舌亜全摘出者の音韻明瞭度改善のためのマルチモーダル声質変換方式の検討

Study of multimodal voice conversion method to improve speech intelligibility for glossectomy patients

荻野 聖也

Seiya Ogino

岡山大学 阿部研究室

Abe Laboratory, Okayama University

概要 本研究では舌亜全摘出者が発声する音声の音韻明瞭度改善を目的として，音響情報と口唇情報を用いたマルチモーダル声質変換方式を提案する．音響情報のみを用いたベースライン方式では，舌亜全摘出者の音韻明瞭性の改善は十分ではない．そこで，Microsoft Kinect v2 の顔検出機能により得られる特徴点を，口唇情報として加えたマルチモーダル声質変換で音韻明瞭性の更なる改善を目指す．本研究では主観評価実験により，提案方式の有効性を明らかにした．

1 はじめに

人は舌の動きや口の開閉により，構音を可能にしている．舌亜全摘出者は癌治療などのために手術で舌の半分以上を摘出した人であり，構音機能に大きな障害が残る．

声質変換とは，ある特定の話者が発声した音声を，発話内容を保持しつつ，別の話者が発声した音声に聞こえるように変換する技術である [1]．声質変換の代表的な統計的手法として，近年では NN（Neural Network）が用いられる．NN は非線形関数をモデル化することができ，声質変換においても入力話者と目標話者の非線形な特徴量の対応関係を精度良く表現可能である [2]．

口唇を動かすことによって構音がおこなわれ，読唇術で発声内容を推測できることから，口唇の動きに関する情報を用いて発話内容の推定が可能であると考えられる．よって，口唇情報は音声情報の補助情報として有効であると考えられる．

これまで，GMM（Gausian Mixture Mode）や DNN（Deep Neural Network）に基づいた声質変換を用いて舌亜全摘出者の音声を健常者の音声に変換することで，舌亜全摘出者の音韻明瞭度の改善を行ってきた．これにより破裂音等の音韻明瞭度の改善を報告した [3][4]．しかし，未だ十分な音韻明瞭度の改善は達成できていない．そこで，我々は声質変換においても，口唇の動き（口唇情報）を補助情報として用いることで従来よりも舌摘出者の音韻明瞭度が改善できると考えた．本研究では，口唇情報として Microsoft Kinect v2 の顔検出機能によって得られる 3 次元座標を用いた，マルチモーダル声質変換によって舌亜全摘出者の音韻明瞭度改善を目指す．主観評価実験では，提案方式による音韻明瞭性の改善を明らかにした．

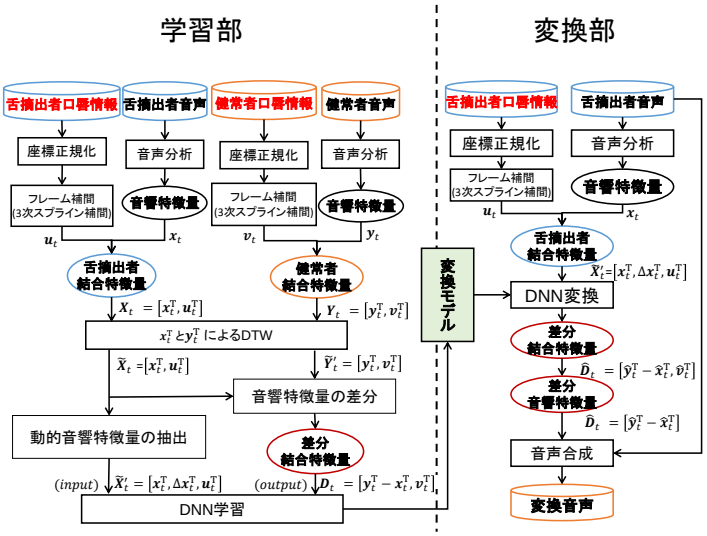

図 1: 提案方式 概要図

2 提案方式

本研究では音響特徴量と，Kinect より得られる 3 次元顔座標から求めた口唇特徴量を統合した結合特徴量を用いたマルチモーダル声質変換により，音響特徴量のみを用いるベースライン方式よりも音韻明瞭度を改善する．提案方式の概要図を図 1 に示す．

