話者正規化に基づく構音障害者の音声認識*

◎松政宏典, 滝口哲也, 有木康雄 (神戸大・工), 李義昭 (追手門大), 中林稔堯 (神戸大・発達)

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非常に少ない．

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．そこで，我々は PCA (Principal Component Analysis) による発話変動にロバストな特徴量抽出法を提案してきた [3]．

本稿では，さらなる改善として，各話者の音素毎の置換，挿入の傾向を音声認識の過程に組み込むことが可能な Metamodel[4] との統合を試み，その有効性を示す．

2 Metamodel

単語認識を行う場合，音素集合を $\mathcal{P}$，入力音声データを $\boldsymbol{A}$ とした場合，単語 w は以下の式で求められる．

$$Pr(w|\boldsymbol{A}) = \sum_{p\in\mathcal{P}} Pr(w|p)Pr(p|\boldsymbol{A}) \tag{1}$$

式（1）は，音素認識列 $\boldsymbol{p^*}$

$$\boldsymbol{p^*} = \arg\max_{p\in\mathcal{P}} Pr(p|\boldsymbol{A}) \tag{2}$$

を用いることで，以下のように近似される．

$$Pr(w|\boldsymbol{A}) \simeq Pr(w|\boldsymbol{p^*})Pr(\boldsymbol{p^*}|\boldsymbol{A}) \tag{3}$$

よって，単語 $\acute{w}$ は以下の式で求められる．

$$\acute{w} = \arg\max_{w\in\mathcal{W}} Pr(w|\boldsymbol{p^*}) \tag{4}$$

音素認識列から，尤もらしい単語の推定を行うために，Metamodel（図 1）を用いる手法がある [4]．音響モデルと同様に，Metamodel も音素毎に構成される．通常の音響モデルの場合は，入力がフレーム単位の特徴量（MFCC など）に対し，Metamodel の場合は，音素認識で得られる音素列を入力として扱う．音素列を入力として扱うことで，状態 A, C では発話前後の挿入を，状態 B では置換を考慮することが可能である．各状態は離散出力確率を持ち，初期出力確率として confusion matrix $Pr(p_{出力}|p_{入力})$ を与える．学習は，Baum-Welch アルゴリズムによって行われる．

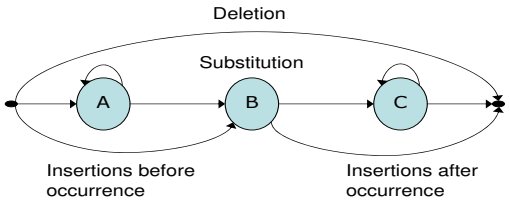

Fig. 1 Architecture of the metamodel of a phoneme

3 提案手法

3.1 PCA を用いた特徴量抽出

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 $X_n(\omega)$，クリーン音声を $S_n(\omega)$ とする．ここで発話不安定な音声を以下の式で表現する．

$$X_n(\omega) = S_n(\omega)H(\omega) \tag{5}$$

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．

$$\log X_n(\omega) = \log S_n(\omega) + \log H(\omega) \tag{6}$$

ここで観測信号 X 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できる．ここで，主な D 個の固有値に対応する固有ベクトルを $\boldsymbol{V} = [\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(D)}]$ とすると，この $\boldsymbol{V}$ を用いて以下のようなフィルタを考える．

$$\hat{\boldsymbol{S}} = \boldsymbol{V^t X} \tag{7}$$

このフィルタリングによって付加成分 $H(\omega)$ を抑圧することが出来る．また主軸 $\boldsymbol{V}$ の計算をクリーン音声のみから行えば，クリーン音声の構造のみを考慮したフィルタを作成することが出来る．本稿では，主軸 $\boldsymbol{V}$ 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．

本手法において，安定した音声成分は低次に，発話スタイル変動成分は高次に集まると期待できる．この結果 PCA により発話スタイル変動成分の抑圧が行われ，より有効なスペクトル情報の抽出が可能となる．

3.2 Metamodel と音響モデルの統合

文献 [4] では，構音障害者に対して，Metamodel を用いて話者適応を行っている．本稿では，発話スタイルの変動により生じると考えられる音の挿入，置換に対して，抑圧方法として Metamodel を音響モデ

*Speech Recognition of a Person with Articulation Disorders based on Speaker Normalization. by MATSUMASA Hironori, TAKIGUCHI Tetsuya, ARIKI Yasuo (Kobe Univ.) and LI Ichao (Otemon Univ.) and NAKABAYASHI Toshitaka (Kobe Univ.)

