社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2004－SLP－54（34）
2004／12／21

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

鮫島　充†　李　晃伸†　猿渡　洋†　鹿野　清宏†

† 奈良先端科学技術大学院大学 情報科学研究科
〒630-0101 奈良県生駒市高山町 8916-5
E-mail: †{mituru-s,ri,sawatari,shikano}@is.naist.jp

あらまし　子供音声は自由な発話形式のものが多く，既存の読み上げ音声コーパスでは対応が難しい．また一般に子供音声の収録には多大な労力やコストがかかるため，整った音声コーパスを作成することも難しい．本研究では，音声情報案内システムにおいて自動収集した子供の音声に基づく音響モデルの作成および認識性能の評価を行った．また，自動収集した子供音声に対する自動話者クラスタリングを提案し，それを用いた十分統計量に基づく教師なし話者適応を行った．収集した子供音声から作成した音響モデルにより，年齢層ごとに幼児 53.6%，低学年子供 82.1%，高学年子供 77.6%の認識性能が得られ，既存のモデルに比べ大幅に認識性能の改善が得られた．また提案した自動話者クラスタリングとそれを用いた十分統計量に基づく教師なし話者適応の結果，子供用不特定話者モデルに対して一定の認識性能の改善が得られ，年齢層ごとの MAP 適応モデルよりも高い認識性能が得られたことを示す．
キーワード　子供音声，音響モデル，十分統計量に基づく教師なし話者適応，自動話者クラスタリング

Evaluation of Acoustic Models and Adaptation Methods Based on Collection of Spontaneous Speech for Child Speech Recognition

Mitsuru SAMEJIMA†, Akinobu LEE†, Hiroshi SARUWATARI†, and Kiyohiro SHIKANO†

† Nara Institute of Science and Technology
630-0192 Takayama-Cho, Ikoma, Nara, JAPAN
E-mail: †{mituru-s,ri,sawatari,shikano}@is.naist.jp

Abstract　Acoustic modeling in current speech recognition system requires a large amount of speech database that are correctly uttered and transcribed. However, this methodology can not be easily applicable for the recognition of child speech. Children's utterances are usually not well-articulated, spontaneous. Controlling them to read sentences precisely for collection of database is difficult, and also the resulting utterances may be far from spontaneous speech. In this research, we evaluate the acoustic models and adaptation methods for child speech recognition based on a natural child speech database automatically collected through actual spoken dialogue system "Takemaru-kun". Also we propose a speaker clustering method to perform unsupervised speaker adaptation based on HMM Sufficient Statistics (HMM-SS) on automatically collected database where no user tag is available. The acoustic HMM trained by 59,966 spontaneous child speech achieved the word accuracy of 53.6% for the infants (pre-school children), 82.1% for elementary schoolers, 77.6% for junior-high schoolers, which substancially outperforms an adult female model and a child model trained by read speech. Furthermore, experiment of the proposed speaker adaptation method based on automatic speaker clustering and HMM-SS resulted in a slight improvement of recognition accuracy, that was better than age-class dependent MAP adaptation.
Key words　Children speech, Acoustic model, Speaker adaptation, HMM Safficient Statistics, Automatic speaker clustering

1. まえがき

近年，音声認識技術の発展により，カーナビゲーションや音声ポータルなど音声インターフェースを備えたシステムが様々な分野へ応用されているが，多くは成人の音声を主な対象としており，子供の音声に対応したものが少ない．しかし，小学生や中学生が情報端末などの機器に触れる機会が増えており，子供の声をも認識できることが望まれている．

従来の子供音声を対象とした研究では，子供の音声を収集して音響モデルを再学習することで認識精度を向上できることが報告されている [1] [2] が，整った子供音声を収録するには多大なコストや労力がかかる．また実際の子供音声は成人に比べてより自由な発話であることが多く，スムーズに発話することができずに言い直しや不明瞭な発話が多い．そのため読み上げの子供音声コーパスから学習した音響モデルではミスマッチが起こりやすい．

本研究では，実環境音声情報案内システム「たけまるくん」[3] より，システムに対する子供の自然な音声を自動収集し，それに基づく音響モデルの作成と音声認識実験を行った．子供の年齢層ごとの音声認識性能の傾向を示すとともに，子供音声に対して更なる音声認識性能の改善を目指し，収集した子供音声に対する自動話者クラスタリングを提案し，それを用いた十分統計量に基づく教師なし話者適応 [4] を行う．適応モデルを用いた認識実験より，全学習データより作成した音響モデルに対して一定の認識性能の改善が得られることを示す．

