2005－SLP－58（1）
2005／10／21

音素長伸縮による対話音声認識性能の向上手法

山田 善之, 宮島 千代美, 伊藤 克亘, 武田 一哉

名古屋大学大学院情報科学研究科

〒 464-8603 名古屋市千種区不老町 1
{y-yamada,miyajima,itou,takeda}@sp.m.is.nagoya-u.ac.jp,

あらまし　大語彙自由発話音声認識においては，発話者の話速の変動が大きな問題となる．本研究ではこの問題に対して，入力となる音声信号の音素長を時間軸において伸縮処理をすることで話速のばらつきを抑え，変動による影響を軽減する手法を提案する．本手法は，音素アライメント情報から各音素の伸縮率を決定し，音声の時間軸での圧縮・伸長手法である PICOLA を応用することで音素長を伸縮するものである．本手法を用いた結果，自由発話特有の言い淀みなどによって引き延ばされた母音の圧縮によって，遅い発話の認識率を，70.1%から 83.3%と大きく改善できることが確認された．

キーワード　音声認識，発話速度，PICOLA，自由発話

A spontaneous speech recognition method by adjusting phoneme lengths

Yoshiyuki YAMADA, Chiyomi MIYAJIMA,
Katsunobu ITOU and Kazuya TAKEDA

Graduate School of Information Science, Nagoya University

Furo-cho 1, Chikusa-ku, Nagoya 464-8603, JAPAN
{y-yamada,miyajima,itou,takeda}@sp.m.is.nagoya-u.ac.jp

Abstract　Variation in speech rate is one of the largest problems in large vocabulary spontaneous speech recognition. In order to reduce effects of speech rate variation, we apply a method of adjusting phoneme lengths of input speech by signal processing. Using phoneme alignment information, this method decides the rate of extension/compression of each phoneme, and adjusts phoneme lengths using PICOLA(Pointer Interval Controlled OverLap and Add)'s algorithm. Using this method, we improved the recognition rate from 70.1% to 83.3% for slow speech. The improvements included vowel sounds elongated by hesitation or searching for next words, characteristics of spontaneous speech.

Keywords　speech recognition, speeking rate , PICOLA , spontaneous speech

1 はじめに

カーナビゲーションシステムやロボットコミュニケーションにおいて音声認識システムの応用が期待されている．これらの応用のためには，自由発話の音声認識において高い性能を実現しなければならない．自由発話は読み上げ音声に比べて，話者間での発話の多様性や発話速度の変動が大きな問題となる．そこで本研究では，自動車内での情報提供タスクを想定して収録した CIAIR 車内音声データベース [1] の音声を対象に，認識性能が発話速度によって劣化しない音声認識を目標としている．

発話速度の変動の問題に対して，従来の研究では速い発話の頑健な認識に主眼をおいた音響モデル化手法が多くなされてきた [2, 3]．しかし，これらは全ての発話に対して画一的にモデル適用を行うものであり，大きな精度の改善は得られていない．これらに対して，発話速度に応じてモデルを選択し，発話速度の変動にも対応できるようにした研究もなされている [4]．また，遅い発話に対しては，挿入ペナルティを操作することで認識性能を改善することができることもわかっている [5]．しかしこれらの研究の問題点は，発話中の速度変化に対応しきれない点である．自由発話の特徴として，語尾に向かって発話速度が速くなっていく，発話内容を考える際に有声休止が生じる等が挙げられる．これらの現象より，発話速度を発話単位で平均的に扱うのは処理として不十分であるといえる．

提案手法ではこれらの問題に対して，発話速度の変動を音素単位で伸縮して調整することで対応した．認識システムに入力する波形を信号処理し，発話速度のばらつきを抑えることで速度変動による影響を軽減して認識した．

