構音障害者を対象とした
混合正規分布モデルに基づく統計的声質変換に関する研究*

☆石井良，滝口哲也，有木康雄（神戸大）

1　はじめに

近年，健常者以外に対する声質変換の研究が始められている．例えば，文献[1]では喉頭摘出者を対象とした声質変換を行っている．本稿では，脳性麻痺によって構音機能に障害を持つ被験者の声を対象として声質変換することを目的としている．脳性麻痺は未成熟脳における非進行性の欠損，もしくは病変に起因する運動，および姿勢の疾患として定義されている．特に，意図的な動作を行う際や緊張状態にあるときに，筋肉の制御が難しくなり，不随意運動を伴って正しく構音出来ない場合がある．

また，声質変換には高品質な音声を合成するための音声分析合成方式が不可欠である．高品質な音声を合成する音声分析変換合成方式として，STRAIGHT[2]が挙げられる．

今回の報告では，STRAIGHT ベースの混合正規分布に基づく統計的声質変換に関して，構音障害者を対象に実験を行った結果について述べる．

2　混合正規分布に基づく声質変換

2.1 学習ステップ

入力特徴量，出力特徴量を

$$\mathbf{X}_t=[x_1,...,x_d]^T,\quad \mathbf{Y}_t=[y_1,...,y_d]^T$$

とする．ここで，d は次元数で，T は転置を表す．健常者と構音障害者の同一発話の特徴量を，動的計画法によりタイムアラインメントをとり，GMM を学習する．結合確率密度は次式にてモデル化される．

$$p(\mathbf{X}_t,\mathbf{Y}_t,|\boldsymbol{\lambda})=\sum_{m=1}^{M}\alpha_m N(\mathbf{X}_t,\mathbf{Y}_t;\boldsymbol{\mu}_m^{(\mathbf{X},\mathbf{Y})},\sum_m^{(\mathbf{X},\mathbf{Y})})$$

$$\boldsymbol{\mu}_m^{(\mathbf{X},\mathbf{Y})}=\begin{bmatrix}\boldsymbol{\mu}_m^{(\mathbf{X})}\\ \boldsymbol{\mu}_m^{(\mathbf{Y})}\end{bmatrix}\qquad \sum_m^{(\mathbf{X},\mathbf{Y})}=\begin{bmatrix}\sum_m^{(\mathbf{XX})} & \sum_m^{(\mathbf{XY})}\\ \sum_m^{(\mathbf{YX})} & \sum_m^{(\mathbf{YY})}\end{bmatrix}$$

ここで，$N(\mathbf{X};\boldsymbol{\mu},\boldsymbol{\Sigma})$は平均ベクトル$\boldsymbol{\mu}$，共分散行列$\boldsymbol{\Sigma}$の正規分布を表し，$\alpha_m$は m 番目の分布重みで，$M$ は混合数である．また GMM のパラメータは EM アルゴリズムで推定される．

2.2 変換ステップ

構音障害者の音声特徴量を健常者の音声特徴量へ変換する関数は，次式で表すことができる[3,4].

$$\hat{\mathbf{y}}_t=\sum_{m=1}^{M}h_m(\mathbf{x}_t)[\boldsymbol{\mu}_m^{(\mathbf{Y})}+\sum_m^{(\mathbf{YX})}(\sum_m^{(\mathbf{XX})})^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_m^{(\mathbf{X})})]$$

$$h_m(\mathbf{x}_t)=\frac{\alpha_m N(\mathbf{x}_t;\boldsymbol{\mu}_m^{(\mathbf{X})},\sum_m^{(\mathbf{XX})})}{\sum_{i=1}^{M}\alpha_i N(\mathbf{x}_t;\boldsymbol{\mu}_i^{(\mathbf{X})},\sum_i^{(\mathbf{XX})})}$$

$\mu_m^{(x)}$，$\mu_m^{(Y)}$，はそれぞれ構音障害者，健常者のクラス m における平均ベクトル，$\boldsymbol{\Sigma}_{\mathbf{m}}^{(\mathbf{XX})}$，$\boldsymbol{\Sigma}_{\mathbf{m}}^{(\mathbf{YX})}$はそれぞれ，構音障害者のクラス m における共分散行列，健常者と構音障害者のクラス m の相互共分散行列を表す．

3　実験

3.1 実験条件

音声分析変換合成 STRAIGHT によって抽出された平滑化スペクトル(標本化周波数 12000 Hz)は，離散コサイン変換することで高次成分がほとんど 0 となるため，本実験では，その低次 24 次元分を GMM の学習，テストに用いる．学習時は，抽出された特徴量を，健常者と構音障害者の 105 単語から成る同一発話データセットに関して，ユークリッド距離に基づく DP マッチングを行い，時間アラインメントをとる．またテスト時は，学習に用いていない 10 単語の音声データを用いて声質変換を行った．

