メタモデルと音響モデルの統合による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

言語障害の原因の一つとして，脳性マヒが考えられる．そのため，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 [1]．また，構音障害者の場合は，調音を意図した通りに行うことが困難であり，そのため，意図していない音の発声，挿入が起こりやすいと考えられる．

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[2] との統合を試み，複数話者に対してその有効性を示す．

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．

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

式（1）は，音素認識列 $\boldsymbol{p^*}$

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

を用いることで，以下のように近似される．

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

よって，単語 $\acute{w}$ は以下の式で求められる．

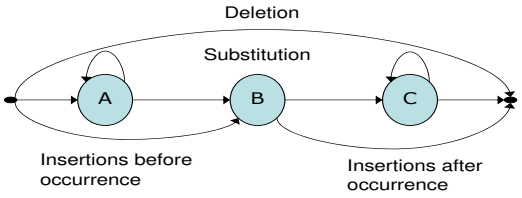
$$\acute{w} = \arg\max_{w\in\mathcal{W}} Pr(w|\boldsymbol{p^*}) \quad (4)$$

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 [2]．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，学習データを音素認識したものを用いて，Baum-Welch アルゴリズムによって行われる．

3 Metamodel と音響モデルの統合

文献 [2] 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 を音響モデルと統合して用いる．システムの概要図を図 2 に示す．特徴量抽出に PCA を行わない場合は，$\hat{\boldsymbol{S}} = \boldsymbol{V^t X}$ のステップは省略される．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 の尤度を表し，N_{frame} , N_{phone} は，それぞれ音響特徴量のフレーム数，音素認識結果の音素列数を表す．統合の精度を高めるために，単語認識結果の N-best 単語に対してのみ統合を行う．

$$\begin{aligned} & L_{Aco+Meta}^{\acute{w}_{N-best}} \\ &= (1-\alpha)\cdot L_{Aco}^{\acute{w}_{N-best}} + \alpha \cdot L_{Meta}^{\acute{w}_{N-best}} \\ &= (1-\alpha)\cdot Pr(\boldsymbol{A}|\acute{w}_{N-best})/N_{frame} \\ &\quad +\alpha \cdot Pr(\boldsymbol{p^*}|\boldsymbol{\acute{w}_{N-best}})/N_{phone} \end{aligned} \quad (5)$$



Discrete probability distribution

Phoneme / State	A	B	C
a	0.1	0.7	0.2
i	0.4	0.1	0.3
u	0.3	0.05	0.3
e	0.1	0.05	0.1
o	0.1	0.1	0.1

Fig. 1 Architecture of the metamodel of a phoneme

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 3 名のデータを収録した．話者 A は発話内容として 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 回連続発声した．話者 B,C は，210 単語とさらに，ATR 単語発話データベース（2620 単語）から 1400 単語を選択した．収録は各単語を 3 回連続発声した．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[3] を用いた．

*Integration of Metamodel and Acoustic Model for Dysarthric Speech Recognition. by MATSUMASA Hironori, TAKIGUCHI Tetsuya, ARIKI Yasuo (Kobe Univ.) and LI Ichao (Otemon Univ.) and NAKABAYASHI Toshitaka (Kobe Univ.)

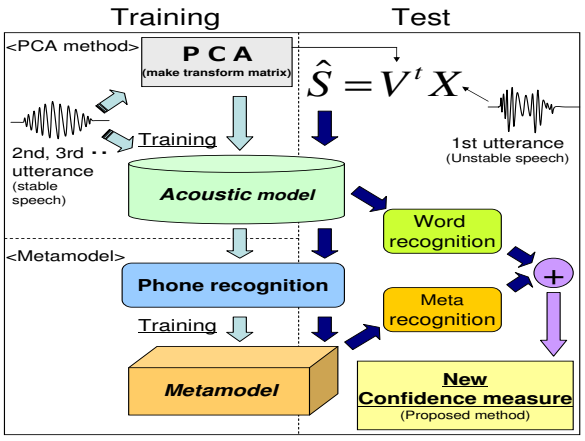


Fig. 2 Integration of Metamodel and Acoustic model

Table 1 発話回数毎の認識率 (%)[話者 A]