2 CIAIR コーパスにおける発話速度分析

CIAIR 車内コーパス [1] は 800 人を超えるドライバのデータが記録されたマルチメディア信号 1.4TB 以上から構成されている．本研究で用いた信号は接話マイクを用いて収録された音声で，3 つの異なるシステム（HN システム，WOZ システム，ASR システム）との対話及び，バランス文読み上げ音声が収録されたものである．HN システムは，助手席に座っている人間（ナビゲータ）がドライバの情報検索に関する質問に応えるシステムである．このシステムは３つのシステムの中でドライバが最も機械を

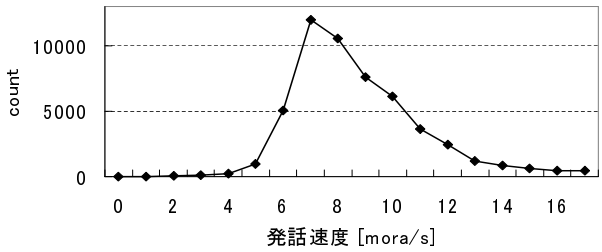

図 1: CIAIR コーパスにおける発話速度の分布 (学習セット).

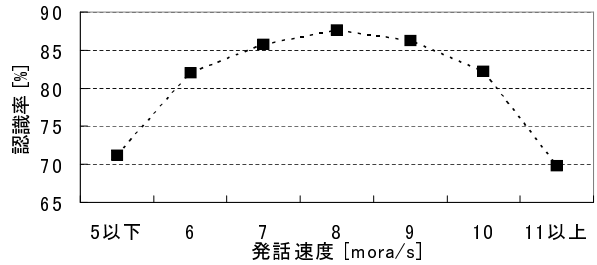

図 2: 発話速度に対する認識率

意識せずに対話できるシステムである．WOZ システムは，人間が検索を行い，その応答を音声合成で出力することで対話を行うシステムである．音声が機械であるため，擬似対話システムといえるシステムである．ASR システムは，1500 語からなる語彙のバイグラムを用いた音声認識システム Julius[6] によってドライバの発話を認識し，その認識結果に基づいて対話処理を行うシステムである．

本研究では発話速度を 1 秒あたりに発話するモーラ数と定義した．発話速度は書き起こしの音素ラベルと各音素 HMM を与え，ビタビアルゴリズムを用いて推定した音素アライメントから推定した．図 1 に本コーパスの発話ごとの発話速度の分布を示す．この分布における平均及び標準偏差はそれぞれ 8.87, 4.59 であった．先行研究 [4] より，日本語話し言葉コーパス (CSJ：Corpus of Spontaneous Japanese)[7] の学会講演音声における発話速度の平均及び標準偏差はそれぞれ 8.70 と 2.10，JNAS の読み上げ音声の発話速度の平均及び標準偏差は 6.27，0.97 である．これらより CIAIR コーパスは，自由発話の特徴が顕著に現れているコーパスであるといえる．また標準偏差が CSJ より高いのは，CIAIR コーパスがカーナビゲーションを想定した対話タスクである上に運転中のため，話す内容をその場で考えなければならないことが多いため，発話速度の変動が大きくなったと考えられる．

発話速度と認識率の関係を図 2 に示す．5[mora/s]

以下の遅い発話と，11[mora/s] 以上の速い発話のどちらにおいても，15%程度低下しているのがわかる．これらの発話数は全体の 18.0%を占めるため，これらの改善を計ることでシステムとしての性能を向上させることができるといえる．

3　音素長伸縮による音声認識

3.1　発話速度変換

本報告では，音声認識における発話速度の変動による影響を除去するために，各音素継続長を一定範囲の長さになるよう伸縮する方法を提案する．発話速度変換は過去に多く研究がなされてきた．それらの手法を応用し，音素単位で時間軸での収縮を可能とすれば，発話ごとにばらついている速度を一定にすることができ，発話速度による影響を軽減することができると考えられる．

