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.

Buried Markov Modelを用いた構音障害者の音声認識の検討

宮本　千琴†　　駒井　祐人†　　滝口　哲也†　　有木　康雄†　　李　　義昭††

† 神戸大学工学研究科　〒 657–8501 兵庫県神戸市灘区六甲台 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
E-mail: †{miyamoto,komai}@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@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，アテトーゼ型脳性麻痺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，筋肉の緊張のため発話が不安定になりやすい．これに対し，本研究では，時間変化による依存関係を考慮することで，不安定な発話に対する音声認識精度の改善を試みる．従来用いられている HMM による音声認識は，はっきりと発話された音声に対しては高い精度で認識を行うことができるが，複雑な事象を表現するには適しておらず，雑音を含む音声や，連続的に発話された音声を認識する際には，精度が著しく低下する．この問題に対し，過去の観測と現在の観測の間の依存関係を表現できる確率モデルである Buried Markov Model を用いた音声認識モデルが，J. Bilmes によって提案された．本研究では，構音障害者の音声認識の実現に向けて，この Buried Markov Model を用いて時間的依存関係を考慮し，音声認識精度の向上を目指す.

キーワード　構音障害，Buried Markov Model

A Study on Dysarthric Speech Recognition using Buried Markov Model

Chikoto MIYAMOTO†, Yuto KOMAI†, Tetsuya TAKIGUCHI†, Yasuo ARIKI†, and Ichao LI††

† 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
E-mail: †{miyamoto,komai}@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@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 speech tends to become unstable due to strain on speech-related muscles, and that causes degradation of speech recognition. Therefore, we consider temporal dependence to solve this problem. Though HMM makes it possible to recognize clear utterance with high accuracy, the speech including the noise or the continuous utterance causes degradation of speech recognition. To solve this problem, J. Bilmes proposed buried Markov model which contains the conditional independence between the observation nodes. In this paper, we perform phone recognition experiments using buried Markov model.

Key words　articulation disorders, Buried Markov Model

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].

音声認識技術は近年，飛躍的に進歩し，様々な環境や場面での利用が期待されている．例えばカーナビゲーションの操作や会議音声の議事録化など様々な分野に応用されている．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おり，文献 [6] では，特徴量抽出や音響モデルの構築を行っているが，言語障害者などの障害者を対象としている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 4 万 2000 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 [7].

言語障害の原因の一つとして，脳性麻痺が考えられる．脳性麻痺の定義として，1968 年の厚生労働省脳性麻痺研究班は「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.」としている．

脳性麻痺とは，筋肉の動きをつかさどる脳の部分が受けた損傷が原因で筋肉の制御ができなくなり，けいれんや麻痺，そのほかの神経障害が起こる症状のことである．出生前，出生時，出生直後の脳への酸素供給，出生前の胎内感染，妊娠中毒症，分娩時の外傷，仮死状態，未熟出生，出生後の脳を覆う組織の炎症や外傷性損傷などが原因として考えられる．

脳性麻痺は，脳の損傷部分によって主に痙直型（大脳皮質），アテトーゼ型（中脳もしくは脳基底核），失調型（小脳），混合型（脳の広範囲）に分類される．痙直型は正常な筋の伸張反射が過度になる，アテトーゼ型はアテトーゼと呼ばれる筋肉の不随意運動を伴う，失調型は協調運動の障害，混合型はそれぞれの症状が混合して現れる，というような症状が見られる．

本稿では，アテトーゼ型の脳性麻痺による構音障害者を対象としている．アテトーゼ型は，脳性麻痺患者の約 20 % に発生する．筋肉の随意運動や姿勢の調整を行っている大脳基底核（大脳皮質，視床や脳幹を結び付けている神経核の集まり）に損傷を受けたことにより，筋肉が不随に動き，正常に制御できないアテトーゼと呼ばれる症状が見られる．とくに意図的な動作を行う場合や，緊張状態にある時に見られ，この運動障害の一つとして，正しく構音できない場合がある．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].