以下，2節では，収集した子供音声データについて述べる．3節では2節の子供音声を用いて音響モデルを作成し，作成した音響モデルで評価実験を行った結果を示す．4，5節では子供音声に対する適応手法について述べ，6節では，適応モデルで評価実験を行った結果を示す．最後に，7節において本報告のまとめを述べる．

2. 実環境音声情報案内システムにおけるデータ収集

2.1 音声情報案内システム「たけまるくん」

音声情報案内システム「たけまるくん」（図1）は，「生駒市北コミュニティーセンター ISTA はばたき」の館内で稼動している [3]．コミュニティーセンター内の施設案内，生駒市の観光案内，周辺の情報案内などの各種案内を行うシステムである．ユーザの発話による質問に対し，合成音声とアニメーションを用いて，ソフトウェアエージェントが応答する一問一答形式の対話機能を持つ．このシステムを用いて利用者の音声を収集し，データベースとして整備している．

図1 音声情報案内システム「たけまるくん」
Fig. 1 Speech-oriented information system "Takemaru-kun".

表1 収集した子供音声データの内訳
Table 1 Breakdown of collected children's speech data.

	学習用データ	評価用データ
幼児	9982	400
低学年子供	38111	400
高学年子供	11873	400
合計	59966	1200

今何時？
図書館はどこですか？
コミュニティーセンターは何処？
たけまるくんは何が好きですか？
今日のニュースは？
こんにちは。

図2 評価用データの例
Fig. 2 Example sentences of evaluation data.

2.2 子供音声の収録状況

センターが開所した2002年11月6日から現在までの開館時間中に，利用者の音声を収録している．ただし，これらの中には雑音のみのデータや発話が不明瞭で聞き取れないものも含まれる．そこで，発話データの整備として，人手による雑音データのラベル付け，発話内容のテキストへの書き起こし，収録音声からの主観による発話者の性別及び年齢のラベル付けを行っている．

このうち発話内容が聞き取れる発話データを学習用データ，評価用データとして用いる．学習用データ，評価用データについては幼児，低学年子供 (小学校3年生くらいまで)，高学年子供 (中学生くらいまで) に分類した．各ラベルは書き起こし者の主観に従って音声のみから判断されている．各年齢層の内訳を表1に示す．

評価用データについては，書き起こしテキストをキーにしてソートを行い，同じ発話内容のデータを間引いて除いた後の発話から，その発話内容の出現頻度に基づく上位400個のデータを各年齢層ごとに選んだ．評価用データの例を図2に示す．なお，評価用データには，学習用

表 2　たけまる子供用不特定話者モデルの作成条件
Table 2　Conditions used in Acoustic Modeling.

初期モデル	JNAS 性別非依存モデル (PTM [6] 2000 状態 64 混合)
学習データ	音声情報案内システムの子供音声 59966 個
標本化/量子化	16kHz/16bit
窓シフト長	10msec
特徴量	MFCC(12 次元), Δ MFCC, Δパワー

データは含んでいない.

3. 子供音響モデルの作成と認識実験

2 節で収集した子供データを用いて音響モデルを作成し，音声認識実験を行った.

3.1 たけまる子供用不特定話者モデルの作成

たけまる子供用不特定話者モデルの作成条件を表 2 に示す. たけまる子供用不特定話者モデルは JNAS 性別非依存モデルを初期モデルとし，子供の年齢層を区別しない全学習データ 59966 個を用いて EM 学習を繰り返し行い作成した.

3.2 実験条件

作成したたけまる子供用不特定話者モデルの認識性能を評価するために，音声認識実験を行った. 音声認識実験条件を表 3 に示す.

比較対象として，既存モデルである JNAS 性別非依存モデルおよび CSRC [5] 子供モデルを用いる. JNAS 性別非依存モデルは，JNAS データベース (成人男女の音声データベース) から構築した性別非依存 PTM triphone HMM モデルである. CSRC 子供は「連続音声認識コンソーシアム (CSRC)」の 2002 年度版ソフトウェアに含まれる，子供の読み上げ音声より作成された PTM triphone HMM である.

評価用データは表 1 で示した各年齢層ごとの 400 文を用い，言語モデルは文法適用 3-gram を使用した.