入力特徴量を舌亜全摘出者の結合特徴量，出力特徴量を口唇特徴量を健常者の差分結合特徴量とした DNN により変換モデルの学習をおこなう．まず，舌摘出者と健常者のパラレルコーパスに対してそれぞれ WORLD による音声分析をおこない，人の声道特性を表す音響特徴量を抽出する．あるフレーム t における舌摘出者の静的音響特徴量ベクトル $\boldsymbol{x}_t$，健常者の静的音響特徴量ベクトルを $\boldsymbol{y}_t$ とする．また，舌摘出者の口唇特徴量ベクトルを $\boldsymbol{u}_t$，健常者の口唇特徴量ベクトルを $\boldsymbol{v}_t$ とすると，舌摘出者の結合特徴量系列と健常者の結合特徴量系列はそれぞれ，$\boldsymbol{X}_t = [\boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{u}_t^{\mathrm{T}}]^{\mathrm{T}}$，$\boldsymbol{Y}_t = [\boldsymbol{y}_t^{\mathrm{T}}, \boldsymbol{v}_t^{\mathrm{T}}]^{\mathrm{T}}$ のように表せる．$^{\mathrm{T}}$ は転置を表す．動的時間伸縮（DTW: Dynamic Time Warping）を用いて，$\boldsymbol{X}_t$ と $\boldsymbol{Y}_t$ をフレームごとに対応付けをおこなう．そして，$\boldsymbol{x}_t$ から動的音響特徴量ベクトル $\Delta\boldsymbol{x}_t$ を算出し，舌摘出者の結合特徴量系列を $\tilde{\boldsymbol{X}}_{\boldsymbol{t}} = [\boldsymbol{x}_t^{\mathrm{T}}, \Delta\boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{u}_t^{\mathrm{T}}]^{\mathrm{T}}$，健常者の結合特徴量系列を $\tilde{\boldsymbol{Y}}_{\boldsymbol{t}} = [\boldsymbol{y}_t^{\mathrm{T}}, \boldsymbol{v}_t^{\mathrm{T}}]^{\mathrm{T}}$ とする．差分結合特徴量は $\boldsymbol{D}_t = [\boldsymbol{y}_t^{\mathrm{T}} - \boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{v}_t^{\mathrm{T}}]^{\mathrm{T}}$ で表される．最後に，$\tilde{\boldsymbol{X}}_t$ を入

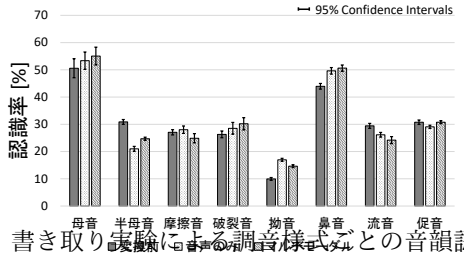

図 2: 書き取り実験による調音様式ごとの音韻認識率

力特徴量，$\boldsymbol{D}_t$ を出力特徴量とし，DNN による変換モデルを学習する．

変換部では舌摘出者の音声を学習済み変換モデルを用いて健常者の音声へ変換する．入力された任意の発話内容の舌摘出者音声に対して WORLD で音声分析をおこない，音響特徴量を抽出する．また，同じ発話内容の舌摘出者の口唇情報から口唇特徴量を得る．得られた音響特徴量と口唇特徴量から結合特徴量を生成し，これを変換モデルによって差分結合特徴量 $\widehat{\boldsymbol{D}}_{\boldsymbol{t}}$ に変換する．最後に，舌摘出者の入力音声と変換後差分結合特徴量の音響特徴量成分を MLSA フィルタにより合成し，舌摘出者の音声を健常者の音声へ変換する．

