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

2006－SLP－62　(1)
2006／7／7

# 有声・無声休止区間の自動検出に基づく自由発話音声認識の性能改善手法

緒方 淳 †　　後藤 真孝 †　　伊藤 克亘 ‡

† 産業技術総合研究所
‡ 法政大学

†{jun.ogata,m.goto}@aist.go.jp
‡itou@k.hosei.ac.jp

**あらまし**　自由発話音声認識においては，不明瞭な発声や口語表現，言い淀み，発話速度の変動など，様々な要因により認識性能が劣化してしまう．本研究では，その中でも特に，現在の音声認識では扱うことが困難な，有声休止，無声休止の 2 つの非言語情報に着目する．本報告では，自然発話中の有声休止，無声休止の音響的特徴をボトムアップな信号処理にて検出し，それらを認識時に考慮することで，両休止に対する頑健な音声認識手法を提案する．CIAIR 車内音声コーパスを用いた自由発話連続音声認識実験を行い，提案手法の有効性を確認した．

**キーワード**　自由発話，連続音声認識，音響モデル，有声休止，無声休止

# Improvements of Spontaneous Speech Recognition by Using Automatic Filled and Silent Pause Detection

Jun Ogata†　　Masataka Goto†　　Katsunobu Itou‡

†National Institute of Advanced Industrial Science and Technology (AIST)
‡Hosei University

**Abstract**　The accuracy of a spontaneous speech recognition system depends on many factors, such as various pauses, unclear pronunciation, spoken expressions, and speaking rates. In this work, we focus on filled and silent pauses, which are hesitation phenomena that degrade the accuracy of continuous speech recognition systems. We propose a speech recognition method that can handle both filled and silent pauses simultaneously. These pauses are automatically detected by using bottom-up acoustical analysis, and the detected results are incorporated into the decoding process. In our experiments using the CIAIR spontaneous speech corpus, the effectiveness of the proposed method was confirmed.

**Keyword**　spontaneous speech, continuous speech recognition, acoustic model, filled pause, silent pause

## 1　まえがき

実環境において音声認識システム，アプリケーションが幅広く利用されるようになるためには，自由発話，話し言葉を頑健に認識する技術が必要不可欠となる．自由発話は，読み上げ音声と比べ，不明瞭な発声や口語表現，言い淀み，発話速度の変動など，現状の音声認識システムでは取り扱うことが困難な様々な要因を含んでいる．このような自由発話音声に関する研究を目的とし，近年では，日本語話し言葉コーパス (CSJ)[1] や CIAIR 車内音声コーパス [2] など，大規模な自由発話音声データベースが構築されており，自由発話音声認識に関する研究の進展が期待されている．

本研究では，自由発話特有の様々な現象に対して頑健な音声認識システムを構築することを目的としている．本報告では，このような現象として**有声休止**，**無声休止**の 2 つを取り上げる．本稿では，有声休止を，「えー」,「あのー」等の会話において場つなぎ的な役割を果たすフィラーに含まれる、母音の引き延ばしの意味で用いる。有声休止は単語や音節の引き延ばしなどの言い淀み現象でも現れ、音声対話において，発話権の保持や心的状態・思考状態の表出といった役割を果たす [3]．一方、無声休止は，基本的には，発話中における無音 (発声がなされていない区間) を表すが，本研究で対象とする無声休止は，読み上げ音声において発生するそれとは意味や性質が異なるといえる．自由

発話，音声対話においては，発話者が発声中に思考状態であることが頻繁にあり，そのため，発話中の様々な箇所に比較的長い無音が挿入される．しかも単語間だけでなく単語内においても無声休止が発生することもあり，誤認識を引き起こす原因となり得る．