言語モデルには，Web ページテキストと，システムを想定して人手で収集し質問テキストから学習した言語モデルをベースとし，文法の二単語間の接続制約を用いた N-gram 確率の強化を行っている [7]. 本実験では，この言語モデルと収集した子供音声の書き起こしテキストより作成した言語モデルを用いて重み 0.8 で融合し，子供に適応した言語モデルとしている.

3.3 実験結果

認識実験で得られた各年齢層の評価用データに対する単語認識精度 (Word Accuracy) を図 3 に示す. 図 3 より，各年齢層とも収集した発話データで学習したたけまる子供用不特定話者モデルが，既存のモデルに比べて高い認識性能を示していることがわかる.

本研究で作成したたけまる子供用不特定話者モデルは，

表 3　音声認識実験条件
Table 3　Conditioin of speech recognition experiment using SI model.

認識エンジン	Julius ver.3.4.2
タスク	音声情報案内システムの子供音声
言語モデル	文法適用 3-gram(単語数 40k)
音響モデル	1) JNAS 性別非依存モデル 2) CSRC 子供モデル 3) たけまる子供用不特定話者モデル

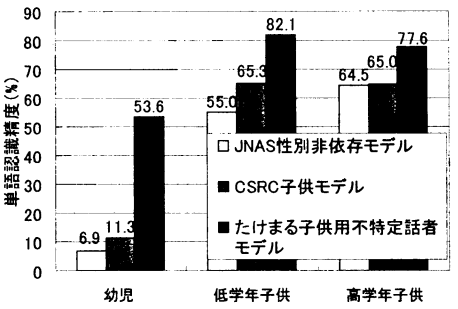

図 3　年齢層ごとの子供音声認識結果
Fig. 3　Results of children's speech recognition using SI model.

比較対象モデルである JNAS 性別非依存モデルを初期モデルとしている. これより，成人モデルに対して子供の発話データを大量に学習させることで，各年齢層において認識性能が大きく改善できたことが言える.

また CSRC 子供モデルについては，子供の音声を学習しているが，読み上げ音声と自由な発話の音声という発話の性質の違いがあるため，両者の認識性能には，各年齢層において大きな差があることがわかる.

次に，各年齢層ごとの傾向をみると，低学年子供，高学年子供に対してはたけまる子供用不特定話者モデルでそれぞれ 82.1%，77.6%と比較的高い認識性能が得られた. 両者とも既存のモデルに比べ認識性能が改善されたが，高学年子供が低学年子供よりも認識性能が悪いことがわかる. これは高学年子供になると，声変わりや男女の性別差の影響が出始めるため，たけまる子供用不特定話者モデルでカバーされない評価話者がいたことが考えられる.

一方で，幼児の子供に対しては，たけまる子供用不特定話者モデルを用いることで 53.6%の認識性能が得られ，既存のモデルに比べては格段に認識性能が改善されたが，他の年齢層と比べては特に認識性能が劣ることがわかった. これは幼児の音声が他の年齢層に比べて特に舌ったらずな発話が多く，他の年齢層の発話様式とは大きく異なるためではないかと考えられる.

4. 十分統計量に基づく教師なし話者適応と自動話者クラスタリング

子供は声の変動が激しく，話者によって音声の特徴が大きく異なるため，話者に適応したモデルを構築することは有効であると考えられる．

子供への適応を考えた場合，従来の教師あり適応のように発声者の数十文の発話データが必要な適応法は実行が難しい．

本研究では，話者の負担が少ない適応法として，任意の一発声文から教師なし話者適応が可能な，十分統計量に基づく教師なし話者適応 [4] を用いる．この適応法は対象話者の特徴に近い既存の話者データベースを用いて適応するため，子供に対しても負担なく適応できる．しかし，適応の際には話者ごとに大量の学習データが必要となるが，音声情報案内システムから自動的に収集した子供音声においては話者を特定できないため，話者ごとにデータベースを整えることが難しい．

そこで収集した子供音声に対する自動話者クラスタリングを行うことで，音響的に近い発話データ集合を形成して十分統計量に基づく教師なし話者適応を行うことを提案する．以下，手法の詳細を述べる．

4.1 十分統計量に基づく教師なし話者適応

十分統計量とはデータベースを表すのに十分な統計量のことであり，特に HMM においては平均，分散，EM カウントのことを示す．この十分統計量をあらかじめデータベースから話者ごとに作成することで，高速に音響モデルの学習を行うことができる．話者適応の際には，任意の一発声からデータベース上の近傍話者を選択し，その十分統計量から音響モデルを再構築する．上記の処理は，発声話者に音響的に近い話者データを用いて，適応元モデルから 1 回学習して音響モデルを作成することと同等である．