発話速度変換技術は，PICOLA[8] や NHK 方式 [9]，STRAIGHT 分析合成 [10] 等で過去に研究されてきた．PICOLA は処理が単純なために手を加えやすく，波形をピッチ周期単位で直接処理するために，スペクトル構造が変化しにくい利点があるため，本報告では PICOLA のアルゴリズムを応用し，各音素に対して異なる伸縮率で，音声波を伸縮処理できるようにした．

3.2　PICOLA の原理

PICOLA(Pointer Interval Controlled OverLap and Add) とは，音声波を音声の基本周期単位で分析し，ピッチを変えずに音声波を伸縮する手法で，速記補助や，放送における時間長調整などの目的に開発された．処理が単純だが，自然な時間長伸縮を可能とする．冗長性を前提とした処理のために，ピッチが安定している母音に対して特に効果的である．このアルゴリズムを用いて，音素継続長を伸縮した．

まず分析ポインタと分析フレームを図 3 のように設定する．この際，分析フレーム長は予想される最大ピッチ周期とほぼ同じくらいにする．本報告では 8[ms](最小基本周波数 125[Hz]) とした．この分析フレームに対して自己相関関数を計算し，その最大値をとる時間遅れを T_p とし，1 ピッチ波長を推定する．

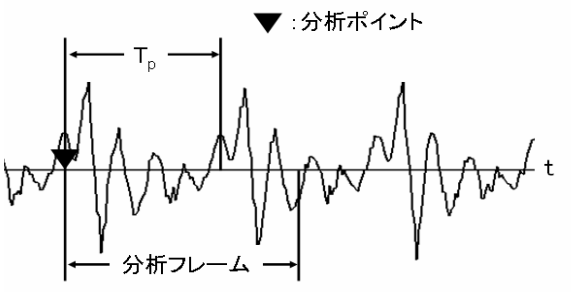

図 3: 分析ポインタと分析フレーム例

波形圧縮の例を図 4 に示す．分析ポインタから，推定した 1 ピッチ波長分の三角窓を 1 から 0 になるようにかけ，同様に次のピッチには 0 から 1 になるようにかける．これらを A，B とする．この A と B の波形を足し合わせて C の波形を作り，A，B と置き換える．この際に用いた三角窓は，C の前後との接続点における連続性を保つために設けたものである．次に，分析ポインタを C 上で $L_C = RT_p/(1-R)$ だけ移動し，同様の操作を行う．

波形伸長の例を図 5 に示す．分析ポインタから，推定した 1 ピッチ波長分の三角窓を圧縮の場合と逆の，0 から 1 になるようにかけ，同様に次のピッチには 1 から 0 になるようにかける．これらを A，B とし，足し合わせた波形を C とする．そして，A と B の間に C を挿入することで波長を伸長する．ここでの三角窓は，圧縮の場合と同様に，C の前後の連続性を保つために設けたものである．次に，分析ポインタを C 上で $L_S = T_p/(R-1)$ だけ移動し，同様の操作を行う．

3.3　音素継続長の伸縮法

まず音声認識結果に基づいて各音素の音素アライメントを抽出する．このアライメント情報から，適切な音素継続長になるように伸縮率を決定する．本報告では音素継続長 60[ms] 以下の音素を 60[ms] 伸長し，90[ms] 以上の音素を 90[ms] に圧縮するように伸縮率を算出した．これらの値は，全音素の継続長を用いて k-means 法で 3 つのクラスにクラスタリングした結果の境界値から決定した．そして分析ポ

