高次元理論を用いた発話障害者の発話スタイルの解析及び音声認識への適用
- Utterance analysis using higher-order theory and speech recognition for a person with articulation disorders -

(登録番号)
研究者代表　神戸大学都市安全研究センター・講師　滝口哲也

[研究の目的]

近年，家庭生活においても様々な機器の情報化が進み,情報家電が浸透しつつある.しかし，そのような機器は操作が複雑であり，障害者が使いこなすには困難である場合が多い.

重度障害者の自立支援機器として，環境制御装置は重要な役割を担っている.これまでにも，幾つかの環境制御装置が開発，実用化されている．従来の環境制御装置では操作方法として，押しボタンスイッチを用いたものや，呼気を用いたもの，音声認識を用いて操作を行うものもある.

人が最も使いやすいインターフェースとしては,「声」があげられる．声を使って機器の操作が可能になれば，障害者にとっては，ユビキタス社会での自立した生活への期待が持てる.

これまでの音声認識の研究は，発話障害を持たない成人男女の発話に対するものがほとんどであった．そのため発話障害者などの発話スタイルが異なった場合では，従来の音響モデルでは認識が困難であると考えられる．本研究課題においては，これまでになかった新しいソフト面でのユニバーサルデザインとして発話障害音声に注目し，ユビキタス社会における発話障害者の自立支援に役立てる.

[研究の内容，成果]

1. 健常者音響モデルでの認識

脳性麻痺の症状における発話障害のレベルは様々であり，例えば次のような分類が考えられる．1) 痙直型，2) アテトーゼ型，3) 失調型，4) 緊張低下型，5) 固縮型， それぞれの症状が混合して現れる混合型もある．これらのうちアテトーゼ型脳性麻痺では，知能障害を合併していないケースや比較的知能障害の程度が軽いケースが多いため，まず我々はこのタイプの音声データに注目した．収録した障害者音声データを，健常者の音声データで学習した音響モデルを使用している汎用の音声認識ソフトウェア Julius を使い 210 単語認識を行ったところ，2.4%しか認識出来なかった．一方，健常者の音声データに対しては 88.4%の単語正解精度（認識率）が得られた．従来の健常者用音響モデルでは発話スタイルが全く異なるため，認識が困難である事が分かる．実験結果を図 1 に示す.

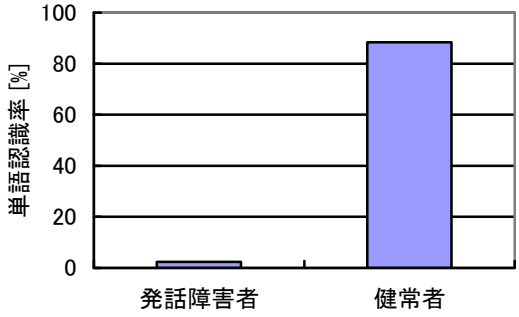

図 1. 健常者用不特定話者音響モデルでの認識結果

Fig. 1: Recognition results with a speaker-independent model of unimpaired persons

健常者の不特定話者音響モデルでの音声認識が困難であることから，発話障害者の音響モデルを作成し認識実験を行った．実験用データとして，各音素バランス単語(210 単語)を 5 回連

続発声し，各発話を手動で切り出したファイルを使用した．1 回目の発話の認識を行う場合は 2～5 回目の発話を用いて音響モデルを作成した．これを各発話に対して行う．各発話における認識率を図 2 に示す．使用した音声特徴量は MFCC12 次元+ΔMFCC12 次元である．図 2 より，1 回目の発話が 77.1%と他の発話に比べると著しく低下している．これは 1 回目の発話は最初の意図的な動作であり，他の発話よりも緊張状態に陥っていると考えられる．そのためアテトーゼが生じて調音が困難になり，発話スタイルが不安定となることから認識精度が低下したと考えられる．

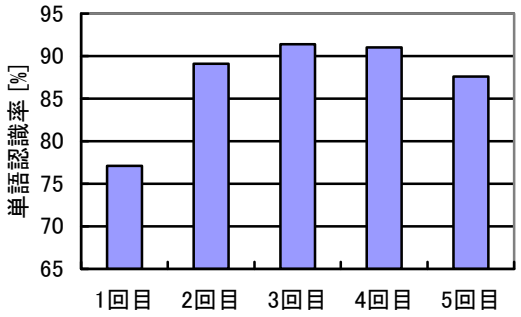

図 2. 発話回数毎の認識率
Fig. 2: Recognition results for each utterance

2. 主成分分析による発話バリエーションの正規化

まず，本研究課題においてアテトーゼ型に見られる，筋肉の緊張から生じる発話スタイルのバリエーションの正規化手法を検討した．従来の音声認識システムでは MFCC (Mel Frequency Cepstral Coefficient) が音声特徴量として用いられている. MFCC ではメル尺度フィルタバンクの短時間対数エネルギー出力系列に対して，離散コサイン変換（Discrete Cosine Transform: DCT）を適用し，ケプストラムが得られる．そして音声のスペクトル包絡成分に対応する低次ケプストラムのみを抽出し，特徴量として音声認識に用いられる．本研究課題では，発話スタイルの変動にロバストな特徴量抽出法として, 離散コサイン変換の代わりに, 主成分分析（Principal Component Analysis）を用いた手法を検討した（図 3）.

