式（提案方式）

	CNN	BLSTM
学習データ	ATR音素バランス A～Dセット 200文	ATR音素バランス A～Iセット 400文
評価データ	ATR音素バランス Jセット 53文	ATR音素バランス Jセット 53文
入力特徴量	メルケプストラム + 3次元顔座標	メルケプストラム + Δメルケプストラム ③(+ 音素ラベル) ④(+ 推定音素ラベル)
出力特徴量	音素ラベル	差分メルケプストラム
活性化関数	ReLU	PReLU
損失関数	Sigmoid cross entropy	平均二乗誤差
ミニバッチサイズ	20	1024
最適化手法	Adam	Adam

客観評価実験

□評価指標

メルケプストラム距離：入力音声と目標音声の距離を表す



4. まとめと今後の課題

まとめ

- 声質変換による舌摘出者の音韻明瞭度改善のための推定音素ラベルを用いた声質変換の検討
- 評価実験結果より，提案方式の有効性を示すことはできなかったが，今後の課題が明らかになった

今後の課題

- 音素ラベルの推定精度の向上
  - ✓ BLSTMを合わせたフレーム間での連続性を考慮し，推定
- 音素ラベルのクラスタリング方式の改善