そこでこれまでにも，自由発話音声認識の性能向上を目的として，前述した休止情報を取り扱った研究が幾つかなされている．例えば，有声休止に関連する研究としては，日本語の典型的な幾つかのフィラーを語彙として追加登録することにより，連続音声認識システムにおいて扱えるようにする手法 [5] や，サブワード認識を用いた照合処理に基づいて，フィラーの箇所を未知語とみなす手法 [6] などが提案されている．しかし，これらは単独で発生するフィラーに対して有効な手法であり，単語末尾の有声休止 (単語の引き延ばし) や，単語内部における有声休止には対処できない．一方，無声休止に関連する研究としては，連続音声認識システムを用いたときの，離散発声に対する頑健な認識手法 [7]，言い直し発話における音節強調発声に対する性能改善手法 [8] などが提案されている．これらの手法はいずれも，個々の音響モデルをそれぞれの発話スタイルに合わせて精密化するアプローチであり，有声休止も含んだあらゆる休止を扱うことはできない．

また，最近では，大規模な自由発話音声データベースを用いた研究も行われている．[9] では，CSJ を用いた，発音モデリング，言語モデリングの枠組みにおいて，単語内部・末尾の有声休止や，単語内部の無声休止に対処している．しかしながら，この手法は，音声データに対する各休止や発音変形などの詳細な情報が付与された書き起こしテキストを用いることにより，初めて実現される方法である．そのため，タスク依存性が強く，あらゆる自由発話に対応できる手法とはいえない．実際，一口に，自由発話に関する音声データベースといっても，主に講演音声を取り扱っている CSJ と，車内音声対話を取り扱っている CIAIR とでは，発話速度などの自由発話に関する特性が大きく異なっていることが示されている [10]．また，[10][11] では，自然発話中の話速の変動に対処するため，発話中の個々の音素を波形レベルで伸縮させる手法が提案されており，遅い発話に対して改善が得られている．以上の研究では，主として，自由発話音声認識における全体の認識率の改善について報告されているが，有声休止，無声休止に関する個々の解析，認識性能への影響についてはこれまで報告されていなかった．

本研究では，以上述べた従来研究とは別のアプローチとして，自然発話中の有声休止，無声休止の音響的特徴をボトムアップな信号処理にて検出し，それらを認識時に考慮することで，両休止に対する頑健な音声認識手法を提案する．CIAIR 車内音声コーパスを用いた実験により提案手法の評価を行い，有声休止検出，無声休止検出それぞれの音声認識での効果，また，両休止を同時に考慮したときの音声認識での効果について報告する．

## 2　有声・無声休止の自動検出に基づく音声認識手法

有声休止，無声休止を含む音声に対して頑健な認識を行うには，それらの頑健な検出手法，また，休止区間を考慮したデコーディング手法や音響モデリング手法が特に重要となる．以下では，まず，本研究の提案手法の概要について述べ，有声・無声休止区間の検出手法，休止区間スキップを用いたデコーディング手法について説明する．

### 2.1　システムの概要

提案するシステムの概要を図 1 に示す．まず，入力音声に対し，後述する有声休止区間検出，無声休止区間検出をそれぞれ実行し，有声休止，無声休止の区間情報 (始端時刻，終端時刻) を算出する．次に，得られた休止区間情報を音声認識の探索過程に考慮することで，有声・無声休止に頑健なデコーディングを実行し，認識結果を出力する．

また，更なる高精度化のために，音響モデルの学習の段階にも，有声・無声休止区間検出を適用することを考える．音響モデルの学習データの全発話に対し，同様の各休止検出を実行し，区間情報を得る．それらの時刻の情報をもとに，各発話の音響特徴量データ (MFCC) の中から，休止区間に相当する箇所を除去する．そして，休止区間が除去された特徴量データを用いて，音響モデルの再学習を行う．

### 2.2　有声休止の検出手法

有声休止の検出には，後藤らによって提案されたリアルタイム有声休止検出手法 [3],[4] を採用した．本手法では、有声休止が自然な発話において不可避なのは、それが思考プロセスが発話プロセスに追い付かない場合に表れる現象であるからだという仮説に基づく。すると有声休止は、調音器官がほぼ一定のまま声帯が振動し続けるときの音声、すなわち、音韻的に変化が少ない持続した母音の引き延ばしを伴っていると仮定できる。そこで、そうした有声休止が持つ 2 つの音響的特徴 (基本周波数の変動が小さい，スペクトル包絡の

