社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

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† 　李　 義昭†† 　中林　稔堯†††

† 神戸大学工学研究科　〒 657–8501 兵庫県神戸市灘区六甲台町 1–1
†† 追手門学院大学経済学部　〒 567–8502 大阪府茨木市西安威 2–1–15
††† 神戸大学発達科学部　〒 657–8501 兵庫県神戸市鶴甲 3–11
E-mail: †mattu28@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki,nakaba}@kobe-u.ac.jp,
†††chao55@res.otemon.ac.jp

あらまし　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．しかし，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．本稿では，アテトーゼ型脳性マヒによる構音障害者の音声認識の検討を行う．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．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メタモデル（Metamodel）との統合を試み，その有効性を示す．
キーワード　構音障害，言語障害，脳性マヒ

Integration of Metamodel and Acoustic Model for Speech Recognition

Hironori MATSUMASA†, Tetsuya TAKIGUCHI†, Yasuo ARIKI†, Ichao LI††, and Toshitaka NAKABAYASHI†††

† Graduate School of Engineering, Kobe University　1–1 Rokkodaicho, Nada-ku,Kobe,Hyogo, 657–8501 Japan
†† Faculty of Economics, Otemon Gakuin University　2–1–15 Nishiai, Ibaraki,Osaka, 567–8502 Japan
††† Faculty of Human Development, Kobe University　3–11 Tsurukabuto, Nada-ku, Kobe,Hyogo, 657–8501 Japan
E-mail: †mattu28@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki,nakaba}@kobe-u.ac.jp,
†††chao55@res.otemon.ac.jp

Abstract　Recently, the accuracy of speaker-independent speech recognition has been remarkably improved by use of stochastic modeling of speech. However, there has been very little research on orally-challenged people, such as those with speech impediments. Therefore we have tried to build the acoustic model for a person with articulation disorders. The articulation of the first utterance tends to become unstable due to strain of a muscle and that causes degradation of speech recognition, where MFCC (Mel Frequency Cepstral Coefficients) is used as speech features. Therefore we proposed a robust feature extraction method based on PCA (Principal Component Analysis) instead of MFCC. In this paper, we discuss our effort to integrate a Metamodel and Acoustic model approach. Metamodel has a technique for incorporating a model of a speaker's confusion matrix into the ASR process in such a way as to increase recognition accuracy. Its effectiveness is confirmed by word recognition experiments.
Key words　articulation disorders,cerebral paralysis

1. はじめに

情報技術が向上し，近年，福祉分野への情報技術の適用が行われている．例えば，画像認識技術を用いた手話認識 [1] や，文書内の文字の音声化などが行われている [2]．また，音声合成を用いて，発話障害者支援のための音声合成器の作成なども行われている [3]．

音声認識技術は現在，車内でのカーナビの操作，会議においての書き起こしなど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．対象者が子供である場合などには精度が低下することがわかっている [4]．文献 [5] 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．現在，日本だけでも構音障害者も含まれる言語障害者が 3 万 4000 人もいることから十分なニーズがあり，研究の必要性があるといえる [6] [7]．

言語障害の原因の一つとして，脳性マヒが考えられる．脳性マヒとは新生児 1000 人あたりおよそ 2 人の割合で発生する．脳性マヒの定義として厚生省は「受胎から生後４ 週以内の新生児までの間に生じた，脳の非進行性病変に基づく，永続的な，しかし変化しうる運動および姿勢の異常である．その症状は満 2 歳までに発現する.」（1968 年）と定義している．

脳性マヒの原因は，中枢神経系の損傷によるものであり，それに伴う運動障害であると考えられている．発症時期として出生前（胎内感染，母体の中毒，栄養欠損など），分娩時（脳外傷，脳出血，脳の無酸素状態，胎児黄疸，仮死状態，未熟での出産など），出生後（脳内出血，脳炎など）の 3 つの要因が考えられている．