現在の音声認識システムでは，Hidden Markov Model (HMM) が音響モデルとして広く用いられている．HMM は隠れ状態と観測のみからなる単純な構造の確率モデルである．このような単純な構造は，“状態は 1 フレーム前の状態のみによって決まる” という仮定と，“観測はフレームごとに独立であり各フレームの状態のみによって決まる” という仮定の，2 つの独立仮定の上に成り立っている．この 2 つの独立仮定により音声の生成構造が単純化されているため，扱いやすいモデルであり，はっきりと発話された音声に対しては高い精度で認識を行うことができるが，複雑な事象を表現するには適しておらず，雑音を含む音声や，連続的に発話された音声を認識する際には，精度が著しく低下する．

この問題に対し，HMM の独立性の仮定を緩和することでその表現能力を高めたモデルである Buried Markov Model (BMM) [10] が，J. Bilmes によって提案された．HMM の隠れ状態と観測の構造に加え，各フレームの観測系列間の依存関係をモデルに埋め込んでいるため，より複雑な事象を表現することができる．この BMM の依存関係は従来，データを用いて相互情報量を計算し，それを用いて学習することで求められている．相互情報量を計算するために，初期学習した HMM のパラメータが必要となる．この方法では HMM のモデリング精度によって得られる依存関係が影響を受ける可能性がある．これを改善するために，HMM の初期学習なしでパラメトリックに BMM の構造を学習するための手法として，独立性検定を用いた構造学習法が提案されている [11]．本稿では，構音障害者の音声認識に BMM を適応し，不安定な発話に対する音声認識精度の改善を検討する．また，BMM の構造を学習する手法として独立性検定を導入する．

2. Buried Markov Model

2.1 概　　要

Buried Markov Model は HMM の各フレームの観測系列間に依存関係を加えた図 1 のようなグラフィカルモデルによって表現される．Hidden Markov Model という名前は，実際に観測できるのは出力信号系列だけであり，マルコフ過程に従う隠れ状態系列は観測でないことに由来するが，Buried Markov Model は，この隠れ状態のマルコフ過程が出力系列間の依存関係によって HMM よりさらに隠れた状態 (buried) になっていることから，そう呼ばれている．

また，BMM の同時分布 $Pr(x_{1:T})$ は式 (1) のように表される．

$$Pr(x_{1:T}) = \sum_{q1:T} \prod_{t} Pr(x_t|z(q_t), q_t) Pr(q_t|q_{t-1}) \qquad (1)$$

T は時間長，x_t は t 番目のフレームにおける出力，q_t は観

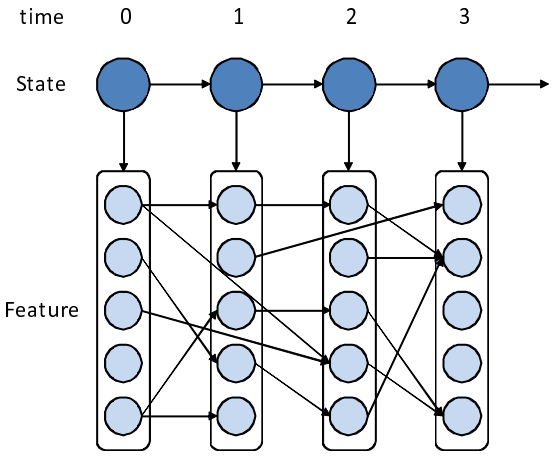

図 1　Buried Markov Model の構造

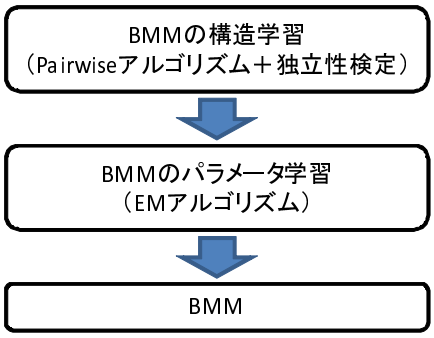

図 2　Buried Markov Model の学習の流れ

測に対応する状態を表す．z はフレーム t における状態 q_t によって一意に決まる出力間のエッジの集合を決定する関数である．BMM の観測系列間の依存関係の構造は，参照するフレームの隠れ状態によって決まるため，各隠れ状態ごとに学習が行われ，また，別のフレームからの情報が埋め込まれたモデルを構築することができる． 依存関係の構造は，各状態ごとに各状態における参照フレームの出力と過去や未来のフレームの出力の間の相関関係を調べることで学習される．BMM の簡単な学習の流れを図 2 に示す．