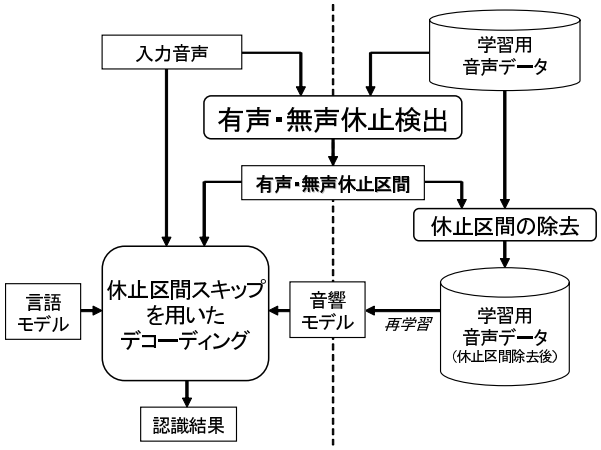


図 1: 提案システムの概要

変形が小さい) をボトムアップな信号処理によってリアルタイムに検出する．そのため，任意の母音の引き延ばしの開始点を終了点を、言語非依存に検出できるという特長を持っている．

## 2.3 無声休止の検出手法

無声休止 (無音) の検出法としては，従来から様々な手法が存在するが，本実験では，音声信号中のパワー情報に基づく検出法，HMM(GMM) に基づく検出法，の 2 種類を検討する．前者は単純に，振幅スペクトルのパワー値に対して閾値処理を行うことにより，無声休止区間の決定を行う方法である．後者の方法としては，本研究では，無音のモデルを含む 245 種類の音節 HMM(後述) を用いた連続音節認識を行うことにより，無声休止の検出を行う．Viterbi アルゴリズムにより求められた，無音モデルに対するセグメンテーション結果により，無声休止区間を決定する．

## 2.4 休止区間スキップを用いたデコーディング手法

以上の手法にて検出された有声・無声休止区間を用いたデコーディング手法について述べる．

発話中に有声休止あるいは無声休止が検出され，その区間情報 (始端時間，終端時間) が与えられると，以下の処理が実行される．フレーム同期ビーム探索において，認識処理が検出された休止区間の開始時刻に到着すると，音声認識器の動作を一時停止し，現時点の認識処理過程 (それまでの仮説情報，探索空間での現在の位置情報等) を保存する．休止区間のフレームは，認識処理の対象とならず，スキップされる．認識処理が休止区間の終端時刻に到着すると，保存された各情報をもとに認識処理が再開される．

本手法は，基本的に，入力音声中の検出された休止

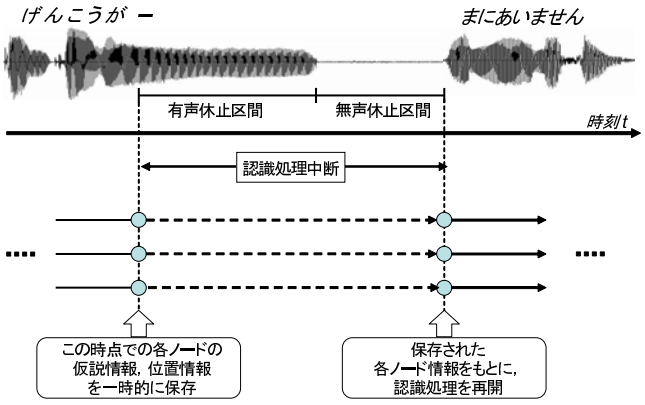


図 2: 休止区間スキップを用いたデコーディング

区間を無視してデコーディングを行う方法である．ただし，無声休止の場合，検出された全ての休止区間を完全に無視すると，単語内において発生した無声休止に対しては有効に働くと考えられるが，単語間に発生した無声休止の場合，実際の認識の際に単語の区切りを同定することが困難になり，認識性能を劣化させる可能性がある．そこで，検出された休止が無声休止の場合は，休止区間全てを無視するのではなく，ある一定の長さだけ休止区間を残すようにする．ここでは，単語内での無声休止に対する効果も考慮し，検出された無声休止を比較的短い時間長に統一する．