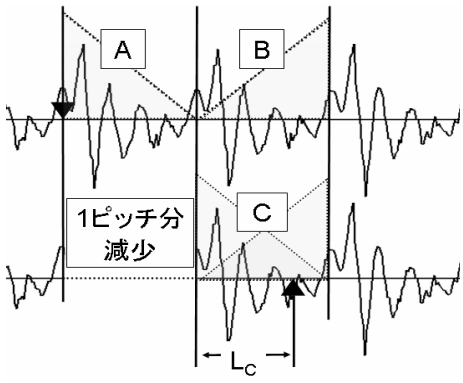

図 4: 波形圧縮例

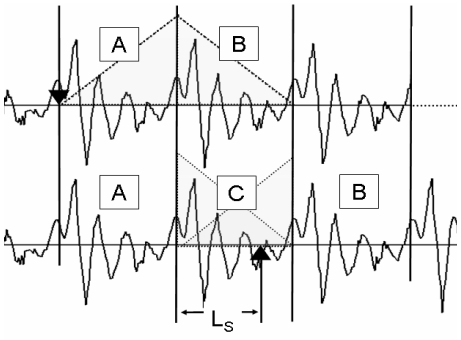

図 5: 波形伸長例

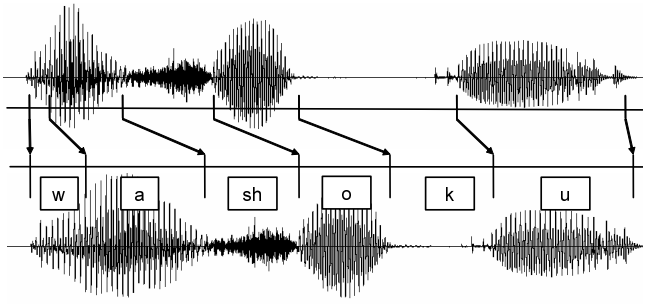

図 6: 上：「和食」という発話に対する伸縮処理前，下：処理後の音声波形.

インタが音素間の境界を越えるまで，PICOLA の伸縮処理を行う．PICOLA は分析ポインタの前後で波形の連続性が保たれるため，伸縮率はアライメント境界を越えた段階で変更するのではなく，ポインタが境界を越えてフレーム処理が終了した段階で次の音素の伸縮率に変更した．図 6 に，音素伸縮処理を行う前後の音声の例を示す．提案手法による認識実験では，このように音素を伸縮させた音声で学習させたモデルを用いる．

4 認識実験

4.1 実験概要

本手法の効果を調べるため，4 種類の方法で音声認識を行い，認識率を調査した．ベースラインシステム，従来法 (発話速度別デコーディング [4])，本手法の 3 種類に加えて，書き起こしラベルによるアライメントを用いた伸縮音声に対しても音声認識を行った．書き起こしラベルを用いる場合は正しい音素数なので，提案手法での目的に沿った伸縮処理ができると考えられる．しかし提案手法では認識結果をラベルとして用いるので，実際の音素とは音素数が異なるために音素アライメントが誤って，想定した処理ができない可能性がある．認識結果による音素数が正解よりも多い場合は，アライメントが実際より細かくなるため，伸長処理が多くなり，不必要な挿入誤りを誘導する可能性がある．また，音素数が少ない場合はアライメントが粗くなるため，圧縮される部分が増え，圧縮する必要のない音声まで圧縮してしまい，正確に認識できなくなることが考えられる．このようにアライメント誤りによる影響が提案手法に与える影響を調査するために，書き起こしラベルによる伸縮音声の認識も行った．

4.2 ベースラインシステム

ベースラインの音響モデルには，CIAIR データベースの HN システム，WOZ システムによって収録した音声，及びバランス文読み上げによる音声によって学習した PTM(Phonetic Tied-Mixture)triphone モデル [11] を用いた．学習データ量は 456 話者，時間である．音素は 43 種類とし，2000 状態，64 混合でモデル化した．音響分析は，フレーム長 25ms のハミング窓，フレーム周期 10ms で行い，各フレームごとに MFCC(12 次元)，Δ MFCC(12 次元)，Power(1 次元)，Δ Power(1 次元) の 26 次元の特徴量ベクトルを求めた．

言語モデルは，CIAIR データベースの HN システム，WOZ システム，ASR システムによって収録した音声を書き起こしたものを用いて作成した．デコーダには認識エンジン Julius[6] を用いた．