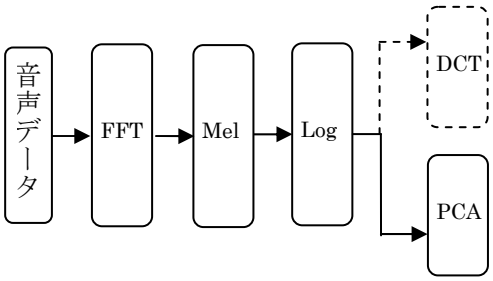

図 3. PCA を用いた特徴量抽出
Fig. 3: Feature extraction using PCA

前述した通り，アテトーゼ型の発話障害の場合，最初の動作において緊張状態により通常よりも不安定な場合がある．そこで提案手法では 2 回目発話以降の安定した発話を用いて主軸を求め，1 回目の発話（調音不安定音声）に対して対数スペクトル上で PCA を適用する．本手法において, 安定した音声成分は低次元空間に, 一方調音不安定成分は高次元空間に集まると期待出来る．この結果 PCA により調音不安定成分の抑圧が行われ，より有効な（安定した）スペクトル情報の抽出が可能となる．具体的な処理を以下に述べる．

短時間分析によって得られたフレーム n，周波数 ω の観測音声を $O_n(\omega)$，クリーン音声を $S_n(\omega)$ とする．ここで，1 回目の音声を以下の式で表現する．

$$X_n(\omega) = S_n(\omega)H(\omega)$$

1 回目発話には発話スタイル変動成分 H が畳み込まれていると考える．次に対数変換を行い，観測信号を S と H の加算で表す.

$$\log X_n(\omega) = \log S_n(\omega) + \log H(\omega)$$

ここで観測信号 O に対して PCA を適用すると，

- クリーン音声 S の主なエネルギーは D 個の主な固有値に集中する.
- それ以外の固有値に対応する主な成分は，付加成分である.

と期待出来る．ここで，主な D 個の固有値に対

応する固有ベクトルを V とすると, この V を用いて以下のようなフィルタリングを考える.

$$\hat{S} = VO$$

このフィルタリングによって付加成分 H を抑圧する事が出来る．本研究課題では，主軸 V の計算には 2 回目以降の発話音声を用いて，上記フィルタリングを 1 回目の調音不安定音声に適用する．図 4 に主成分が 17 個の場合の認識結果を示す(比較用に MFCC17 次元の結果も示す). DCT (MFCC) の代わりに PCA を適用し，高次空間における不安定要素を抑圧することにより，1 回目発話において 83.8%まで認識率が改善された．DCT よりも PCA の方が 1 回目発話の調音不安定音声において，より有効的な特徴量抽出法であるといえる.

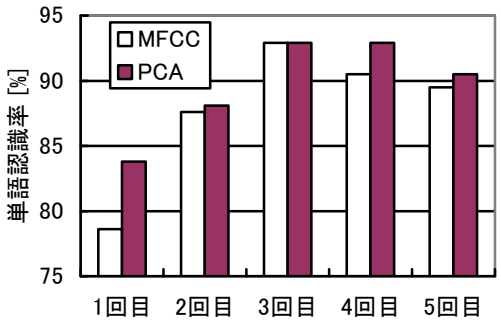

図 4. PCA を用いた発話バリエーションの正規化

Fig. 4: Recognition results by feature extraction using PCA

3. 障害者発話区間検出

従来の音声認識器では，周囲雑音の影響を受けないようにするために, 音声入力スイッチ(マイクロフォンスイッチ）が必要となっている．脳性麻痺障害者にとっては，発話障害だけでなく，身体的な障害を伴っている場合があり，マイクロフォンスイッチ等を操作するのは非常に困難である．本研究課題において，機械学習法の一つである AdaBoost を用いた障害者発話区間検出を検討した.

AdaBoost は，単純な識別器を複数組合わせることによって精度の高い識別器を構成する Boosting 法の中でも顕著な性能を示す方法である．本研究課題においては，AdaBoost を発展させた Real AdaBoost を用いて障害者発話の検出を行った.

学習データ数を n , 繰り返し回数を M とする. これらの値はあらかじめ決定しておく必要がある．学習データを x_i $(i=1,...,n)$，各データの重みを w_i $(i=1,...,n)$ とする． x_i は音声特徴量を表し, ここでは MFCC を使用した. 各データ x_i にはあらかじめラベルとして $y \in \{-1,1\}$ を与えておく．すなわちデータ x_i が音声であれば $y=1$，データが非音声であれば $y=-1$ とする.