以上のデコーディング手法は，先に述べた音響モデルの再学習時と同様に音響特徴量データから休止区間をあらかじめ除去し，除去したデータに対して通常のデコーディングを直接行う場合と基本的には同等の効果をもたらす手法といえる．ただし，提案手法のようにデコーディングの過程において，休止区間の情報を直接組み込むことによって，様々な発展，応用が考えられる．例えば，休止の前後での文脈に関する統計情報を学習しておき，実際の認識時の休止区間においてそれらをデコーディングの過程に組み込む方法や，休止中やその前後では話速が変化する傾向があることから，休止区間に基づいてデコーディングパラメータを動的に決定する手法，などが考えられる．このようなデコーディングの高精度化については今後の課題とする．

# 3 評価実験

提案手法の効果を調べるため，実環境の自由発話音声データを用いて評価実験を行った．

### 3.1 データベース

本実験では，使用するデータベースとして，CIAIR 車内コーパス [2] を用いた．CIAIR は，800 人を超えるドライバのマルチメディアデータで構成される大規模な車内データベースであり，音声データとしては，実際の対話音声，音素バランス文の読み上げ音声が収録されている．文献 [10] によると，CIAIR の対話音声は，自由発話の特徴が顕著に現れている音声データであり，発話速度の変動も他のコーパス (CSJ, JNAS) に比べて大きいことが報告されている．また，本研究で対象としている有声休止，無声休止が頻繁に発生した音声データとなっている．これは，CIAIR はカーナビゲーションを想定した対話であり，発話者は運転中のため，発話内容をその場で考えなければならないことが多いためである．

### 3.2 ベースライン認識システムと実験用音声データ

本実験で使用したベースライン音声認識システム，ならびに学習，評価用データについて述べる．

音響モデルには，長母音化を考慮した日本語音節モデル [12] を用いた．本モデルは，自由発話の特徴である長母音化を個々のサブワード単位内に考慮したもので，日本語自由発話音声認識において，一般的な triphone モデルと同等以上の性能を持つことが示されている．サブワード単位数は 245，そのうちの 3 つは無音モデル (発話開始の無音，発話終端の無音，単語間ショートポーズ) である．サブワード間のコンテキスト依存はなく，mono 音節モデルとなっている．音響分析には，39 次元の特徴ベクトル (12 次元の MFCC とパワー，およびそれぞれの $\Delta$，$\Delta\Delta$) を用いた．学習データには，CIAIR の音声データのうち，401 名のドライバにより発声された音声データ 79093 発話 (音素バランス文読み上げ音声も含む) を用いた．

言語モデルは，CIAIR の対話音声の書き起こしテキスト 94306 文を用いて学習した単語 bigram である．形態素解析には ChaSen Ver.2.3.3 を用いた．構築した bigram の単語数は 4305 である．

デコーディングには，back-off 制約を考慮した探索ネットワークによる $N$-best 探索手法 [13] を用いた．これにより，1-pass デコーディングであっても，単語対近似と同等の性能で，かつ効率的な認識が可能となっている．

評価用データとしては，まず上記の学習データとは別のデータ 11021 発話に対して，有声，無声検出器を実行し，その中から，より多くの休止が含まれている発話 (発話中の休止区間長の合計が長い発話) を順に 600 発話選択した．

### 3.3 ベースラインシステムの実験結果

まず，ベースラインの認識システムを用いて実験を行った．

一般的に，音声認識システムを構築する際，特に本研究で対象とするデータのように無声休止が頻繁に発生する場合には，通常の音声だけでなく，発話中の無音区間 (ショートポーズ) をいかにモデル化するかが認識性能に大きく影響すると考えられる．本実験では，無音のモデリング手法として以下の 3 つのパターンを比較した．

