

スパース表現に基づく構音障害者の発話スタイル変動にロバストな特徴量抽出*

吉岡利也, 高島遼一, 滝口哲也, 有木康雄 (神戸大・工)

1 はじめに

音声認識技術は現在, 様々な環境下や場面において使用される機会が増加している. 文献 [1] では, 構音障害者音声を対象とした音響モデル適応の検証を行っているが, 言語障害者などの障害者を対象としたものは少ない. 言語障害の原因の一つとして, 脳性マヒが考えられる. その分類の一つとして, 不随意運動を伴うアテトーゼ型がある. アテトーゼは緊張時や意図的な動作を行う際に出現しやすい [2]. そのため, アテトーゼ型の構音障害者の場合, 最初の動作において緊張状態により, 通常よりも発話が不安定になる場合がある.

本稿では, 最初の発話スタイルが不安定になるという課題に対して, NMF (非負値行列因子分解) をベースとした発話スタイル変動にロバストな特徴量抽出法を提案する. 提案手法では, 不安定な発話のスペクトルを, 安定した発話のスペクトルパターンのスパースな線形結合 (結合係数のほとんどがほぼ 0) で近似表現し, その結合係数を特徴量として音声認識を行う.

2 構音障害者音声データ

実験用データとして, 構音障害者 1 名のデータを収録した. 発話内容として ATR 音素バランス単語 (216 単語) から 210 単語を無作為に選択した. アテトーゼ型脳性マヒの構音障害者は, アテトーゼの影響により, 最初の発話が他発話よりも不安定になる場合がある. そこで本研究では, 収録方法として, 同一単語の複数回連続発話を行った (図 1). その後, 各発話を手動で切り出した. これは, アテトーゼによる発話スタイルの変動の影響を調べるためである.

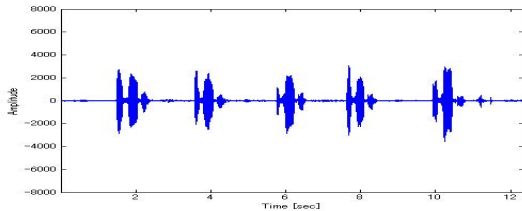


Fig. 1 収録音声データ

3 NMF

3.1 NMF の概要

NMF [3] とは, 全ての要素が非負値である行列 $Y (\in R^{M \times N})$ を二つの非負値行列因子 $A (\in R^{M \times K})$ と $B (\in R^{K \times N})$ に近似分解する手法であり, 以下のように表される.

$$Y \simeq AB \quad (1)$$

音響信号の分解を行う場合, Y は観測信号を短時間フーリエ変換することによって得られた時間周波数領域における振幅スペクトログラムであり, A は元のスペクトログラムに含まれる様々なスペクトルパターン (以下, 基底行列と呼ぶ), B は各スペクトルパターンがそれぞれ現れるタイミングとその大きさ (以下, アクティビティ行列と呼ぶ) を表す. 本研究では, NMF の目的関数として, Y と AB 間のユークリッド距離 J_{NMF} を用いる.

$$J_{NMF} = \|Y - AB\|_F^2 \quad (2)$$

J_{NMF} が最小となるように, 以下の更新式を用いて A, B を求める.

$$a_{i,l}^{[p+1]} = \frac{[YB^T]_{i,l}}{[ABB^T]_{i,l}} a_{i,l}^{[p]} \quad (3)$$

$$b_{l,j}^{[p+1]} = \frac{[A^TY]_{l,j}}{[A^TAB]_{l,j}} b_{l,j}^{[p]} \quad (4)$$

3.2 教師あり NMF

NMF は, その本質として行列を機械的に分解しているに過ぎないので, 本来意図した結果が得られない可能性がある. そこで, 予め基底行列を知識として与えておくことで, 欲しい信号のアクティビティ行列だけを取り出し, 音源分離や解析を行う研究がある [4]. 本研究では, このように, 予め用意した基底行列を用いてアクティビティ行列だけを推定する解析手法を, 教師あり NMF (Supervised NMF: SNMF) と定義する.

4 提案手法

提案手法は, 図 2 のように, 学習ステップと解析ステップに分けることができる.

学習ステップでは, 2 回目以降の安定した発話を用いて基底行列を学習する. まず第 2 ~ 第 5 発話のメル

*Robust feature extraction to utterance fluctuations of articulation disorder based on sparse expression.
by Toshiya YOSHIOKA, Ryoichi TAKASHIMA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe Univ.)

