GMM に基づく声質変換を用いた舌亜全摘出者の音韻明瞭性改善の検討

Study of phinetic clarity improvement for a glossectomy patient via GMM-based voice conversion

田中 慧

Kei Tanaka

岡山大学 阿部研究室

Abe Laboratory, Okayama University

概要 本研究では特徴量を変換し話者性を変化させる GMM に基づく声質変換技術を舌摘出者の損なわれたスペクトルの復元を行うために応用し，舌亜全摘出者の音韻明瞭度の改善を目指す．本報告では研究における最初の試みとして舌摘出者のために最適な学習条件を検討するとともに，健常者同士の変換音声と舌摘出者，健常者間の変換音声を MFC（Mel-Frequency Cepstrum）距離を用いて比較することで本手法の課題と有効性を明らかにした．

1 はじめに

音声は人がコミュニケーションをとるための主要な手段であり，日々の生活の質を維持するために重要な役割を果たしている．これは発声に障碍を持つ人々も同様で，本研究では特に舌切除，あるいは舌の運動障害を持つ患者の発声を改善するための研究を行っている．こうした構音障碍への現在の治療法として舌接触補助床や人工舌 [1] が挙げられる．しかしこれらの治療法は器具を直接口内に取り付けるため食事中や口内が荒れている状態では使えないなどの欠点を持っている．そこで本報告では声質変換アルゴリズムを使用した新たな音韻明瞭度改善方法を提案する．

声質変換は話者の声の個人性を変化させる技術である．話者 A が発した音声を別の話者 B が発したように聞こえる音声へと変換を行う．声質変換技術は食道発声の音声支援など様々な用途に応用されている．本研究においては A を舌摘出者，B を健常者として舌摘出者の損なわれた特徴量を健常者に近づけるというアプローチを取っている．この際，健常者間との比較で問題の難しさを計ることを容易にするため声質変換には従来アルゴリズムである GMM（Gaussian mixture model）に基づく声質変換アルゴリズム [2] を採用している．

本報告では研究を始めるにあたっての基礎的な実験として舌摘出者の声質変換における最適な学習条件を検討した．また舌摘出者から健常者への変換の効果を評価するために，一般的な研究対象である健常者から他の健常者への変換結果と比較した．

2 GMM に基づく声質変換アルゴリズム

$\mathbf{x}_t$ と $\mathbf{y}_t$ をそれぞれ D 次元の元話者と目的話者の特徴量ベクトルとする．元話者と目的話者の同時確率密度は GMM によって次のようにモデル化される．

$$P(\mathbf{z}_t|;\boldsymbol{\lambda}^{(x)}) = \sum_{m=1}^{M} w_m N((\mathbf{z}_t;\boldsymbol{\mu}_m^{(x)},\boldsymbol{\Sigma}_m^{(x)})x \qquad (1)$$

ここで $\mathbf{z}_t$ は結合ベクトル $[\mathbf{x}_t^T,\mathbf{y}_t^T]$ であり，T はベクトルの転置を意味する．m は混合成分指標，M は総混合数であり，w_m は m 次の混合成分の重みを示す．また平均 $\boldsymbol{\mu}$，分散 $\boldsymbol{\Sigma}$ の正規分布は $N(*;\boldsymbol{\mu},\boldsymbol{\Sigma})$ と表記されている．GMM のパラメータセットは $\boldsymbol{\lambda}$ であり，重み，平均ベクトル，個々の混合成分の共分散行列で構成されている．結合ベクトル $\mathbf{z}_t$ は元話者と目的話者が同じ文を発声したパラレルコーパスを使用して DTW（dynamic time warping）によって生成される．最終的に，N はコーパスから与えられた学習データの総フレーム数となる．

M 次の混合成分の平均ベクトル $\boldsymbol{\mu}_m^{(z)}$ と共分散行列 $\boldsymbol{\Sigma}_m^{(z)}$ は以下のようにかける．

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix} \qquad (2)$$

ここで $\boldsymbol{\mu}_m^{(x)}$ と $\boldsymbol{\mu}_m^{(y)}$ はそれぞれ元話者と目的話者の m 次の平均ベクトルである．行列 $\boldsymbol{\Sigma}_m^{(xx)}$ と $\boldsymbol{\Sigma}_m^{(yy)}$ はそれぞれ元話者と目的話者の m 次混合成分の共分散行列である．GMM では学習データセットにおいて，DTW で自動的に揃えられた結合ベクトルに EM（expectation-maximization）アルゴリズムを用いて学習を行う．

3 評価実験

3.1 実験条件

本実験を行うにあたっての音声データの条件を表 1 に示す．また，3.2 では舌摘出者と健常者間，男性健常者同士の変換を行っており，3.3 ではそれらに加え男女健常者間の変換を行っている．また 3.3 ではフレーズ毎に発声した音声で学習した場合と文毎に発声した音声で学習した場合のそれぞれの音声の差を比較している．