4.3 発話速度別デコーディング [4]

比較のため，発話速度別デコーディングシステムを作成した．これは以下に述べる 4 種類のモデル・パラメータを発話速度に応じて選択的に適用するものである．発話速度は正解ラベルから求めたものを用い，x[mora/s]$(N \leq x \leq N+1; N = \{6,7,8,9,10\})$ 及び，6[mora/s] 以下と 11[mora/s] 以上の 7 クラスに分類した．この 7 クラスに対して，最適な条件を選択した．

1. 音響分析のフレーム長・フレーム周期の変更

音響分析のフレーム長やフレーム周期を短くすることで速い発話に対処する．音素 HMM が 3 状態で，フレーム周期が 10ms の場合，一つの音素を表現するのに最短で 30ms 必要となる．しかし実際には 30ms 未満の音素継続長しかない音素も存在するため，フレーム長の操作は短い音素を認識するのに効果的であると考えられる．ここでは，フレーム長を 20ms，フレーム周期を 8ms（ともにベースラインの 80%）に変更する．

2. 飛び越し遷移を許すモデル

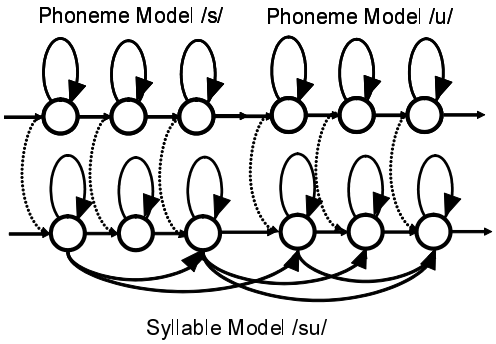

図 7: 音節モデル HMM の遷移図.

各音素 HMM(3 状態 left-to-right) の第 1 状態から第 3 状態に遷移を付加することで，短い音素セグメントに対応する．フレーム周期が 10ms の場合，第 1 状態から第 3 状態に遷移することで，短い音素を 20ms で表現することが可能となる．ただし，音素の中には最低音素継続長が長い音素も存在するため，全ての音素に飛び越し遷移を付加すると認識率が低下すると考えられる．そこで，中心の音素継続長が 30ms 以下になりうるトライフォンを調査し，それらのトライフォンに対してのみ遷移を付加した．学習の方法としては，各音素 HMM の初期モデルに飛び越し遷移を許し，各分布の平均・分散・重み，及び遷移確率の全てのパラメータを推定する．

3. 音節モデル

速い発話では，音素が消失していることがあるため，音節単位でモデル化することで対応を図る．継続時間長が短くかつ頻出する音節を求め，それらに対して音節モデルを作成した．音節モデルは図 7 に示すように音素 HMM を結合して作成し，飛び越し遷移を付与した上で全てのパラメータを推定した．また音節モデルを追加する際には学習量が足りなくなるため，分布共有型 (tied-mixture) によってモデル化する．まず，モノフォンモデルを 64 混合まで学習した上で，対象となる音節の子音と母音の分布を重み付けした上で共有させて，もう一度学習を行う．このモデル化は学習量不足の問題に頑健である．

4. 単語挿入ペナルティの変更

遅い発話に対しては単語挿入誤りが増えるため，単語挿入ペナルティを厳しくして対処する．今回はデフォルトの-2 から-20 まで変化させたとき一番認識率がよかった-14 の値を用いた．

4.4 実験結果

伸縮処理を行う前後における，発話速度の分布を図 8 に示す．この図より，伸縮処理によって発話速度の変動を減少させることに成功したことがわかる．また，各手法による認識率を図 9 に示す．従来法だと全体で 0.4%，提案手法で 3%，正解ラベルを与えると 4%改善される．この結果から提案手法は，アライメント誤りの影響はあるが，話速の変動に対して効果的であることがわかる．