脳性マヒは次のように分類される．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．

本稿ではアテトーゼ型の脳性マヒによる構音障害者を対象としている．アテトーゼ型とは脳性マヒ患者の約 10〜15%に発生する症状であり，大脳基底核と呼ばれる視床下部，脳幹，小脳と関連を持ち随意運動，姿勢，筋緊張を調節する働きをしている部位に損傷をうけたためアテトーゼと呼ばれる不随意運動が伴う型である．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．症状は軽度から重度まで様々であり，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も多いのが特徴である [8] [9]．そこで本稿では，まず知能障害を合併していないアテトーゼ型に着目した．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，通常よりも不安定になる場合がある．そこで本稿では，複数回連続発話の収録を行ない，アテトーゼによる発話スタイルの変動の影響を調べる．

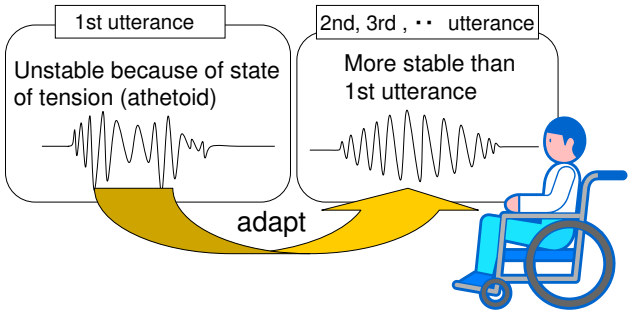


図 1　構音障害者に対する改善概略図

従来の音声認識では，対数スペクトルに対し離散コサイン変換を適用した MFCC を特徴量として用いるが，我々は離散コサイン変換ではなく 2 回目以降のより安定したデータを利用した（図 1），PCA (Principal Component Analysis) による発話変動にロバストな手法を提案してきた [10]．

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 メタモデル [11] [12] との統合を試み，その有効性を示す．

2. メタモデル [11] [12]

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 $\boldsymbol{p}^*$

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

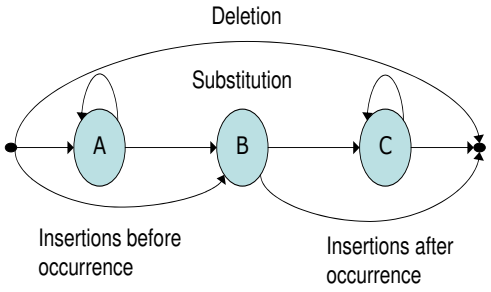
を用いることで，以下のように近似される．

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

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，メタモデルを用いる手法がある．文献 [11] [12] で



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

図 2　メタモデル概形図

は，認識仮説の精度向上や話者適応のために，メタモデルを使用している．メタモデルは離散 HMM から形成され（図 2），音響モデルと同様に，メタモデル 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，メタモデル 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 $Pr(p_{\text{出力}}|p_{\text{入力}})$ を与える．学習は，Baum-Welch アルゴリズムによって行われる．

3. 提 案 手 法

3.1 PCA を用いた特徴量抽出

音声認識システムにおいて従来は 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 Transformation: 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．我々は，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 PCA を用いた手法を提案してきた（図 3）．文献 [13] では残響下でのロバストな特

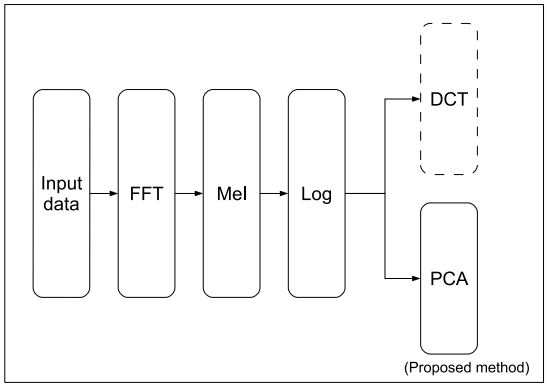


図 3　PCA を用いた特徴量抽出

徴量抽出として対数スペクトル上で，DCT を行わず PCA を行う事によって有効なスペクトル情報の抽出を可能としている．本稿では 2 回目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．

3.2 発話スタイル変動成分の抑圧

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 $S_n(\omega)$ する．ここで 1 回目の音声を以下の式で表現する．