- 発音辞書中の各単語に対し，各音韻系列の終端にショートポーズ (sp) を付与したエントリを新たに追加する．(**sp,dict**)
- 言語モデルの学習段階で，1 つの単語のエントリとしてショートポーズをモデル化する．(**sp,lm**)
- 言語モデル，発音辞書のいずれにおいてもショートポーズを考慮しない．(**nosp**)

sp,dict は，辞書中の全ての単語の末尾にショートポーズが発生する可能性を考慮したモデリング手法である．sp,lm は，学習データを用いて，ある一定の長さのショートポーズが発生する確率を言語モデルの枠組みにてモデル化する手法である．sp,lm を実現する際には，まず，全学習データに対して書き起こしテキストを用いた Viterbi アラインメントを行い，自動的にショートポーズの区間情報を得た．次に，得られた区間情報をもとに，ショートポーズを表す単語「sp」を言語モデル学習テキスト中に付与し，単語 bigram を構築した．ただし，以上の手法は，いずれも単語間のショートポーズに対するモデル化であり，単語内のショートポーズには対処することはできない．

各ショートポーズモデリング手法ごとの，ベースライン認識結果を表 1 に示す．結果より，ショートポーズをモデル化しない場合 (nosp) には，その他 2 つの手法に比べて認識率が大きく劣化していることがわかる．sp,dict と sp,lm を比較すると，sp,lm の方が若干認識率が高い結果となった．sp,dict においては，通常では発声直後に無音が発生しないような単語に対してもショートポーズを考慮してしまう．その結果，認識の際に余計な音響パターンが増えてしまったことが原因と考えられる．以降の実験では，sp,lm の結果をベースラインとした．また，全体の結果として認識率は低い結果となっている．これは，今回の評価用データが，言い淀みやフィラー，単語末尾の有声休止，単語内部の無声休止などを多く含むためと考えられる．

表 1: ベースラインシステムの認識率 (%)

| | 単語正解精度 |
|---|---|
| sp,dict | 67.41 |
| sp,lm | 67.87 |
| nosp | 55.71 |

特に，単語末尾の有声休止，単語内部の無声休止に関しては，ほとんどのケースで誤認識となっていた．

### 3.4 実験結果

有声・無声休止区間検出に基づく音声認識の評価実験を行った．2.4 節で述べた，無声休止検出後の区間長は予備実験の結果より，0.1sec とした．

実験結果を図 3 に示す．ここで,「ベースライン」はベースライン音声認識の結果,「有声」は有声休止検出のみを用いた場合,「無声 (パワー)」はパワー情報に基づく無声休止検出のみを用いた場合,「無声 (HMM)」は HMM に基づく無声休止検出のみを用いた場合,「有声・無声」は両方を用いた場合の結果である．提案手法の 4 種類では，検出結果に基づいて音響モデルを再学習したときの結果も示している．なお,「有声・無声」の実験には，無声休止検出として，2 種類の検出器 (パワー，HMM) による結果の比較 (後述) をもとに，パワー情報に基づく検出器を用いている．

まず，有声休止区間検出を考慮した音声認識結果とベースラインの結果を比較すると，音響モデルの再学習も併用することで約 3%の性能向上がみられた．認識結果を調べたところ，改善されたパターンとしては，単語末尾の有声休止，単語内部の有声休止に対するものが特に多かった．CIAIR のような対話タスクの場合，ある特定の名詞 (店名や地名，製品名等) において，有声休止が発生することが頻繁にあり (例: コンビニ ⇒ コンービニ)，これを言語モデルや発音モデルなどの事前学習の枠組みのみで解決することは困難と考えられる．また，極端に長く発声されたフィラーに関しても，本手法によって湧き出し誤りが抑えられ，多くの改善が得られた．