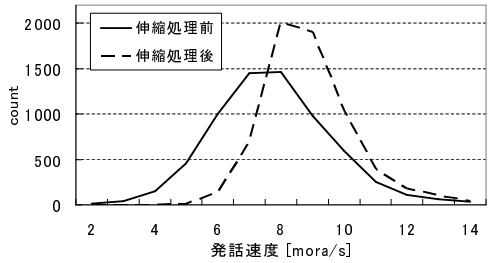

図 8: 伸縮処理を行う前後での発話速度の分布 (テストセット)

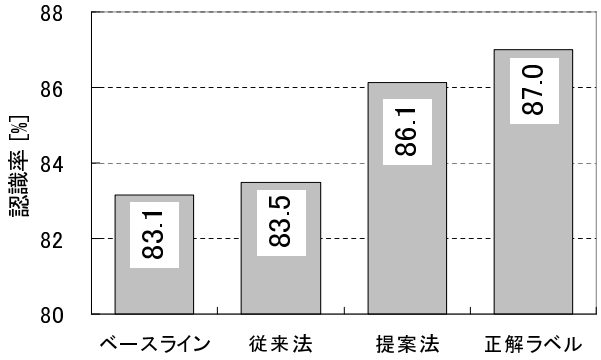

図 9: 手法別の認識率

図 10 に各手法における話速別の結果を示す．どの手法も平均的話速である 8[mora/s] が最もよく，その速度から速くても遅くても認識率が低下している．提案手法は，速い発話については若干従来法のほうが認識率がよかったが，遅い発話に対して大きな改善を見ることができる．平均的話速の 8[mora/s] と，遅い発話である 5[mora/s] との認識率の差が，ベースラインでは 17.6%だったものが，提案手法では 5.9%の差まで改善されている．従来法 (挿入ペナルティの強化) と比べても高い認識率であることがわかる．

挿入ペナルティを操作する場合，発話に対して均一に処理を施してしまうため，局所的に速度が変化するような発話に対応しきれない．例えば,「えーと，喫茶店で.」という発話で,「えーと」の部分で遅く,「喫茶店で」の部分で標準的な速さとなるような発話に対して，挿入ペナルティを強化してしまうと,「えーと」という部分での挿入誤りは解決できても,「喫茶店で」の部分までペナルティを課してしまう．

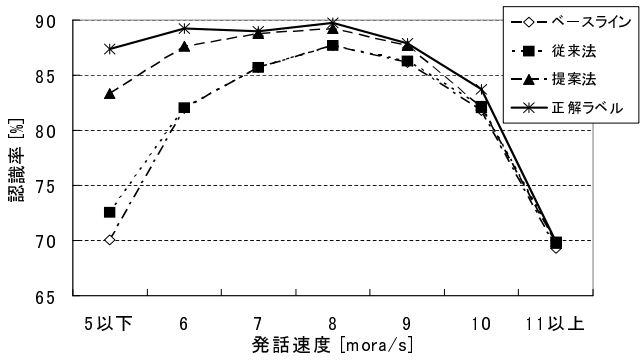

図 10: 話速に対する各手法の認識率

これに対して提案手法では，不必要に長く発音している部分だけを圧縮することができる．上の例では，「えーと (/e: t o/)」の部分，そして仮に語尾が延びてしまった場合,「でー (/d e:/)」の部分も圧縮することができる．その結果，標準的な速さの発話として処理することができるようになる．この局所的処理が認識率向上に貢献したと考えられる．

次に，正解ラベルによる伸縮と提案手法とを比べる．どちらも先述したように遅い発話に対して効果を発揮し，正解ラベルと提案手法の差は，5[mora/s] 以下で 4.1%，6[mora/s] で 1.7%，7[mora/s] で 0.2%となった．また，標準より速い発話ではほぼ同じ認識率となっている．先述したように，提案手法ではアライメント誤差が影響すると考えられるが，遅い発話であるほど大きく影響することがわかる．これは，挿入ペナルティなどのパラメータ調節が平均的な発話に適したものになっているので，発話が遅いと挿入が増え，アライメント誤りが多くなることが原因であると考えられる．