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

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．

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

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．

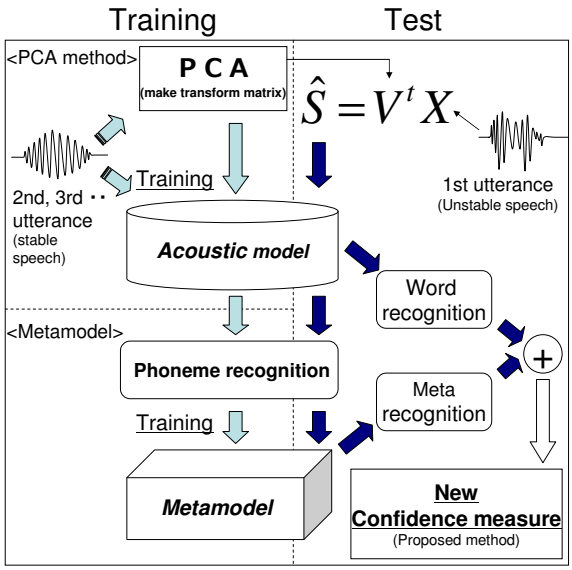


図 4　メタモデルと音響モデルの統合

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．

3.3　メタモデルと音響モデルの統合

文献 [12] では，構音障害者に対して，メタモデル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 メタモデルを音響モデルと統合して用いる．文献 [12] では，適応データが少ない場合に優位性が見られたが，本稿では，十分なデータを用いて音響モデルを作成するため，統合の場合に優位性が現れると考えられる．システムの概要図を図 4 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，メタモデルの尤度を表す．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（図 5）．

$$
\begin{aligned}
&L_{Aco+Meta}^{\acute{w}_{N-best}} \\
&= (1-\alpha) \cdot L_{Aco}^{\acute{w}_{N-best}} + \alpha \cdot L_{Meta}^{\acute{w}_{N-best}} \\
&= (1-\alpha) \cdot Pr(\boldsymbol{A}|\acute{w}_{N-best}) + \alpha \cdot Pr(\boldsymbol{p^*}|\boldsymbol{\acute{w}_{N-best}})
\end{aligned}
\tag{8}
$$

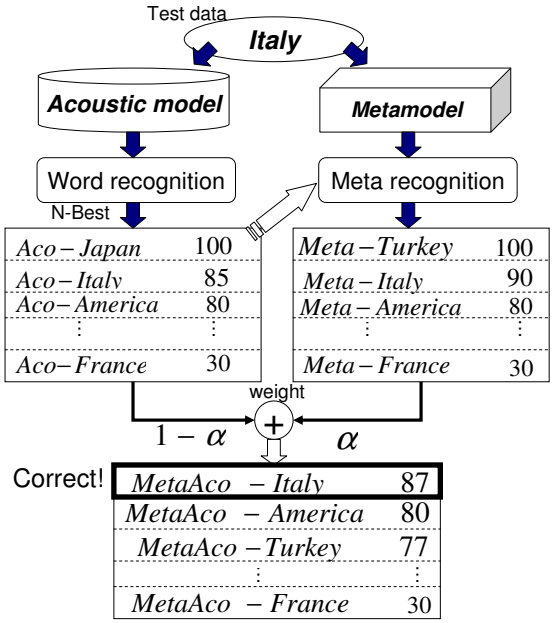


図 5　認識時の統合例

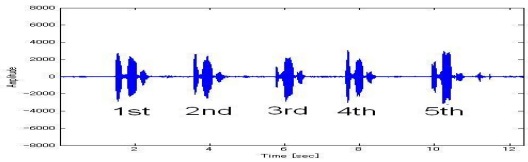


図 6　収録データ

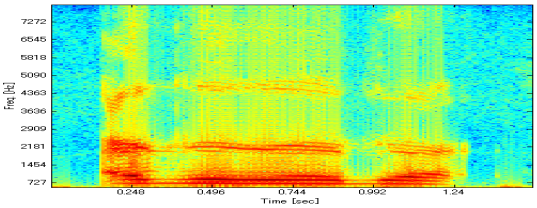


図 7　構音障害者のスペクトログラム例//a k e g a t a

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（図 6），その後，各発話を手動で切り出した．図 7 に構音障害者，図 8 に健常者のスペクトログラム例を示す．構音障害者の場合，子音など高域のパワーが弱く，明瞭度が劣化している．

