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

Study of Auxiliary Information for Improving The Speech Ineligibility of Glossectomy Patients via Voice Conversion

村上 博紀
Hiroki Murakami

岡山大学 阿部研究室
Abe Laboratory, Okayama University

概要 本研究では，声質変換を用いた舌亜全摘出者の音韻明瞭度改善を目指す．舌摘出者の発話は曖昧なものであり，音響特徴量の変換だけでは十分な音韻明瞭度の改善は達成できない．本報告では，音響特徴量の補助情報として音韻ラベルを用い，声質変換における補助情報の有効性について検討した．評価実験から，従来手法よりも補助情報を用いた提案手法の方が高い音韻明瞭度の改善を実現できることを明らかにした．

1　はじめに

舌は人間が構音をおこなうための重要な器官である．舌摘出者は, 癌などの治療のために手術で舌を摘出した人であり, 構音機能に大きな障害が残る．特に，舌の半分以上を切除した患者のことを舌亜全摘出者と呼ぶ．舌亜全摘出者が発声する音声は健常者のものと比べると非常に聞き取りづらいものとなる．患者にとってこの問題は会話による円滑なコミュニケーションをおこなう上で深刻な障害となる．

声質変換は, 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, あたかも別の特定の話者が発声した音声に聞こえるように変換をおこなう技術である [1]．声質変換を用いた舌亜全摘出者の発話支援の研究の最初の試みとして，我々は GMM に基づく声質変換を用いた手法を提案した [2]．声質変換によって舌摘出者の声道特性を健常者のものに変換することで，聞き取りやすい音声に変換する．変換音声の一部に破裂音や摩擦音等の音素が復元できたことが報告されている．また，我々は差分スペクトル法 [3] による声質変換を提案し，舌摘出者の音声の自然性を改善した [4]．

しかしながら，声質変換による音韻明瞭度改善方式では, 未だ十分な音韻改善の効果は得られていない．この原因として，舌摘出者の曖昧な発音による多対一変換が起きていることが挙げられる．そこで我々は，声質変換において音響特徴量と共に補助情報を利用することで，多対一変換の問題を解決できるのではないかと考えた．そこで，我々は舌亜全摘出者の「音声情報」「口唇周りの動画像情報」「3 次元顔座標情報」を同時に収録したマルチモーダルデータベースを構築した [5]．また，我々はそのデータベースを用いた舌亜全摘出者のマルチモーダルな声質変換提案し，主に母音の発話明瞭度が向上を実現した [6]．一方で，摩擦音等の子音については未だ十分な成果が達成されていない．そこで，舌摘出者の声質変換に対して音響特徴量と共に補助情報を用いることの有効性について検討することにした．本稿では，補助情報として ATR 音素バランス文 [7] に付属している音韻ラベルを採用し，舌摘出者の声質変換に対して補助情報を用いることの有効性を検証した．

2　音韻ラベルの概要とラベル付与方式

本研究では，音韻ラベルとして ATR 音素バランス文に付属する計 47 種類の音韻ラベルの音声記号層を使用する．ATR 音素バランス文には，音声と音韻ラベル情報がセットで収録されている．音韻ラベル情報には, 音声に対応する区間の [音韻の開始時間][音韻の種類][音韻の終了時間] が収録されている．音韻ラベルの継続時間情報を基に，ATR 音声に対して音声分析をおこない抽出した音響特徴量の各フレームに対してラベル情報を付与する．本稿では，ATR 音素バランス文から 1 名の男性健常者話者を選択し，音響特徴量とラベル情報を結合した音響ラベル系列を作成した．次に，舌摘出者音声に対して音韻ラベルを付与する．まず，舌摘出者の音響特徴量系列と ATR 話者の音響ラベル系列に対して動的時間伸縮 (DTW:Dynamic Time Warping) をおこない，フレームの時間的対応付けをおこなう．対応するフレームに対してラベル情報を付与することで，舌摘出者の音響特徴量に音韻ラベルを付与する．