4.2 自動話者クラスタリング

収集した子供音声から十分統計量に基づく教師なし話者適応を行うために音響的に近い発話データを形成して，一人分の話者データとみなす自動話者クラスタリングを提案する．概念図を図 4 に示す．提案するクラスタリングでは K-means 法 [8] を用い，音韻バランスが整っていない発話間での距離計算を工夫している．

まずクラスタリングに用いる特徴量として，各発話データの母音に関する MFCC12 次元の平均ベクトルを用いる．さらに，どの発話データにも含まれており，一般に雑音に対して頑健な母音情報を用いる．これにより，音韻バランスが整っていない発話間でも共通の情報を用いることができ，かつ比較的雑音の影響を受けずに発話間距離を計算できる．

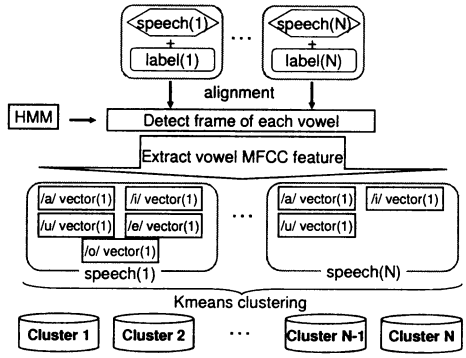

図 4 自動話者クラスタリング
Fig. 4 Automatic speaker clustering.

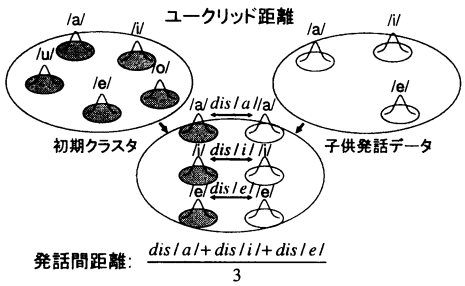

図 5 発話間の距離計算
Fig. 5 Distance measure.

母音の平均ベクトルを抽出する際には，収集した子供音声データから，発話 speech(x) をその書き起こしラベル label(x) が既知の条件でたけまる子供用不特定話者音響モデルに与えることで発声内容を表す音素系列のアライメントを取り，発話中の各母音フレームを検出する．そして得られたフレーム情報を用いて，各母音の平均ベクトルを抽出する．

発話間の距離計算は，発話間で共通する母音の MFCC 平均ベクトルを用いたユークリッド距離で算出する．発話間の距離計算の様子を図 5 に示す．ここで初期クラスタにはすべての母音が揃っている発話文を選択し，初期クラスタと各学習発話との距離は，共通する母音同士の距離の和を計算する．これにより，各学習発話データがどの初期クラスタに対しても必ず共通する母音が存在し，音韻バランスが整っていない発話間でも一定の距離尺度によるクラスタリングが可能になる．

上記の距離尺度を用いて，以下の手順でクラスタリングを行う．

step1: 初期クラスタ (クラスタの代表点) を N 個設定する．この時，音韻のバランスを考慮して，全ての母音が揃っている学習発話を初期クラスタに設定する．

step2: 各学習発話ごとに最近傍の初期クラスタを算出

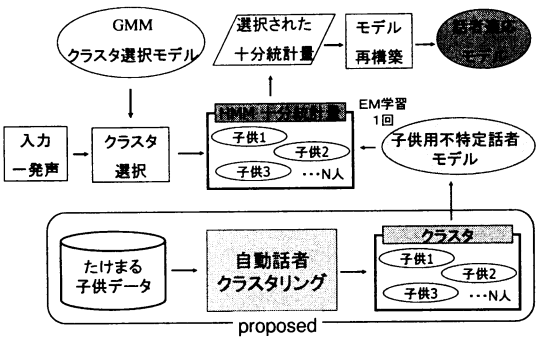

図 6 自動話者クラスタリングを用いた十分統計量に基づく教師なし話者適応

Fig. 6 Unsupervised speaker adaptation based on HMM Sufficient Statistics using automatic speaker clustering.