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	89.0
Meta-PCA	89.0	91.9	90.0	93.8	88.1

4.2 構音障害者音響モデルでの認識実験

汎用モデルでの認識が困難であることから，構音障害者の音響モデルを作成し認識実験を行った．1 回目の発話の認識を行う場合は残りの発話（話者 A の場合；2 回目以降発話，話者 B,C の場合；2,3 回目のテストデータを含まない 1400 単語）を用いて音響モデルを作成した．特徴量は 26 次 MFCC (13 次 MFCC + △) を用いた．話者 A の場合，1 回目の発話は，他の発話よりも高い緊張状態に陥っていると考えられ，認識精度の低下が見られる（表 1）．そこで，メルフィルタバンク出力 24 次元に対し PCA を適用した．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（表 1）．図 3 に話者 A の結果を示し，図 5 に話者 B, C の結果を示す．

4.3 Metamodel と音響モデルの統合

単語認識時の 話者 A ; 3Best 単語, 話者 B,C ; 2Best 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 4.2 節で用いた音響モデルをそれぞれ使用し，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 〜 0.5 として実験を行った．図 4 に話者 A の 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 0.35 の際に，89.0%まで認識率が改善された．話者 B,C に関しては，重みが 0.4 の際に最も良い改善が得られた．図 3 に話者 A の結果を示し，図 5 に話者 B, C の結果を示す．

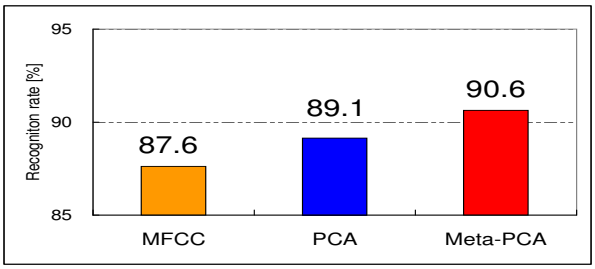


Fig. 3 Recognition results for the speaker-dependent model [Speaker A]

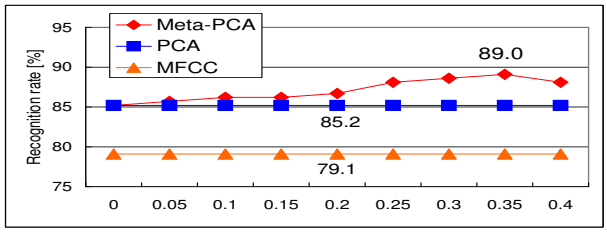


Fig. 4 Recognition rate for the 1st utterance by the proposed method [Speaker A]

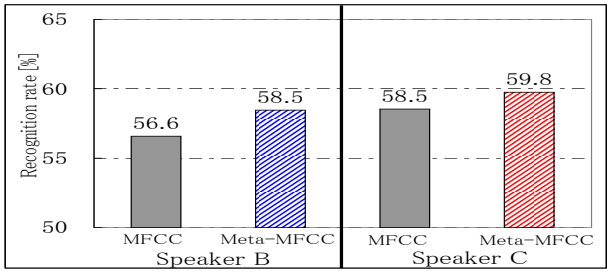


Fig. 5 Recognition results for the speaker-dependent model [Speaker B, Speaker C]

果を示す．

5 おわりに

構音障害者の調音不安定発話に対して，音響モデルと Metamodel の統合を検討した．提案手法によって話者 A の 1 回目の調音不安定音声では，9.9% (79.1%→ 89.0%) の改善が得られた．それぞれの話者において，全発話平均で 1.6 % の改善が得られた．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．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] 松政 他，信学技報，WIT2008-7，37-42，2008.
[2] O. Caballero Morales *et al.*, Interspeech2007, 1565-1568，2007.
[3] S. Young *et al.*, “ The HTK Book, ” Entropic Labs and Cambridge University, 1995-2002.