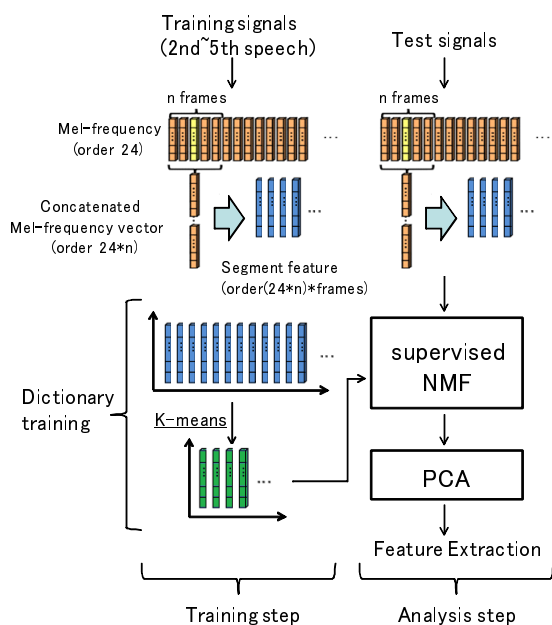


Fig. 2 提案手法による特徴量抽出の流れ

フィルタバンク出力 24 次元のセグメント特徴量を計算し、それを時間方向に連結することで、辞書行列を作成する。さらに、この辞書行列に対して、k-means 法 (k 平均法) を適用しクラスタリングを行ったものを基底行列とする。

解析ステップでは、学習ステップで得られた基底行列に、SNMF を適用することで、入力信号に対するスパースな結合係数 (アクティビティ行列) を推定する。具体的には、基底行列とアクティビティ行列の積と、入力信号の距離を評価関数としたアルゴリズムを用いて、距離が最小となるようにアクティビティ行列の更新を繰り返し、最適なアクティビティ行列を求める。得られたアクティビティ行列に主成分分析 (Principal Component Analysis: PCA) を適用して次元圧縮したものを特徴量とする。

5 認識実験

5.1 実験条件

収録した複数回連続発話データ (210 単語) から 50 単語を無作為に選択して、各単語の第 1 発話に関して単語認識実験を行った。収録データのサンプリング周波数は 16kHz、フレーム幅は 32msec、フレームシフトは 16msec である。特徴量として従来の MFCC (12 次元)、提案手法で得られた特徴量 (13 次元)、MFCC (12 次元) と特徴量 (13 次元) を統合した特徴量 (25 次元) の 3 種類を用い、音響モデルは monophone (54 音素, 3 混合) を使用した。また、セグメント特徴量のフレーム数は 3、k-means のクラスタ数は 100、SNMF の更新回数は 1000 回、主成分の個数は 13 とした。

5.2 実験結果

認識実験の結果を図 3 に示す。従来の MFCC における認識率が 72% であるのに対して、MFCC と提案

手法で得られた特徴量を統合した特徴量を用いることにより、認識率が 84% まで改善された。これにより、スパースな結合係数を従来の MFCC に統合することで、発話スタイル変動の影響を抑圧し、認識率が改善されることが確認できた。しかし、提案手法で得られた特徴量だけでは認識率が 68% となり、MFCC の結果を若干下回ってしまった。その原因として、基底行列に含まれる基底の数や質が不十分であることが考えられる。

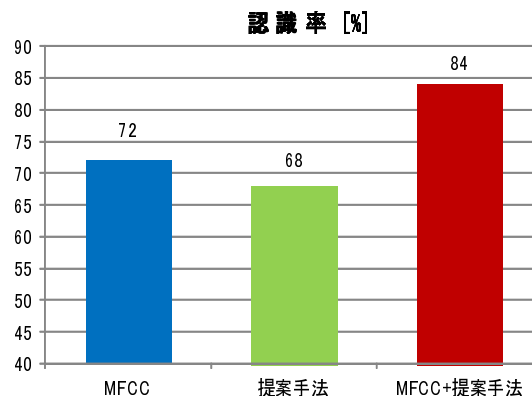


Fig. 3 第 1 発話における認識結果

6 おわりに

本稿では、構音障害者の最初の発話スタイルが不安定になるという課題に対して、スパース表現に基づく発話スタイル変動にロバストな特徴量抽出法を提案し、評価実験によりその有効性を確認した。提案手法で得られた特徴量と従来の MFCC を統合することで、不安定な第 1 発話において 12 ポイント (72% → 84%) の認識率の改善が得られた。今後は、基底行列の作成方法を工夫するなどして、認識率の改善に取り組んでいく予定である。

参考文献

- [1] 中村 他, “ 発話障害者音声を対象にした健常者音響モデルの適応と検証, ” 音講論 (秋), 109-110, 2005.
- [2] S.Terry Canale, 落合 直之, 藤井 克之, “ キャンベル整形外科手術書 第 4 巻 小児の神経障害/小児の骨折・脱臼, ” エルゼビア・ジャパン, 2004.
- [3] D. Lee, and H. Seung, “ Algorithms for non-negative matrix factorization, ” Advances in neural information processing systems, vol.13, 2001.
- [4] 八木浩介, 猿渡洋, 鹿野清宏, 近藤多伸, 高橋祐, “ 拘束条件付き教師あり非負値行列因子分解による目的楽器音抽出とその評価, ” 日本音響学会講演論文集 (秋), 1-15, 2011.