2.2　構造学習方法

BMM の構造学習には相互情報量を用いた Pairwise アルゴリズムが用いられる．Pairwise アルゴリズムはターゲットノード X に対し，相互情報量が次の条件を満たすノード Z を親ノード集合 $\mathbf{Z}$ に加える方法である．q はターゲットノードのフレームの状態，$\delta_{1\cdots3}$ はそれぞれの閾値を表す．

$$I(X, Z \mid q) > \delta_1 \tag{2}$$

$$I(X, Z) < \delta_2 \tag{3}$$

$$I(Z, Z_i) < \delta_3 I(X, Z) \quad Z_i \in \mathbf{Z} \tag{4}$$

式 (2) は相関の高いノードを親とするという条件である．この条件によって，HMM では独立と仮定していた相関の強いノードを特定できる．しかし，相関の高いノードを親としても，他の状態においても同様に相関が高ければ，状態 q を推定するために有益な情報を持ったノードとはいえない．そこで式 (3) の 2 つの変数間の相互情報量の状態平均が小さいノードを選択する条件を導入し，候補ノードを状態 q において X との依存性が特異な値をとるノードに限定する．式 (4) は冗長性の検定を行うための条件である．Z と $\mathbf{Z}$ のノードとの相関が強ければ，$\mathbf{Z}$ のノードから得られる情報と同様の情報しか得られないため，Z を親の候補から除外する．

この手法で用いる相互情報量を計算するために，初期学習した HMM のパラメータが必要となる．この方法では HMM のモデリング精度によって得られる依存関係が影響を受ける可能性がある．これを改善するために，HMM の初期学習なしでパラメトリックに BMM の構造を学習するための手法として，独立性検定を用いた構造学習法が提案されている [11]．本稿では，この構造学習法を用いる．詳しくは次章で述べる．

3.　独立性検定

3.1　スピアマンの順位相関係数

スピアマンの順位相関係数 (Spearman's rank correlation coefficient) は 2 つの変数の順位の間の相関の強さを表す指標であり，次式で表わされる．

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{(n-1)n(n+1)} \tag{5}$$

長さ n の 2 つのデータ系列 X，Y のあるインデックス i におけるデータ X_i ,Y_i の順位の差を D_i とする．ρ はデータの順位のピアソンの相関係数を計算したものである．実際に Pairwise アルゴリズムの中で用いる場合，ρ の絶対値が依存性として用いられる．

3.2　ケンドールの順位相関係数

ケンドールの順位相関係数はスピアマンの順位相関係数と同様の 2 つの変数の順位の相関の強さを表す指標であり，次式で表わされる．

$$\tau = \frac{P - Q}{\frac{1}{2}n(n-1)} \tag{6}$$

長さ n の対応したデータ系列 X，Y から 2 つのデータを選ぶ．選んだデータのインデックスを a，b とすると，P は「$X_a < X_b$ かつ $Y_a < Y_b$」 または「$X_a > X_b$ かつ $Y_a > Y_b$」の条件に当てはまる組み合わせの数，Q は「$X_a < X_b$ かつ $Y_a > Y_b$」 または「$X_a > X_b$ かつ $Y_a < Y_b$」の条件に当てはまる組み合わせの数である．これはすなわち，P が 2 つのデータ間の正の相関性を，Q が負の相関性を表しており，P が大きくなれば正の相関が強くなり τ は 1 に近づき，Q が大きくなれば負の相関が強くなり τ は -1 に近づく．P と Q が同じであれば τ は 0 となりデータ間の相関はないと判断できる．