3　音韻ラベルを補助情報とした声質変換

音韻ラベルを補助情報とした声質変換システムは学習部と変換部の 2 つに分けられる．学習部では，まず入力話者と目標話者の音声それぞれに対して音声分析をおこない, 音響特徴量を抽出する．入力話者の音響特徴量系列ベクトルを $\boldsymbol{X}_t = [\boldsymbol{x}_t^{\mathrm{T}}, \Delta\boldsymbol{x}_t^{\mathrm{T}}]^{\mathrm{T}}$，目標話者の音響特徴量系列ベクトルを $\boldsymbol{Y}_t = [\boldsymbol{y}_t^{\mathrm{T}}, \Delta\boldsymbol{y}_t^{\mathrm{T}}]^{\mathrm{T}}$ とする．$^{\mathrm{T}}$ は転置を表す．Δ は動的特徴量を表す．また，ATR 話者の音響特徴量系列 $\boldsymbol{z}_t$ と音韻ラベル系列 $\boldsymbol{l}_t$ を結合した音響ラベル系列ベクトルを $\boldsymbol{Z}_t = [\boldsymbol{z}_t^{\mathrm{T}}, \boldsymbol{l}_t^{\mathrm{T}}]^{\mathrm{T}}$ とする．2 章と同様に $\boldsymbol{Z}_t$ と音響特徴量系列 $\boldsymbol{X}_t$, $\boldsymbol{Y}_t$ のそれぞれに対して DTW をおこない，音韻ラベルの付与をおこなう．DTW によって，入力話者の音響ラベル系列ベクトル $\boldsymbol{X}'_t = [\boldsymbol{x}_t^{\mathrm{T}}, \Delta\boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{l}_t^{\mathrm{T}}]^{\mathrm{T}}$，目標話者の音響ラベル

系列ベクトル $\boldsymbol{Y}'_t = [\boldsymbol{y}_t^{\mathrm{T}}, \Delta\boldsymbol{y}_t^{\mathrm{T}}, \boldsymbol{l}_t^{\mathrm{T}}]^{\mathrm{T}}$ を得る．そして，差分音響特徴量系列ベクトルとして $\boldsymbol{D}_t = \boldsymbol{Y}'_t - \boldsymbol{X}'_t$ を生成する．この時，$\boldsymbol{l}_t$ は取り除いて計算される．最後に，入力特徴量を $\boldsymbol{X}'_t$，出力特徴量を $\boldsymbol{D}_t$ としてこれらをマッピングする関数を DNN で学習する．

変換部では，まず入力話者の音声に対して音声分析をおこない，入力話者の音響特徴量系列ベクトル $\boldsymbol{X}_t$ を得る．次に，DTW によって $\boldsymbol{X}_t$ に対して音韻ラベルを付与した $\boldsymbol{X}'_t$ を得る．次に，学習済み DNN モデルによって $\boldsymbol{X}'_t$ から推定差分音響特徴量系列ベクトル $\hat{\boldsymbol{D}}_t$ を得る．最後に，入力音声波形に対して $\hat{\boldsymbol{D}}_t$ をメル対数スペクトル近似 (MLSA) フィルタ [8] により直接畳み込むことで音声を変換する．以上のようにして，入力話者の音声を目標話者の音声に変換する．

4　評価実験

4.1　実験条件

評価実験の発話文には ATR 音素バランス 503 文 [7] を使用した．学習データは 400 文，検証データは 50 文，評価データは 53 文とした．舌摘出者の音声については，健常者の舌を固定する器具を作成し，疑似的に舌摘出者の発声を再現することで収録をおこなった．話者は男性話者 1 名 (Male1) を用いた．ここで，Male1 の疑似舌摘出者は SPM1(Simulated Patient Male1) と表記する．評価実験では，ベースライン手法として DNN を用いた差分スペクトル補正に基づく声質変換 (**DNN-DIFF**) [4] ，提案手法として音韻ラベルを補助情報とした DNN-DIFF の声質変換 (**DNN-DIFF-LBL**) の 2 種類の変換手法を比較した．

音声のサンプリング周波数は 20 kHz である．音響特徴量は WORLD [9] 音声分析によって抽出されたスペクトル包絡から近似した 0〜25 次メルケプストラムを用いた．フレームシフト長は 5 ms とした．音声合成フィルタには MLSA フィルタを用いた．音韻ラベルは該当する要素を 1，それ以外を 0 とする 1-of-k の 47 次元ベクトルとした．

DNN は各層のユニット数が [99,1024,1024,1024,52] の全結合順伝播型ネットワークを用いた．活性化関数は ReLU を用いた．損失関数は平均二乗誤差とした．最適化手法は Adam を用いた．学習時，特徴量は各次元ごとに平均 0，分散 1 となるように正規化した．

4.2　スペクトログラムの比較

音韻明瞭度改善の効果はスペクトログラムから観察することができる．図 1 に SPM1 to M1 の声質変換について，DNN-DIFF と DNN-DIFF-LBL の変換音声のスペクトログラムをを比較した結果を示す．スペクトログラムは「さみしそうだった」という発話の音声から分析されたものである．赤枠のドットで囲まれた部分について注目する．健常者音声 (a) と疑似舌摘出者音声 (b) を比較すると，(b) には該当部分の周波数成分