3.2 聴取実験

成人男女５名による聴取実験を行った．評価段階は(1.非常に聴き取りづらい，2.聴き取りづらい，3.聴き取ることができる，4.聴き

* Statistical voice conversion based on GMM for articulation disorders, by Ryo Ishii, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

取りやすい，5.非常に聴き取りやすい)で，静かな部屋でのヘッドホンによる両耳受聴を行った．

3.3 実験結果

Fig. 1 は聴取実験の結果を表したものである．混合数が 16 の場合と 32 の場合で比較すると，わずかであるが，変換音声の評価値が上昇しているのが見て取れる．

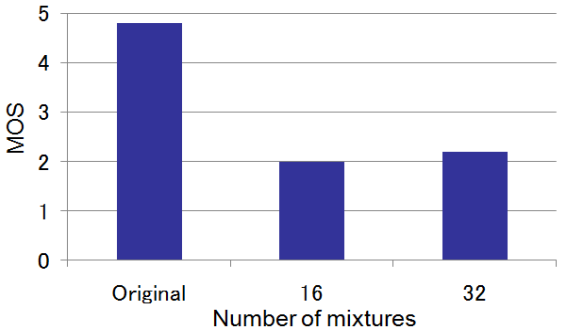

Fig. 1 Result of opinion test

また，音声のスペクトログラムに関して，変換前の構音障害者の音声では，健常者の音声では現れている高周波数域のパワーがあまり現れないが，変換後の音声においては，それが復元された．

結果として，Fig. 1 からわかるように，変換された音声の明瞭度は高くない．そもそも構音障害者の音声というものは，冒頭でも述べたが，体の緊張などと相まって，本来健常者なら現れるはずの子音部分が存在しないことが多々ある．このことから，そもそもの音声特徴量間のデータのアラインメントをとる際に問題が生じているのではないかということが考慮される．学習に用いた構音障害者の音声に現れていない子音部と，健常者の正しい発話音声で音素ごとに時間アラインメントをとるのは困難である．

Fig. 2, Fig. 3 は，それぞれ単語発話「あかちゃん」，「こんにゃく」に関して，健常者の音声特徴量(a)と構音障害者の音声特徴量(b)を時間アラインメントしたものである．(c)がアラインメント後の健常者の音声特徴量，(d)がアラインメント後の構音障害者の音声特徴量である．Fig. 2 に関しては正確にアラインメントがとれていることがわかるが，Fig. 3 に関しては，両発話間の音素対応関係に齟齬がみられる．

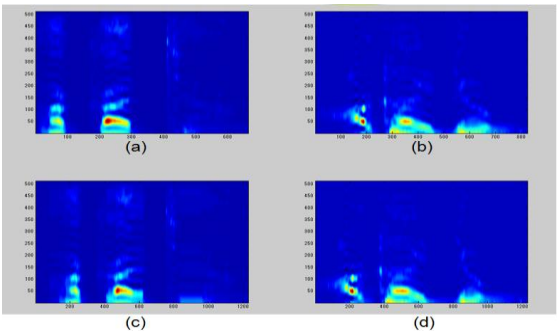

Fig. 2 Result of time alignment 1

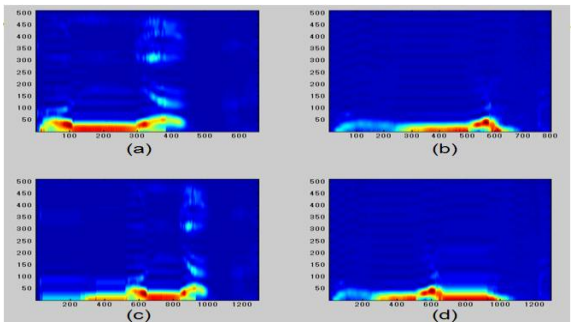

Fig. 3 Result of time alignment 2

4　おわりに

実験結果として，統計的声質変換を行うことで，ある程度の音質の改良を得ることができた．しかし，変換された音声は，構音障害者の方のコミュニケーションを円滑化するにはまだ十分ではなかった．

今回の結果から考えられる課題として，構音障害者と健常者の音声の正しい音素対応関係の学習の重要性が挙げられる．構音障害者の音声において，埋もれてしまっている子音部分をどのように補っていくのかを今後検討していきたい．

参考文献

[1] 土井啓成　他, “1 対多固有声変換に基づく無喉頭音声の音質及び話者性の改善” 情報処理研報, 2010-SLP-82, No. 1, pp. 1-6, 2010.

[2] 河原英紀　他, “時間周波数領域での補間を用いた音声の変換について” 信学技報, EA96-28, pp. 9-16, 1996.

[3] Y. Stylianou et al., “Statistical methods for voice quality transformation,” Proc. EUROSPEECH, pp. 447-450, 1995.

[4] Y. Stylianou et al., “A system voice conversion based on probabilistic classification and a harmonic plus noise model,” Proc. ICASSP, pp. 281-284, 1998.