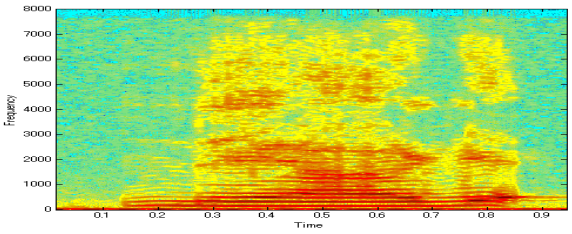

図 3　構音障害者のスペクトログラム例/n e a g e/

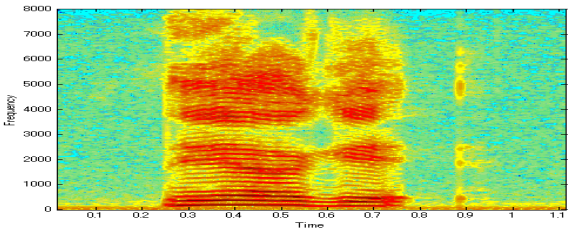

図 4　健常者のスペクトログラム例/n e a g e/

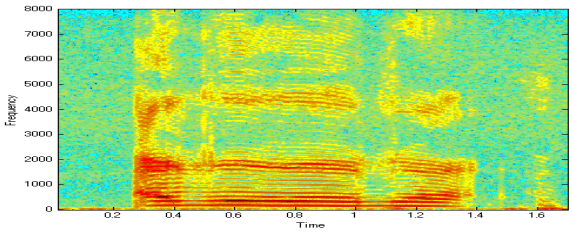

図 5　構音障害者のスペクトログラム例/a k e g a t a/

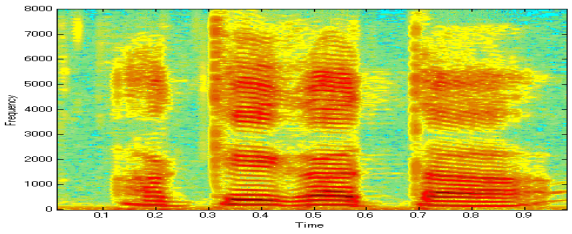

図 6　健常者のスペクトログラム例/a k e g a t a/

ケンドールの順位相関係数は，線形・非線形に関わらず相関関係を判別することができるが，複雑な相関関係を判別することができず，単調関数性を判別する指標となる．実際に Pairwise アルゴリズムの中で用いる場合，スピアマンの順位相関係数同様 τ の絶対値が依存性として用いられる．

4. 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）を使用し，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，フレーム窓長は 25 msec，フレーム周期は 10 msec である．図 3 に構音障害者 (話者 A)，図 4 に同じ発話内容の健常者のスペクトログラム例，図 5 に別の構音障害者 (話者 B)，図 6 に同じ発話内容の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．このデータを用いて，GMTK [12] によりモデルを作成する．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 BMM を作成し，連続音素認識を行う．これを各発話に対して行い平均する．状態には 43 音素 3 状態，各状態 1 混合を用いた．Pairwise アルゴリズムで用いる閾値は全て 0.5 に設定した．構造学習にはケンドールの順位相関係数を用いた．特徴量には 12 次 MFCC ＋Δ MFCC の 24 次元を用いる．以上の条件で，構音障害者の音声認識における BMM の有効性を確認する．

4.2 BMM を用いた認識実験結果

まず，探索過去フレーム数を 5 に固定し，親ノード数の上限を 1 から 7 まで変化させて，モデル構築を行い，その精度を評価した．認識結果を図 7 に示す．

親ノードの上限数が 1 の時 53.3% となり，最も良い結果となった．親ノードの上限数を増やしていくと，認識精度が低下していることが結果よりわかる．これは，今回の実験データを用いた BMM は，複雑な構造よりも簡単な構造が適していたからだと考えられる．

次に，親ノード数の上限をそれぞれ 1，3，5 に固定し，親ノードの探索フレームを過去 7 フレームまで変化させて，モデル構築を行い，その精度を評価した．認識結果を図 8，図 9，図 10 にそれぞれ示す．

親ノード数の上限を 1 とした場合，探索過去フレーム数にかかわらず約 53 %の認識精度を得た．親ノード数の上限を 3，5 とした場合，探索過去フレームを増やしていくと一旦認識精度が低下し，更に増やしてくと認識精度が向上した．これらの結果より，親ノード数が多くなければ，探索過去フレーム数にあまり左右されずに安定して認識を行えるが，親ノード数を増やすと，できるだけ多くの過去を参照し探索を行った方が，高い認識精度が得られると考えられる．

4.3 構造学習に関する考察

本稿では，BMM の構造学習にケンドールの順位相関係数を用いたが，構造学習時に音声データの音素ラベル情報とその時間情報が必要となる．しかし，構音障害者の発話ははっきりしていないため，音素間の境界が曖昧であることや発声されていない子音があるという問題がある．例えば，図 3 と図 4 を比較した場合，構音障害者音声の子音がほとんど表れていないことがわかる．“n e a g e” と発声しているが，実際には “e a e” という発話であるとみなす方が良い可能性がある．また，図 5 と図 6 を比較した場合，健常者に比べて発話障害者は音素間の境界が曖昧ではっきりしていないことがわ