3.2 舌摘出者向けの学習条件の検討

まず舌摘出者の変換を行う際の学習量の影響と適切な混合数を検討した．ここではフレーズ毎の発声音声

表 1: 実験条件

サンプリング周波数	20 kHz
音声分析手法	STRAIGHT[3]
フレームシフト	5 ms
音声特徴量	0–24 次 MFC およびその Δ

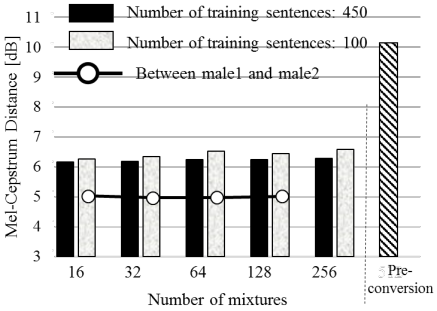

図 1: 混合数，学習数による性能の評価

を学習データとして用い，文毎に発声した音声データを評価に用いた．評価指標には MFC 距離を用いて変換後の特徴量と目的話者がどの程度近づいたか測定した．実験結果を図 1 に示す．変換前の MFC 距離と比較して変換後は 40% 減少したことがわかる．しかし健常者同士の変換に比べると 28% 大きい値となった．学習数では 450 文のほうがわずかに良い性能を示したが，収録の負担の観点から舌摘出者に対しては 100 文で学習するほうがより好ましいと考えられる．さらに最も良い性能の得られる混合数は舌摘出者，健常者間で 16，健常者同士で 32 となることがわかった．

3.3 舌摘出者への提案手法の有効性の検討

発話者と発話方式についても 3.2 と同様に MFC 距離を用いて評価した．この実験では 3.2 の結果から学習数 100，混合数は舌摘出者 16，健常者 32 で実験した．評価する際は 10 分割交差確認を用いて評価した．図 2 に実験結果を示す．変換前との MFC 距離から判断すると舌摘出者の差は健常者同士のものよりはるかに大きく，これは舌摘出者の声道形状が健常者のものと実質的に異なることを示している．MFC 距離は全ての組で 40% の減少が見られたが舌摘出者の変換結果は健常者に比べより距離が大きくなる結果となった．ここで考えられるのは舌摘出者の変換の際には多対一もしくは一対多対応になることで誤変換を招いている可能性である．発話方式の差による結果は話者の差に比べ小さいものであり，患者への負担の少ないフレーズ発話が良いと考えられる．

図 3 は舌摘出者音声（入力音声），変換音声（出力音声），健常者音声（目的話者音声）のスペクトログラムを示している．図 3 で舌摘出者音声と変換音声を比較すると，X で示した範囲においてパワースペクトルが新たに生成されていることがわかる．これは破裂音と呼ばれる舌を使うことではっきりと発音される音素で

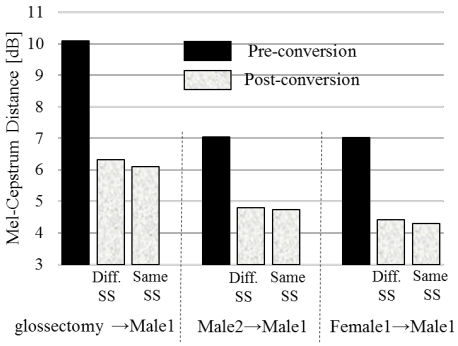

図 2: 話者による変換精度の評価

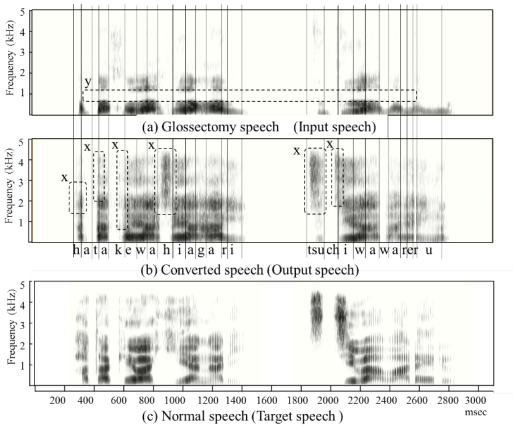

図 3: 各スペクトログラムの比較

あり，目的話者と比較しても適切に再構築されていることがわかる．また Y で示した 1 kHz 帯域でもスペクトルを再構築していることがわかる．

4 まとめ

舌摘出患者の音声に GMM に基づく声質変換アルゴリズムを適用して音韻明瞭度を改善する手法を提案した．実験結果から MFC 距離が 40% 減少することを確認したが，健常者に比べ 28% 高い結果となった．また破裂音の音素での高周波スペクトルが再構築されていることが確認できた．今後の予定として明瞭度の改善を確認する聞き取り評価を実施するとともに，セグメントでの変換を行うことで一対多のような誤った対応付けを改善することを目指す．

参考文献

[1] K. Kozaki, *et al.*, “Structure of a new palatal plate and the artificial tongue for articulation disorder in a patient with subtotal glossectomy,” Acta Medica Okayama. (in printing)

[2] T. Toda, *et al.*, “Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory,” IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, No. 8, pp. 2222–2235, Nov. 2007.

[3] H. Kawahara, *et al.*, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequencybased F0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication 27，pp. 187–207，1999.