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

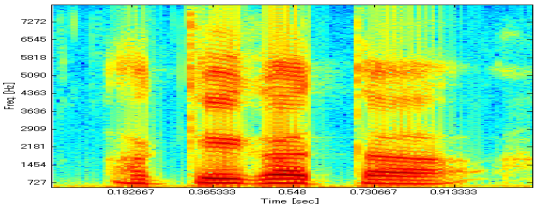


図 8　健常者のスペクトログラム例//a k e g a t a

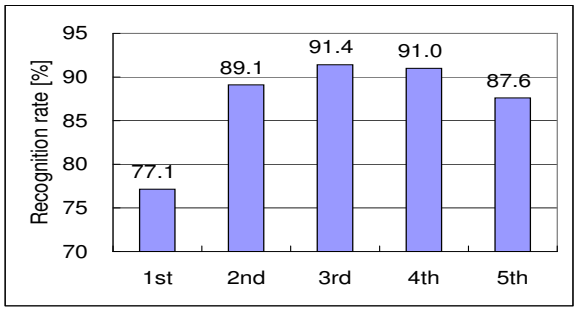


図 9　構音障害者における発話回数ごとの認識率

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．初期モデルの作成，学習，認識には HTK [14] を用いた．実験条件を表 1 に示す．

表 1　実験条件（構音障害者音響モデル）

サンプリング周波数	16 kHz
ハミング窓長	25 msec
フレーム周期	10 msec
音響モデル	monophone (3 状態 54 音素)
特徴パラメータ	12 次 MFCC+Δ 12 次 MFCC
混合分布数	6
テストデータ	1050 (210 単語× 5 回)
辞書	210 単語

特定話者モデルを用いる事で構音障害者において，平均で 87.2%の認識が得られた．構音障害者における発話回数ごとの認識率を図 9 に示す．構音障害者において，1 回目発話の認識率が 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．

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

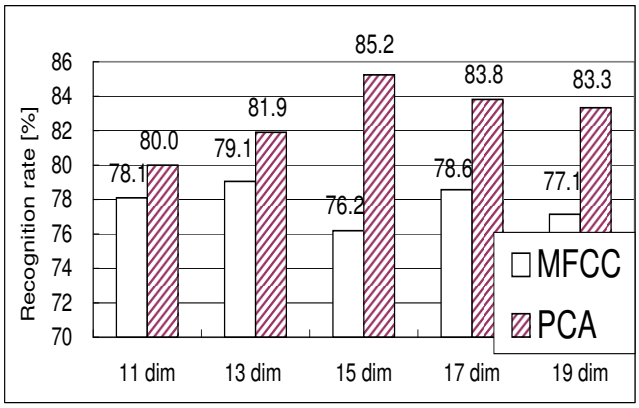


図 10　PCA 手法による認識実験結果（1 回目）

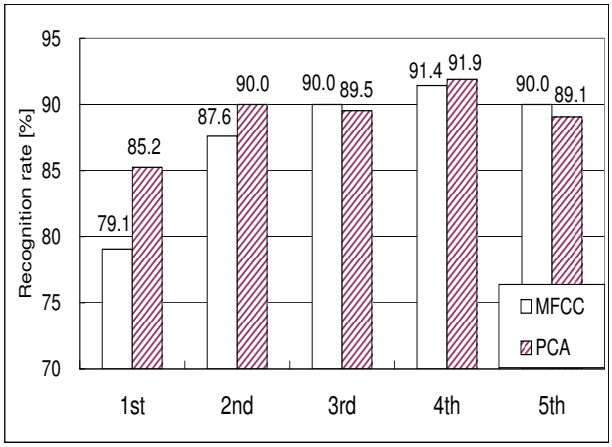


図 11　PCA 手法による発話回数ごとの認識率 (15 次元)

は 11,13,15,17,19 個として実験を行った．図 10 に 1 回目発話における結果を示す．

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．主成分が 15 個の場合の発話回数毎の認識率を図 11 に示す．