次に，無声休止区間検出を考慮した音声認識結果とベースラインの結果を比較する．音響モデルの再学習も併用することで，最高で 2.6%の性能向上がみられた．本手法にて改善されたパターンとしては，まず単語内部の無声休止に対するものが挙げられる．以下に実際に本手法で改善された，単語内無声休止の発声例を示す (発声文中の単語のみを表示，[sp] は無声休止を表す)．

- ホテル ⇒ ほ [sp] てる
- 今池支店 ⇒ いまい [sp] けしてん

実環境での対話タスクにおいては，発話者は発話内容を思考しながら発声することが多いため，単語間だけでなく，単語の内部においてもこのような無声休止が多くみられる．また，[8] で報告されている「言い直し時の音節強調発声」(例 : コンビニ ⇒ コ [sp] ン [sp] ビ [sp] ニ) においても，単語内部 (音節ごと) の無声休止が誤認識を多く引き起こすことからも，このような発声パターンを改善することの意義は大きいといえる．また，単語間の無声休止については，今回の実験では，ベースラインのシステムにおいて，単語のエントリとしてショートポーズを学習しているため，無声休止が発生したことによる直接的な誤認識はみられなかった．しかし，比較的長い無声休止中に，発話者のリップノイズや小さな舌打ちなどのの雑音が発生し，これによる湧き出し誤りが発生していた．パワー情報に基づく無声休止検出では，これらの比較的小さな雑音が閾値処理によって棄却されており，認識改善に寄与していた．一方，HMM に基づく無声休止検出では，これらの比較的小さな雑音を棄却できなかったため，パワー情報に基づく無声休止検出の場合より認識率が低い結果となった (図 3)．

最後に，有声・無声休止の両方の検出結果を考慮した音声認識結果について述べる．ベースラインと比べて，音響モデルの再学習も併用することで，有声休止，無声休止を単独で行うよりも更に改善が得られ，最終的に約 4%の性能向上がみられた．本実験で扱ったような自然発話，特に対話音声においては，有声休止，無声休止が 1 発話中に同時に発生することも頻繁にあると考えられる．また，1 発話中の，ある 1 単語においてこれら 2 つの休止が同時に発生することも考えられる．本実験データにおいて実際に存在した例を以下に示す．

- 正解 : **和食** の お店 に …
- 発声 : **わしょ** [sp] **くー** の おみせ に …
- 従来 : **場所 くう** の お店 に …
- 提案 : **和食** の お店 に …

上から順に，正解単語列，実際の発声，従来手法での認識結果，提案手法での認識結果をそれぞれ示している．この例では,「和食」という単語の内部に無声休止が発生し，末尾に有声休止が発声している．本手法を用いることで，このような，従来の音声認識手法では困難な発声に対しても改善が得られることを確認した．

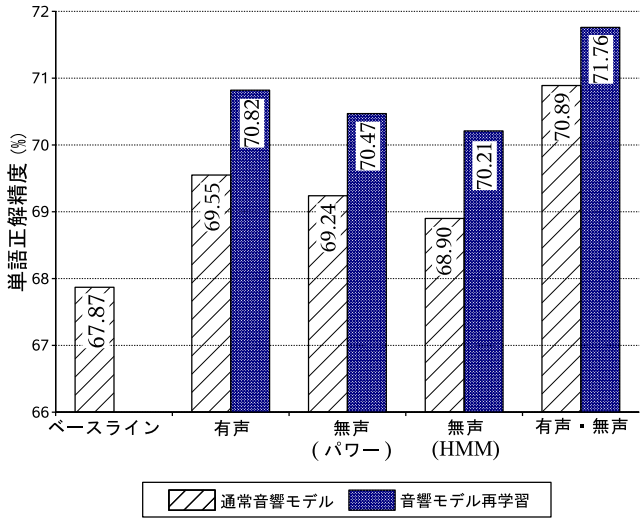


図 3: 各認識手法の認識精度

## 4 まとめ

本報告では，自由発話音声認識の性能改善を目的とし，有声・無声休止区間の自動検出に基づく音声認識手法を提案し，自由発話音声認識実験にてその評価を行った．発話中の有声休止，無声休止をボトムアップな信号処理により検出し，それらの区間をデコーディング時にスキップすることにより，両休止による認識劣化を防ぐことができた．