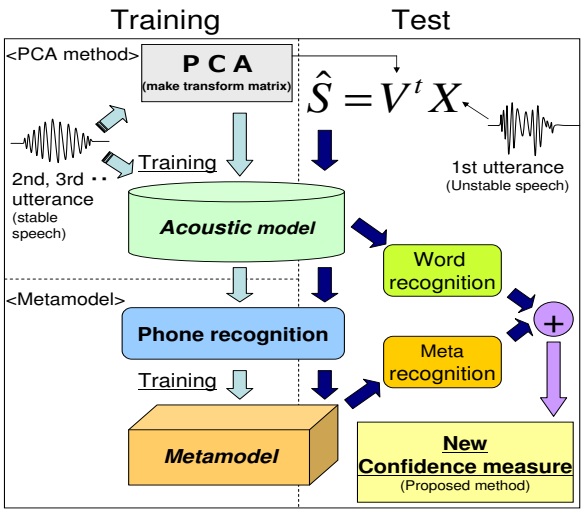

Fig. 2 Integration of Metamodel and Acoustic model

Table 1 発話回数毎の認識率 (%)

特徴量	1 回目	2 回目	3 回目	4 回目	5 回目
MFCC	79.1	87.6	90.0	91.4	90.0
PCA	85.2	90.0	89.5	91.9	89.1

ルと統合して用いる．システムの概要図を図 2 に示す．両モデルの統合により，特徴量抽出時の抑圧だけでなく，認識時の抑圧が可能になると考えられる．認識時において，両モデルの尤度に対して，以下の式で統合を行う．L_{Aco}, L_{Meta} は，それぞれ音響モデル，Metamodel の尤度を表す．

$$L_{Aco+Meta} = (1-\alpha)\cdot L_{Aco} + \alpha \cdot L_{Meta} \qquad (8)$$

4 認識実験

4.1 実験条件

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．収録は各単語を 5 回連続発声し，その後，各発話を手動で切り出した．収録データのサンプリング周波数は 16 kHz である．音響モデルは monophone（54 音素，6 混合）を用い，フレーム周期は 10 ms，ハミング窓長は 25 ms とし，実験には HTK[5] を用いた．

4.2 構音障害者音響モデルでの認識実験

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．特徴量は 26 次 MFCC (13 次 MFCC + △) を用いた．特定話者モデルに対する，発話回数毎の認識率を表 1 に示す．

1 回目発話の認識率が 79.1% 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

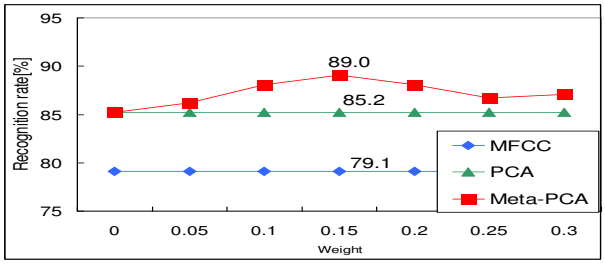

Fig. 3 Recognition rate for the 1st utterance by the proposed method

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．

4.3 PCA を用いた特徴量抽出法による認識実験

メルフィルタバンク出力 24 次元に対し PCA を適用した結果を示す．得られた値を基本係数とし，基本係数+Δ係数を音声認識の特徴量とした．主成分の個数は 15 個を用いた．

DCT の代わりに PCA を適用することにより，1 回目発話において，85.2% まで認識率が改善された．発話回数毎の認識率を表 1 に示す．

4.4 Metamodel と音響モデルの統合

単語認識時の 3Best 単語に対して，Metamodel と音響モデルの統合を行った結果を示す．音素認識（phone-loop）は 4.3 章で用いた音響モデルを使用し，認識された音素列を用いて，Metamodel を作成した．Metamodel に対する重み α を 0.0 〜 0.3 として実験を行った．図 3 に 1 回目発話における結果を示す．

Metamodel との統合によって，重みが 0.15 の際に，89.0%まで認識率が改善された．単語認識のみでは考慮できなかった各音素の置換，削除を Metamodel を用いることによって考慮が可能となり，認識率の改善が得られたと考えられる．

5 おわりに

構音障害者の 1 回目の調音不安定発話に対して，PCA による特徴量抽出法とさらに，音響モデルと Metamodel の統合を検討した．提案手法によって 9.9% (79.1%→ 89.0%) の改善が得られた．今後は，構音障害者特有の特徴量の検討や，様々な音声認識手法の適用を行い認識率の改善に取り組んでいく．また更に対象者を増やしていく予定である．

参考文献

[1] 中村 他, 音講論（秋）, 109-110, 2005.
[2] S. Terry Canale *et al.*, “ キャンベル整形外科手術書 第 4 巻, ” エルゼビア・ジャパン, 2004.
[3] 松政 他, 音講論（春）, 323-324, 2007.
[4] O. Caballero Morales *et al.*, Interspeech2007, 1565-1568, 2007.
[5] S. Young *et al.*, “ The HTK Book, ” Entropic Labs and Cambridge University, 1995-2002.