4.4　メタモデルと音響モデルの統合

単語認識時の 3Best 単語に対して，メタモデル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，メタモデルを作成した．メタモデルに対する重み α を 0.0 〜 0.3 として実験を行った．図 12 に 1 回目発話における結果を示す．

メタモデルとの統合によって，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 メタモデル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．

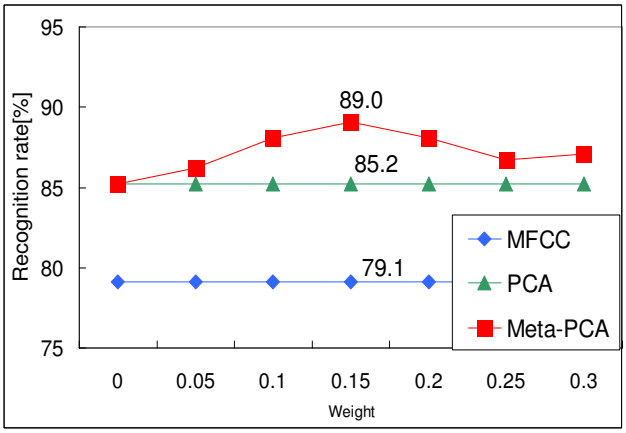


図 12　提案手法による発話回数ごとの認識率 (1 回目)

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 メタモデルの統合を検討した．PCA による特徴量抽出だけでなく，音響モデルと メタモデルの統合によって 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．

文　献

[1] 佐川浩彦，酒匂裕，大平栄二，崎山朝子，阿部正博, “ 圧縮連続 DP 照合を用いた手話認識方式, ”電子情報通信学会論文誌, Vol.J77-D2，No.4，pp. 753-763，1994.

[2] 鈴木悠司，平岩裕康，竹内義則，松本哲也，工藤博章，大西昇, “ 視覚障害者めための環境内の文字情報抽出システム, ”電子情報通信学会技術研究報告，WIT2003-314，pp. 13-18，2003.

[3] 藪謙一郎，伊福部達，青村茂, “ 発話障害者支援のための音声合成器の基礎的設計, ”電子情報通信学会技術研究報告, SP2006-321, pp. 59-64, 2006.

[4] 鮫島充，李晃伸，猿渡洋，鹿野清宏, “ 子供音声認識のための音響モデルの構築および適応手法の評価, ”電子情報通信学会技術研究報告，SP2004-114，pp. 109-114，2004.

[5] 中村圭吾, 田村直良, 鹿野清宏, “ 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, ”日本音響学会講演論文集, 3-7-4，pp. 109–110，2005.

[6] 内閣府“ 平成 18 年　障害者白書, ” http://www8.cao.go.jp/shougai/

[7] 厚生労働省“ 平成 13 年　身体障害児・者実態調査結果, ” http://www.mhlw.go.jp/houdou/2002/08/h0808-2.html

[8] 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.

[9] “ 脳性マヒについて ”, http://www.geocities.co.jp/SweetHome/6954/nouseimahi.html

[10] H. Matsumasa and T. Takiguchi and Y. Ariki and I. LI and T. Nakabayashi,“ PCA-Based Feature Extraction for Fluctuation in Speaking Style of Articulation Disorders,” INTERSPEECH-2007, pp. 1150–1153, 2007.

[11] S. Cox and S. Dasmahapatra, “ High-Level Approaches to Confidence Estimation in Speech Recognition,” IEEE Trans. on SAP, vol. 10, No. 7, pp. 460-471, 2002.

[12] O. C. Morales and S. Cox,“ Modelling Confusion Matrices to Improve Speech Recognition Accuracy, with an Application to Dysarthric Speech,” INTERSPEECH-2007, pp. 1565–1568, 2007.

[13] 滝口哲也, 有木康雄, “ Kernel PCA を用いた残響下におけるロバスト特徴量抽出の検討, ”情報処理学会論文誌，Vol.47, No.6，pp. 1767-1773，2006.

[14] S. Young et. al.,“ The HTK Book, ” Entropic Labs and Cambridge University, 1995-2002.