従来，このような休止情報を音声認識にて扱った研究としては，音響モデルや言語モデル，発音モデルなどにおける事前学習の枠組みによって，休止情報をモデル化するものがほとんどであった．そのため，タスクやデータベースに依存することが避けられず，あらゆる自由発話に対処することは困難であった．本研究は，ボトムアップな信号処理による，言語やタスクに非依存な休止区間検出手法を用いることで，あらゆる自由発話中の休止に対して頑健な手法の実現を目指したものである．

今後の課題としては，2.4 節で述べたデコーディングアルゴリズムの高精度化，他のタスク，データベースによる評価などが挙げられる．また，本実験では，音響モデルとしてコンテキスト独立の音節モデルを用いているが，コンテキスト依存モデルにも対応できるように，本デコーディングアルゴリズムを拡張する予定である．

## 参 考 文 献

[1] 河原達也: “『日本語話し言葉コーパス』を用いた音声認識の進展”, 第 3 回話し言葉の科学と工学ワークショップ講演予稿集, pp.61-66, 2004.

[2] K.Takeda, H.Fujimura, K.Itou, N.Kawaguchi, S.Matsubara and F.Itakura: “Construction and Evaluation of a Large in-car Speech Corpus, IEICE Transactions on Information and Systems, Vol. E88-D, No. 3, pp.553-561, 2005.

[3] 後藤真孝，伊藤克亘，速水悟: “自然発話中の有声休止箇所のリアルタイム検出システム”, 信学論 (D-II), Vol.J83-D-II, No.11, pp.2330-2340, 2000.

[4] Masataka Goto, Katunobu Itou, and Satoru Hayamizu: “A Real-time Filled Pause Detection System for Spontaneous Speech Recognition”, Proc. Eurospeech,pp.227-230, 1999.

[5] 中川聖一，小林聡: “自然な音声対話における間投詞・ポーズ・言い直しの出現パターンと音響的性質”, 音響誌，vol.51, no.3, pp.202-210, 1995.

[6] 甲斐充彦，中川聖一: “冗長語・言い直し等を含む発話のための未知語処理を用いた音声認識システムの比較評価”, 信学論 (D-II)，vol.J80,-D-II, no.10, pp.2615-2625, 1997.

[7] F.Alleva, X.Huang, M-Y.Hwang and L.jiang: “Can Continuous Speech Recognizers Handle Isolated Speech?”, Proc. EuroSpeech, pp.911-914, 1997.

[8] 奥田浩三，松井知子，中村哲: “誤認識時の言い直し発話における発話スタイルの変動に頑健な音響モデル構築法”, 信学論 (D-II), Vol.J86-D-II, no.1, pp.42-51, 2003.

[9] 堤怜介，加藤正治，小坂哲夫，好田正紀: “発音変形依存モデルを用いた講演音声認識”, 信学論 (D), Vol.J89-D, no.2, pp.305-313, 2006.

[10] 山田善之，宮島千代美，伊藤克亘，武田一哉: “音素長伸縮による対話音声認識性能の向上手法” 情処研報, 2005-SLP-58, pp.1-6, 2005.

[11] 山田善之，武田一哉，伊藤克亘，宮島千代美: “発話速度変動に頑健な波形伸縮による音声認識手法の検討” 日本音響学会講演論文集, pp.63-64, 2006-3.

[12] 緒方淳, 有木康雄: “日本語話し言葉音声認識のための音節に基づく音響モデリング”, 信学論 (D-II), Vol.J86-D-II, No.11, pp.1523-1530, 2003.

[13] 緒方淳, 有木康雄: “大語彙連続音声認識における最ゆう単語 back-off 接続を用いた効率的な $N$-best 探索法”, 信学論 (D-II), Vol.84-D-II, No.12, pp.2489-2500, 2001.