し，その初期クラスタに割り当てる．

step3: 各学習発話を初期クラスタに割り当てた後，クラスタごとに，属する学習発話のセントロイドを各母音ごとで求め，それを新しい初期クラスタとする．ここで初期クラスタには常に全ての母音の平均ベクトルが内在するようにしておく．

step4: すべての初期クラスタが変化しなくなるまで step2 から step4 を繰り返す．

4.3 提案するアルゴリズム

収集した子供音声に対する話者適応として，4.2 の自動話者クラスタリングを用いた十分統計量に基づく教師なし話者適応を提案する．提案法の概念図を図 6 に示す．

具体的な適応手順は，以下の通りである．

step1: 子供音声の全学習データに対して自動話者クラスタリングを行い，N 個のクラスタを作成する．

step2: step1 で作成した各クラスタごとに十分統計量，GMM クラスタ選択モデルを算出し保存する．

step3: GMM クラスタ選択モデルを用いて発声話者に音響的特徴が近いクラスタを K 個選択する．

step4: 選択された K 個のクラスタの十分統計量を用いて話者適応モデルを再構築する．

5. MAP 適応による年齢層別モデル構築

提案した話者適応アルゴリズムとの比較のために，各年齢層ごとの学習データから年齢層に適応した音響モデルを作成する．モデル構築の過程を図 7 に示す．適応元モデルには 3 節で作成したたけまる子供用不特定話者モデルを用い，表 1 で示した幼児，低学年子供，高学年子供の各学習用データを用いて MAP 適応を行う．作成したモデルは，それぞれ幼児適応モデル，低学年子供適応モデル，高学年適応モデルとする．

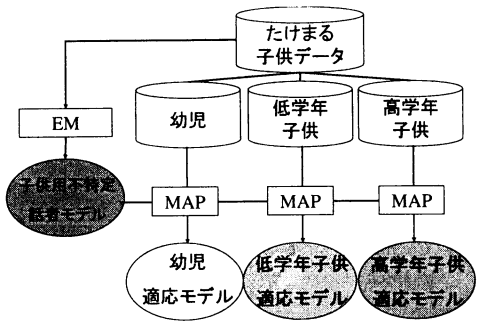

図 7 年齢層別音響モデル

Fig. 7 Acoustic model for each age.

6. 認識実験

提案する話者適応アルゴリズムにより話者適応を年齢層ごとの MAP 適応と比較する．

6.1 話者適応モデルの作成

提案するアルゴリズムにより話者適応を行った音響モデルの作成条件を以下に示す．まず自動話者クラスタリングにおける実験条件を表 4 に示す．自動話者クラスタリングでは，年齢層を区別しない全学習データ 59966 個を用い，200 個のクラスタを作成した．十分統計量に基づく教師なし話者適応モデルの構築に関しては，自動話者クラスタリングにより作成された 200 個の子供クラスタから各年齢層の各評価データに対し音響的特徴が近傍なクラスタを複数選択し，その十分統計量を用いて話者適応モデルを構築した．ここで適応元モデルには，3 節で作成したたけまる子供用不特定話者モデルを使用する．また近傍なクラスタの選択数は，各年齢層における評価データに対して，最適なクラスタ数を用いた．

6.2 実験条件

実験条件を表 5 に示す．提案するアルゴリズムにより話者適応した音響モデル (Suff-Stat 話者適応モデル) の比較対象として，各年齢層ごとに MAP 適応したモデル (MAP 年齢層適応モデル) と 3 節で作成したたけまる子供用不特定話者モデルを用いる．MAP 適応モデルに関しては，評価用データの各年齢層に対応した年齢層に関する適応モデルを用いて認識実験を行う．評価用データ，言語モデルは 3 節の認識実験と同じものを用いる．

6.3 実験結果

実験結果を図 8 に示す．結果より，年齢層別 MAP 適応および十分統計量に基づく教師なし話者適応により，たけまる子供用不特定話者モデルから一定の認識性能の改善を行えた．

年齢層別にたけまる子供用不特定話者モデルからの改善率をみると，幼児においては，MAP 適応で 0.8%，十

表 4　自動話者クラスタリング実験条件

Table 4　Condition of automatic speaker clustering experiment.

使用データ	音声情報案内システムの子供音声
データ数	59966 個
作成クラスタ数	200
特徴量	MFCC12 次元

表 5　音声認識実験条件

Table 5　Condition of speech recognition experiment using the adapted model.