3　評価実験

評価実験の発話文には ATR 音素バランス 503 文 [6] の A，B，J サブセットを使用した．学習データは A，B セット 100 文，評価データは J セット 53 文とした．話者は男性話者 1 名の健常者音声と疑似舌摘出者音声を用いた．音声のサンプリング周波数は 20 kHz，分析のフレームシフトは 5 ms とした．音響特徴量として 0〜25 次元メルケプストラムとその Δ を用いた．DNN は中間層が 4 層で 256 ユニットの順伝播型ネットワークを使用した．活性化関数は ReLU，損失関数は平均二乗誤差，最適化手法は Adam を用いた．

3.1　主観評価実験

主観評価実験ではベースライン方式と提案方式の間の音韻明瞭度改善の差を書き取り実験と AB テストによって評価した．両実験の被験者は学生 9 名である．

3.1.1　書き取り実験

書き取り実験では 4 種類の評価音声をランダムな順番で一度のみ聞かせ，音声終了から 10 秒間の無音区間中に聞こえた通りにひらがなで書き取らせた．書き取られたフレーズから音韻ごとの認識率を算出し，評価指標とする．

調音様式ごとの実験結果を図 2 に示す．結果から提案方式により母音，半母音，破裂音の音韻明瞭度に改善がみられた．これは母音，半母音，/p/，/b/の破裂音の構音は，口唇の動きの影響が大きいためだと考えられる．また，他の調音様式において改善が見られなかった原因として，3 次元座標では舌の動きなどを捉えることが困難であったことが考えられる．

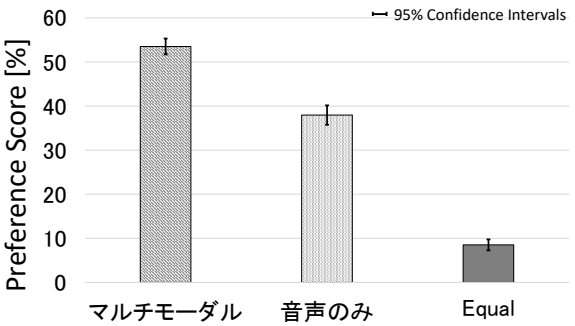

図 3: AB テストによるベースライン方式と提案方式の音韻明瞭性の比較

3.1.2　AB テスト

AB テストではベースライン方式と提案方式による変換音声の音韻明瞭性を評価した．評価音声をランダムな順序で聞き取り，どちらの音声が単語単位で，より明瞭に聞こえたかを評価させた．

実験結果を図 3 に示す．結果から，ベースライン方式と比べて提案方式により音韻明瞭性の改善がおこなわれたことが確認できる．さらに，書き取り実験の結果を踏まえると，母音の明瞭度改善がおこなわれたことにより，文全体の明瞭性が向上したと考えられる．これは文全体において母音が占める割合が大きいためであると考えられる．

4　まとめ

本研究では，舌摘出者の音声の音韻明瞭度を改善するために，音響情報と口唇情報を用いたマルチモーダル声質変換について提案した．主観評価実験により，変換音声の音韻明瞭度が向上していることが明らかとなった．しかし，提案方式では舌の動きにより構音がおこなわれる摩擦音や流音などの音素において，一部の改善が見られたが，改善度は未だ十分ではない．舌亜全摘出者の音韻明瞭度の改善にはこれらの音素の改善が不可欠であり，今後の課題である．

参考文献

[1] M. Abe *et al.*, in *Proc. ICASSP*, pp. 655–658, Apr. 1988.
[2] S. Desai *et al.*, in *Proc. ICASSP*, pp. 3893–3896, Apr. 2009.
[3] 田中慧他, 日本音響学会講演論文集, pp. 141–144, 2-5-8, Sep. 2016.
[4] 村上博紀他, 日本音響学会講演論文集, pp. 297–300, 2-Q-25, Sep. 2017.
[5] K. Kobayash *et al.*, in *Proc. INTERSPEECH*, pp. 2514–2518, Sep. 2014.
[6] A. Kurematsu *et al.*, Speech Communication, vol. 9, pp. 357–363, Sep. 1989.