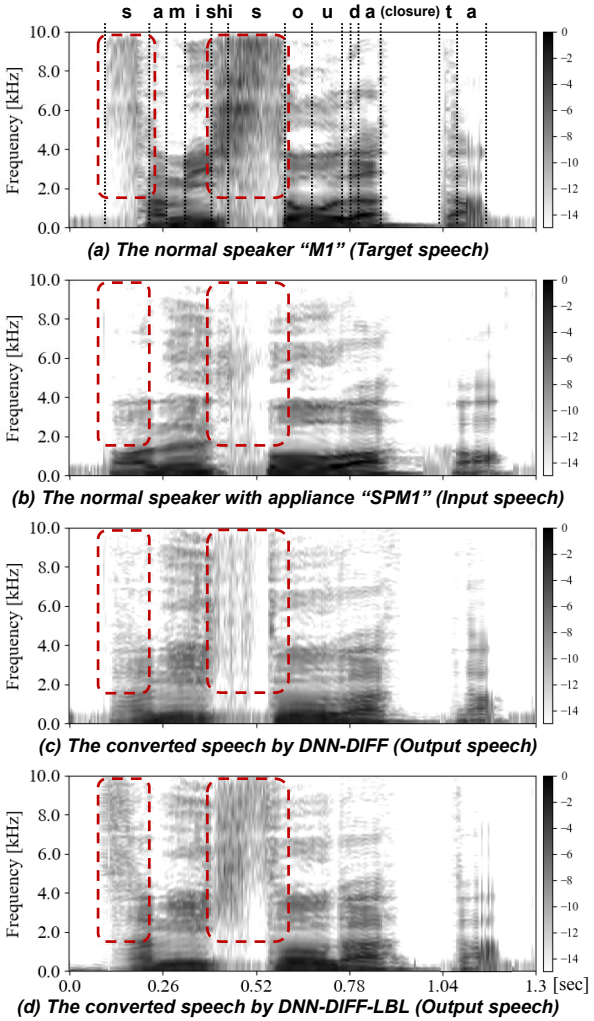

図 1: SPM1 to M1 のスペクトログラムの比較

が失われていることが分かる．(b) とベースライン手法 (c) を比較すると，該当部分にわずかに周波数成分が再構築されていることが分かる．(b) と提案手法 (d) を比較すると，該当部分に強い周波数成分が再構築されていることが分かる．(d) には (c) よりも強い周波数成分が再構築されており，目標である健常者音声 (a) に近いスペクトルとなっていることが分かる．これは，音韻ラベルによる補助情報を用いることで多対一変換の問題が解消され，正しい変換が可能となったためであると考えられる．

5　まとめ

本稿では舌亜全摘出者の音韻明瞭度改善のための補助情報を用いた声質変換について検討した．評価実験から，舌亜全摘出者の声質変換において補助情報を用いることの有効性が明らかとなった．今後は主観評価実験をおこない，音韻明瞭度改善の効果を検証する．また，現状ではシミュレーションによる音韻ラベル付与をおこなっており，実用化が難しい．従って，何らかの外部情報から音韻ラベルを推定する方式を検討する必要がある．

参考文献

[1] M. Abe *et al.*, “Voice conversion through vector quantization,” Proc. *ICASSP*, pp. 655–658, 1988.

[2] K. Tanaka *et al.*, “Speaker Dependent Approach for Enhancing a Glossectomy Patient’s Speech via GMM-based Voice Conversion,” Proc. *INTERSPEECH*, pp. 3384–3388, 2017.

[3] K. Kobayashi *et al.*, “Statistical singing voice conversion with direct waveform modification based on the spectrum differential,” Proc. *INTERSPEECH*, pp. 2514–2518, 2014.

[4] 村上 他, “DNN に基づく差分スペクトル補正を用いた声質変換による舌亜全摘出者の音韻明瞭性改善の検討,” 音講論 (秋) , 2–Q–25, pp. 297–300, 2017.

[5] 村上 他, “舌亜全摘出者の音韻明瞭性改善のためのマルチモーダルデータベースの構築,” 音講論 (秋) , 2–Q–32, pp. 355–358, 2018.

[6] 荻野 他, “音声と口唇形状を用いた声質変換による舌亜全摘出者の音韻明瞭度改善の検討,” 信学技報, SP2018–3, vol. IEICE–118, no. 112, pp. 7–12, 2018.

[7] A. Kurematsu *et al.*, “ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis,” Speech Communication, vol. 9, pp. 357–363, 1990.

[8] 今井 他, 信学論 (A), Vol. J66–A, No. 2, pp. 122–129, 1983.

[9] M. Morise *et al.*, “World: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.