速い発話に対して効果が見られない原因としては，提案手法では音素の消失に対応できない点が考えられる．速い発話では，子音が消失する現象が起きる．実際に速い発話を人が聞く場合，消失した音素を補間するために，内容を理解することができている．提案手法ではこの問題には対応できない．

5 結論

本報告では音声認識システムの性能に影響を与える，話速の変動の問題に焦点を当て，音素継続長伸縮手法の検討とその評価を行った．音素アライメントをもとに，PICOLA のアルゴリズムを用いて音声の時間軸における伸縮処理を施すことで，音素レベルで変動する話速のばらつきを抑え，この問題に対応した．その結果，特に遅い発話に対して有効であることが明らかになった．遅い自由発話に特有の，語尾の長母音化や，話す内容を思考する時間に発話する有声休止の圧縮に成功し，挿入誤りを減少させることができた．提案手法を用いることで，5[mora/s] 以下の非常に遅い発話に対して，10%強の改善を見ることができ，標準的な速度の発話との認識率の差を大きく縮めることができた．また，提案手法は初期認識結果によるアライメントを用いて伸縮率を決定するため，アライメント誤差の影響が大きくなることが予想されたが，特に遅い発話になるにつれてその影響が大きくなることが判明した．また，誤差の影響はあるが，認識システム全体としての性能は向上させることができた．

今後の課題としては，速い発話の認識性能の向上を検討することが挙げられる．先述したように，速い発話に焦点を当てた研究は多くなされている．速い発話に対して効果的な認識手法を見出し，提案手法との選択式にすることでさらなる性能の向上が期待できる．

参考文献

[1] N. Kawaguchi, et. al. ,“ Construction of speech corpus in moving car environment, ”Proc.ICSLP, pp.1281-1284, 2000.

[2] J. Zheng, H. Franco, and F. Weng,“ Word-level rate of speech modeling using rate-specific phones and pronunciations, ”Proc.ICASSP, pp.1775-1778, 2000.

[3] J. Nevel and R. Stern,“ Duration normalization for improved recognition of spontaneous and read speech via missing feature methods, ”Proc.ICASSP, vol.1, pp.313-316

[4] 南條浩輝, 河原達也,“ 講演音声認識のための教師なし言語モデル適応と話速に適応したデコーディング ”, 信学論 (D-II), vol.J87-D-II, no.8, pp.1581-1591, Aug. 2004.

[5] F. Martinez, D. Tapias, J. Alvarez, and P. Leon,“ Characteristics of slow, average and fast speech and their effects in large vocabulary continuous speech recognition, ”Proc. EUROSPEECH, pp.469-472, 1997.

[6] A. Lee et al,“ Continuous Speech Recognition Consortium - an Open Repository for CSR Tools and Models -, ”Proc International Conference on Language Resources and Evaluation (LREC2002), pp.1438-1441, 2002.

[7] 前川喜久雄,“ 言語研究における自発音声 ”, 音講論, 1-3-10, 春季 2001.

[8] 森田直孝, 板倉文忠,“ 自己相関法による音声の時間軸での伸縮方式とその評価 ”, 信学技報, EA86-5, pp.9-16, 1986.

[9] 都木徹,“ 放送における話速変換：話者や音環境の多様性への対応 ”, 音響学会誌 **54**, 533-538 1998.

[10] 河原英紀,“ 聴覚の情景分析が生み出した高品質 VOCODER:STRAIGHT ”, 音響学会誌 **54**, 521-526 1998.

[11] 李晃伸, 河原達也, 武田一哉, 鹿野清宏,“ Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識 ”, 信学論 (D-II), vol.J83-D-II, no.12, pp.2517-2525, Dec. 2000.