1. 各データに対する重みを $w_{1i}=1/n$ で初期化する.
2. $m=1,...,M$ まで，以下を繰り返す.
   (a) w_{mi} を重みとして， x_i から重複を許して n 個, リサンプリングしたものを x_i' とする.
   (b) x_i' に対して弱識別器 $f_m(x)$ を構成する．ここでは， $f_m(x)$ に CART を用いた.
   (c) 得られた弱識別器 $f_m(x)$ を用いて以下の $c_m(x_i)$ を計算する.

$$c_m(x_i) = \frac{1}{2}\log\left(\frac{f_m(x_i)}{1-f_m(x_i)}\right)$$

   (d) 各データ x_i の重みを更新する.

$$w_{(m+1)i} = \frac{w_{mi}\exp\{-y_i c_m(x_i)\}}{\sum_{r=1}^{n} w_{mr}\exp\{-y_r c_m(x_r)\}}$$

3. 弱識別器の線形結合により，強識別器 $F(x)$ を得る.

$$F(x) = \sum_{m=1}^{M} c_m(x)$$

AdaBoost の重要な性質の一つに，誤ったデー

タに対するサンプリング重みを増やし，次回以降の学習でそれらのデータを重点的に学習する，ということがあげられる．それにより前段の弱識別器において誤識別を起こしてしまったデータに対しても，後段の弱識別器により正しく識別する事が期待出来る．図 5 に発話障害者の発話区間検出結果を示す．弱識別器の数は 50，学習用音声データは 340 発話，非音声区間として 80 秒のデータを用いた（男性話者 1 名）．図 5 の縦軸，横軸は各々以下を表す．

FRR (False Rejection Rate)
=「非音声と誤検出されたフレーム数」/ 「音声の総フレーム数」

FAR (False Acceptance Rate)
=「音声と誤検出されたフレーム数」/ 「非音声の総フレーム数」

健常者用モデルと比較評価を行ってみたところ，約 2%から 3%程度の改善が得られた．発話区間検出に関しては，健常者用モデルと発話障害者用モデルとの間に，あまり大きな精度差は見られなかったが，発話障害者においては個人差のばらつきが大きく，更に多くの被験者で評価を行う必要がある．

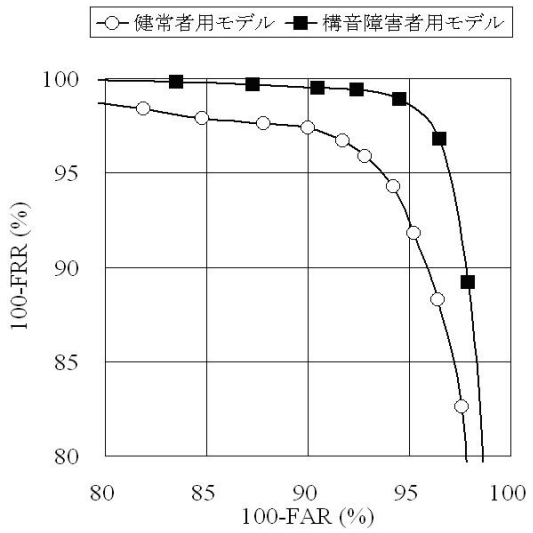

図 5. 発話区間検出精度
Fig. 5: Accuracy of VAD

[今後の研究の方向，課題]

脳性麻痺発話障害者は，一人一人その障害特性が多様であるため，今後さらに様々な話者の多量のデータを蓄積し，発話障害者の解析を行い，音響モデルのカスタマイズ機能の改善を計っていく必要がある．また提案したアルゴリズムの有効性は有限個の単語認識に対してのみ示されている．数十単語程度のコマンド入力だけでは，障害者にとってはシステムの使い勝手が決して十分であるとは言えない．今後，任意の発話文（話し言葉）の認識についても検討を進める予定である．

[成果の発表，論文等]

- Hironori Matsumasa, Tetsuya Takiguchi, Yasuo Ariki, Ichao LI, Toshitaka Nakabayashi, “PCA-Based Feature Extraction for Fluctuation in Speaking Style of Articulation Disorders,” Interspeech2007, pp. 1150-1153, 2007.
- 朴　玄信，滝口　哲也，有木　康雄, “音素部分空間の統合による音声特徴量抽出の検討”, 第 9 回音声言語シンポジウム，SP2007-137, pp.241-246, 2007.
- 松政宏典, 田中克幸, 滝口哲也, 有木康雄, 李義昭，中林稔堯, “情報家電操作における脳性麻痺構音障害者の音声認識評価”, 電子情報通信学会技術研究報告，WIT2007-7，pp33-38，2007.
- 室井貴司，滝口哲也，有木康雄, “フィッシャー重みマップに基づく音声特徴量のロバストネスに関する考察”, 日本音響学会 2007 年秋季研究発表会，1-P-27，pp. 191-192，2007.
- 松政宏典，滝口哲也，有木康雄，李義昭，中林稔堯, “話者正規化に基づく構音障害者の音声認識 ”，日本音響学会 2008 年春季研究発表会, 1-Q-24, pp. 215-216，2008.