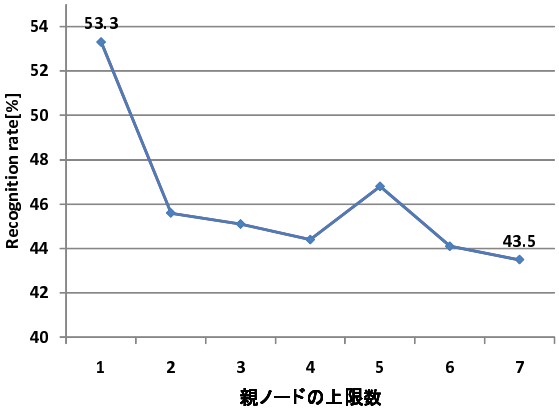

図 7　探索過去フレーム数を 5 としたときの認識結果

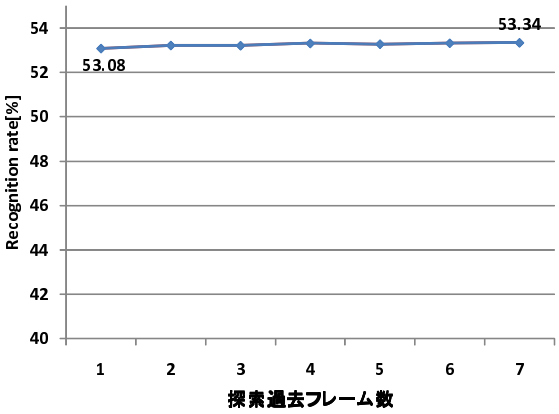

図 8　親ノードの上限を 1 としたときの認識結果

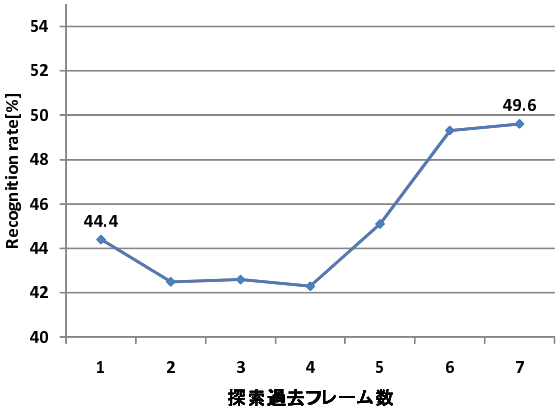

図 9　親ノードの上限を 3 としたときの認識結果

かる．とくにｋｅ → ｇａと発話する際の境界がほとんど表れていないため，正確な時間情報が得られていない可能性が考えられる．今後，学習ラベルを “ｅａｅ” とすることや，音素間の境界の曖昧性を考慮した音素ラベル情報とその時間情報の検討が必要であると考えられる．

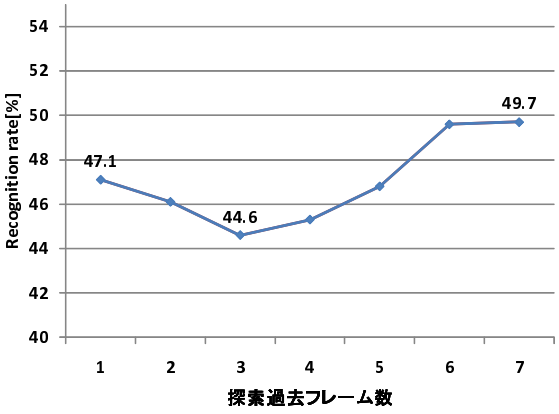

図 10　親ノードの上限を 5 としたときの認識結果

5. おわりに

本稿では，発話が不安定な構音障害者の音声認識精度を改善するために，観測ノード間の時間的な依存関係を記述できる BMM を用いた音声認識手法を検討した．また，音素間の境界が曖昧であることや発声されていない子音があるという問題に対し，これらの問題を考慮した音素ラベル情報とその時間情報が必要であることが分かった．

今後は，スピアマンの順位相関係数など他の相関基準を用いた構築法との比較，構音障害者特有の特徴量の検討や、複数話者に対して有効性を確認していく予定である．また，音声特徴だけでなく画像特徴も共に用いた構築アルゴリズムの検討を行う．

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，“視覚障害者の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，濱篤志，伊福部達，青村茂，“発話障害者支援のための音声合成器—その研究アプローチと設計概念—，” 電子情報通信学会技術研究報告， SP2006-164, pp. 25–30, 2007.

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

[7] 内閣府，“平成 20 年版障害者白書，” http://www8.cao.go.jp/shougai/

[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] Mark H. Beers, 福島雅典, “メルクマニュアル医学百科 最新家庭版,” 日経 BP 社, 2004.
[10] J.A. Bilmes, “Buried Markov models: a graphical modeling approach to automatic speech recognition,” Computer Speech and Language, Volume 17, Issues 2-3, 213-231, 2003.
[11] 山本隆之, 滝口哲也, 有木康雄, “Buried Markov Model を用いた音声認識モデルの構築法の検討,” 情処研報, 2009-SLP-79, No.21, pp.1-6, 2009.
[12] J.A. Bilmes and G. Zweig, “The Graphical Models Toolkit: An open source software system for speech and time-series processing,” ICASSP, 3916-3919, 2002.