認識エンジン	Julius ver.3.4.2
タスク	音声情報案内システムの子供音声
言語モデル	文法適用 3-gram(単語数 40k)
音響モデル	1) Suff-Stat 話者適応モデル 2) MAP 年齢層適応モデル 3) たけまる子供用不特定話者モデル
クラスタ選択数	幼児:20，低学年子供:23，高学年子供:22

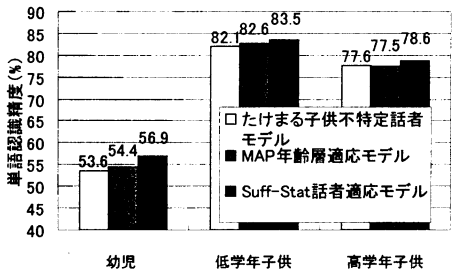

図 8　年齢層ごとの子供音声認識結果

Fig. 8　Results of children's speech recogniton using the adapted model.

分統計量に基づく話者適応で 3.3%，低学年子供においては，MAP 適応で 0.5%，十分統計量に基づく話者適応で 1.4%，高学年子供においては，MAP 適応で 0.1%低下し，十分統計量に基づく話者適応で 1.0%の改善が見られた．

これより自動クラスタリングを用いた十分統計量に基づく話者適応は，いずれの年齢層においても効果的であったといえるが，MAP 適応においては，適応効果が低くあまり有効でないことを示せた．

今回，自動クラスタリングで作成したクラスタ 200 個から入力音声に対して最適なクラスタ数を選択したが，今後はクラスタ数やクラスタの選択数に関する検討も必要である．

7. ま と め

実環境音声情報案内システムで自動収集した自由な発話の子供音声を用いた音響モデルの作成と評価，適応手法の評価を行った．認識実験より，収集した子供音声を用いて作成したたけまる子供用不特定話者モデルにより，子供音声の特徴への対応のみならず，読み上げ音声とは異なる自由な発話にも対応でき，既存のモデルに比較して精度の高い認識ができることがわかった．

話者適応手法に関しては，自動クラスタリングを用いた十分統計量に基づく教師なし話者適応により，収集したデータから作成したたけまる子供用不特定話者モデルに対して一定の効果が得られ，年齢層ごとの MAP 適応よりも高精度な適応効果が得られた．今後はクラスタ数や入力音声に対するクラスタ選択数について更に検討を行う必要がある．

また，話者適応により認識性能の一定の改善が見られたとはいえ，依然幼児に対しては他の年齢層に比べて特に認識性能が低いため，今後は幼児の音声に対して細かく分析し，改善を行っていくことが必要がある．

謝　　辞

松下電器産業株式会社の芳澤伸一博士には研究における多大なる御支援，御助言，御指導を頂きました．心より感謝致します．また本研究は，文部科学省のリーディングプロジェクト「e-Society 基盤ソフトウェアの総合開発」によって行われたものである．

参考文献

[1] 小川 厚徳 他 "小学生音声データベースの構築とそれを用いた子供音声認識の一検討," 電子情報通信学会技術研究報告, SP2002-124, pp.19-24, 2002.

[2] K. Shobaki, J.P.Hosom, R.A. Cole:"The OGI Kid's Speech Corpus and Recognizers," Proc.6th International Conferences on Spoken Language Processing(ICSLP2000), vol.4, pp.258-261, 2000.

[3] 西村 竜一 他 "実環境研究プラットホームとしての音声情報案内システムの運用," 電子情報通信学会論文誌, Vol.J87-D-II, No.3, pp.789-798, 2004.

[4] 芳澤 伸一 他 "十分統計量と話者距離を用いた音韻モデルの教師なし学習法," 電子情報通信学会論文誌, Vol.J85-D-II, No.3, pp.382-389, March 2002.

[5] 河原 達也他 "連続音声認識コンソーシアム 2002 年度ソフトウェアの概要," 情報処理学会研究報告, 2003-SLP-48-1, 2003.

[6] A. Lee, "A New Phonetic Tied-Mixture Model for Efficient Decoding," Proc. ICASSP, pp.1269-1272, 2000.

[7] 鶴身 玲典 他 "タスク文法による N-gram 確率の部分強化を用いた認識アルゴリズムの評価," 情報処理学会研究報告, 2003-SLP-45-13, pp.77-82, 2003.

[8] J.MacQueen. "Some methods for classfication and analysis of multivariate observations," Proc. of the Fifth Berkeley Symposium on Math. Stat. and Prob., volume 1, pp.281-296, 1967.