

構音障害者を対象とした SSMを用いた音声認識の検討*

☆石井良, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

音声認識技術は, これまでの多くは成人を対象としたものであったが, 現在では子供や高齢者などの発話スタイルが成人と異なる人を対象とした場合や[1], また車内や会議室といった様々な環境下や場面を対象とした場合など多様化し, 使用される機会が増加している. 日本には現在 350 万人を超える身体障害者がいること, また近年の高齢化社会の問題から, 福祉分野における情報技術の発展が重要である[2]. しかし, 構音に障害のある人など発話スタイルが健常者とは異なる言語障害者を対象としたものは非常に少ない.

言語障害の原因の一つとして, 脳性麻痺がある. 本研究ではアテトーゼ型の脳性麻痺の患者を対象に研究を行っている. アテトーゼとは患者が緊張状態にあるときや, 意図的行動を行う際に現れる.

発話が困難で手足が不自由な場合など, 音声に頼るほかない場合が考えられ, 構音障害者を対象とした音声認識の実現が期待されており, 障害者の就業機会の増加や講演時の補助等への活用などが望まれる. しかし, 構音障害者における発話スタイルは健常者とは大きく異なるため, 従来用いられている不特定話者モデルでは認識が困難であるのが現状である.

2 SSM

本論文では, 構音障害者を対象とした SSM(Stereo-Based Stochastic Mapping)を用いた音声認識を行う[3]. SSM とは, 観測された特徴量と, 対象となる特徴量間の対応関係を統計的に学習するためのステレオデータに依るアルゴリズムである. 本研究では, 構音障害者と健常者の MFCC 間の結合分布を GMM でモデル化し, 構音障害者のテストデータを健常者の特徴量空間へ変換した後, 健常者の HMM を用いて認識を行う.

3 GMM

3.1 学習ステップ

入力特徴量, 出力特徴量を

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

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, DP マッチングを用いてアライメントを取り, GMM を学習する. 結合確率密度は次式にてモデル化される.

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

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

ここで, $\mathbf{N}(\mathbf{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は平均ベクトル $\boldsymbol{\mu}$, 共分散行列 $\boldsymbol{\Sigma}$ の正規分布を表し, α_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.

3.2 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[4].

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

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

$\boldsymbol{\mu}_m^{(X)}$, $\boldsymbol{\mu}_m^{(Y)}$ 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, $\sum_m^{(XX)}$, $\sum_m^{(YX)}$ 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

* Speech recognition using SSM for articulation disorders, by Ryo Ishii, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

ラス m の相互共分散行列を表す。

4 実験

4.1 実験条件

特徴量として MFCC(12 次元)と Δ MFCC(12 次元)を合わせたものを用い、GMM の学習は構音障害者と健常者の 216 単語から成る同一発話セットで行う。なお、学習時における構音障害者と健常者の MFCC は、ユークリッド距離に基づく DP マッチングによりタイムアラインメントをとったものを用いる。またテストデータに関しては、GMM の学習時と同一の 216 単語を用いた。

4.2 実験結果

Fig. 1 に実験結果を示す。なお Fig. 1 における A は、健常者の 216 単語の MFCC で学習された HMM を用いて構音障害者の MFCC を認識した結果を示したものであり、B は提案手法における認識結果である。

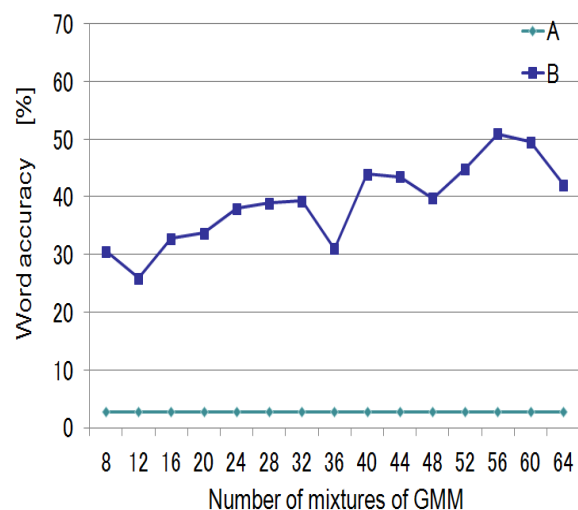


Fig. 1 Word recognition results

Fig. 1 から、提案手法による認識率の向上が確認できた。

また、GMM の学習段階で用いた構音障害者の MFCC を、健常者の MFCC に変換したものでモデル化された HMM を認識に用いて追加実験を行った。なお、この HMM は 108 単語で学習されており、テストデータは HMM の学習に用いていない 108 単語を用いた。以上の条件下での結果が Fig. 2 における D である。なお、Fig. 1 中の B の条件下での GMM の学習単語数を 108 単語にし、テストデータは GMM の学習に用いていない 108 単語で認識を行った結果が Fig. 2 における C で

あり、これを比較対象とした。

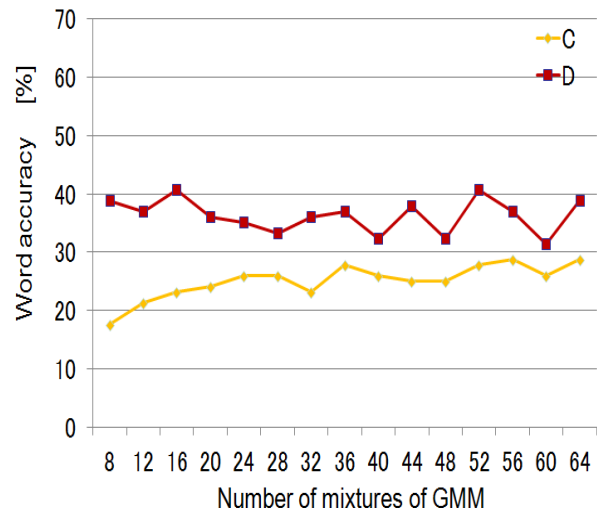


Fig. 2 Word recognition results

Fig. 2 から、GMM の混合数にかかわらず D が C を上回っていることが分かる。

5 おわりに

実験結果として、SSM によって構音障害者を対象とした音声認識率の向上が確認できた。しかし、構音障害者の音声認識を正確に行うためには不十分な結果である。

今後の課題として、更なる認識率の向上を図るために、特徴量のセグメント化による GMM の学習、変換や、MLLR 等の話者適応化と提案手法との比較について検討していきたい。

参考文献

- [1] 馬場朗 他, “高齢者音響モデルによる大語彙連続音声認識” 電子情報通信学会論文誌, Vol. J85-D-II, No. 3, pp. 390-397, 2002.
- [2] 市川薫 他, “福祉と情報技術,” オーム社, 2006.
- [3] M. Afify et al., “Stereo-based stochastic mapping for robust speech recognition,” Proc. EUROSPEECH, pp. 217-220, 2001.
- [4] Y. Stylianou et al., “Statistical methods for voice quality transformation,” Proc. EUROSPEECH, pp. 447-450, 1995.