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

2006－SLP－63（8）
2006／10／20

分散型音声認識の商用システム構築

加藤 恒夫† 　河井 恒† 　宇都宮 栄二‡

†株式会社 KDDI 研究所 〒356-8502 埼玉県ふじみ野市大原 2-1-15
‡KDDI 株式会社 〒102-8460 東京都千代田区飯田橋 3-10-10
E-mail: †{tkato, Hisashi.Kawai}@kddilabs.jp, ‡ei-utsunomiya@kddi.com

あらまし　携帯電話アプリケーションの日本語入力を支援するため、2006 年 1 月よりコンシューマ向けに分散型音声認識のサービスを開始した。携帯電話マイクに入力された音声は携帯電話機上で音響特徴量に変換され、パケット通信で音声認識サーバに送信される。携帯電話が音声認識サーバから受信した認識結果は画面表示されるため、ユーザは瞬時に認識結果を確認し、認識を誤った場合にも誤認識箇所を部分的に修正することができる。音声認識に対するストレスや不安を軽減するため、携帯電話機上の特徴量抽出処理をリアルタイム化し応答時間を数秒に短縮するとともに、誤認識の可能性が高い場合に「声が大きすぎます」、「雑音が大きすぎます」、「発声が早すぎます」と 3 種類のアラームを発生する機能を追加した。また、ネットワークのコンテンツに日々追加される新しいキーワードを認識できるようにするため、サービスを停止せずに単語辞書・文法を更新する機能を開発した。

キーワード　分散型音声認識, 携帯電話

Development of a Commercial System of Distributed Speech Recognition

Tsuneo KATO† 　Hisashi KAWAI† 　and 　Eiji UTSUNOMIYA‡

†KDDI R&D Laboratories Inc. 2-1-15 Ohara, Fujimino-shi, Satitama, 356-8502 Japan
‡KDDI Corporation 3-10-10 Iidabashi, Chiyoda-ku, Tokyo, 102-8460 Japan
E-mail: †{tkato, Hisashi.Kawai}@kddilabs.jp, ‡ei-utsunomiya@kddi.com

Abstract　To assist Japanese text input for applications on cellphones, a distributed speech recognition service for consumer applications was launched in January 2006. Speech input to a microphone is processed for acoustic feature extraction on the cellphone, then the features are transmitted to a speech recognition server by packet exchange, and recognition results received from the server are displayed on the screen. The recognition results are confirmed by sight, and partial correction of misrecognized words is possible if any. To reduce stress and unfamiliarity to speech recognition technology, response time from the server was shorten to a few seconds by real-time acoustic feature extraction on the cellphones, and warning function of three alarms, “Voice too loud ”, “Noise too loud”, and “Uttered too early”, were added to the client software. Moreover, a function of reloading new grammars and lexicons through a nonstop operation is equipped on the speech recognition server to enable frequent update of grammars and lexicons for network contents.

Keyword　Distributed Speech Recognition, Cellphone

1. はじめに

近年、携帯電話は電話機としてだけでなく、メール、ウェブブラウズ、スケジュール管理等が行える携帯情報端末として広く利用されている。メール作成やウェブにおける URL や検索キーワードの入力、電話帳やスケジュール帳の編集など 10 キーによるテキスト入力は日常的に発生している。テキスト入力の負荷を軽減するために、予測変換機能を中心とした日本語入力システムの改良が進んでいるが、10 キーの操作に煩わしさを感じるユーザは多い。こうしたユーザのテキスト入力を支援するため、KDDI は携帯電話に分散型音声認識を実装し、「声 de 入力」機能として 2006 年 1 月にコンシューマ向けの音声認識サービスを開始した。

従来の電話音声認識とは異なり、携帯電話のマイクから入力された音声は携帯電話機上で音響特徴量に変換され、音響特徴量はパケット通信で音声認識サーバに送信される。音声認識サーバの認識結果は同じくパケット通信で携帯電話機に返却され、画面表示される。電話音声認識のように認識結果を音声で確認する代わりに画面上で確認できるため、認識を誤った場合でもユーザは誤った箇所を瞬時に見分け、部分的に修正することができる。

また、計算量や消費メモリの多い認識処理をサーバ

で実行するため、携帯電話のリソースに制限されることなく大語彙連続音声認識を提供することができる。さらに単語辞書と文法がサーバに置かれるため、ウェブ等のネットワークコンテンツと連携しやすい。コンテンツの変化にあわせて単語辞書と文法を更新することも容易である。

商用システムの構築にあたり、音声認識に対するストレスや不慣れなユーザが抱く不安を軽減するために、高速な応答性とユーザフレンドリーなインタフェースの開発に重点をおいた。

前者については、携帯電話とサーバ間の通信時間を含めてストレスのない応答時間を実現する必要がある。そのため、発声終了後まもなく携帯電話からサーバに音響特徴量が送信されること、音声認識サーバの認識処理が 2 秒以内に完了することが必要条件になった。

後者については、従来の音声認識サービスでは認識結果が誤った場合や認識結果が得られなかった場合に、なぜ認識を誤ったのか、なぜ認識結果が得られなかったのか、ユーザは原因を推測する情報さえ与えられなかった。そのため、不適切な発声を繰り返したり、音節を区切って発声することにより却って認識されにくくなったりすることが多かった。認識結果が正しいかどうか、認識結果が誤った場合の原因、認識結果が得られなかった場合の原因、を正確に特定することは困難であるが、ユーザに情報が通知されない状況を改善するため、誤認識の可能性が高い場合に推測される理由をユーザに通知する機能を開発した。通知する内容は、「声が大きすぎます」、「発声が早すぎます」、「背景雑音が大きすぎます」の 3 種類である。

前述のとおり、分散型音声認識はネットワーク上のデータベース検索に適している。データベースに日々登録される新しいキーワードを認識できるようにするには、音声認識サーバの単語辞書と文法（以降辞書と呼ぶ）にも新しいキーワードを登録する必要がある。これまで辞書更新時には音声認識サーバを停止しなければなかった。しかし、辞書更新のたび音声認識サービスを停止するのでは、頻繁な更新が難しくなるため、サービスを停止せずに辞書更新する機能を開発した。

以上の特徴をもつ分散型音声認識の最初のアプリケーションは、その便利さを広くユーザに訴えかけられ、普及の促進に繋がることを条件として選定された。具体的には、

- ユーザが多いこと
- 従来の 10 キーによるテキスト入力が多いこと
- 急いでいたり歩きながらで使用される場合に、10 キー操作が難しく、音声入力が選ばれやすいこと
- 音声認識を適用した場合に適当な語彙サイズと定型的な発声パタンが得られること

である。その結果、乗換検索サービスと目的地検索サービスが最初の適用サービスとされた。

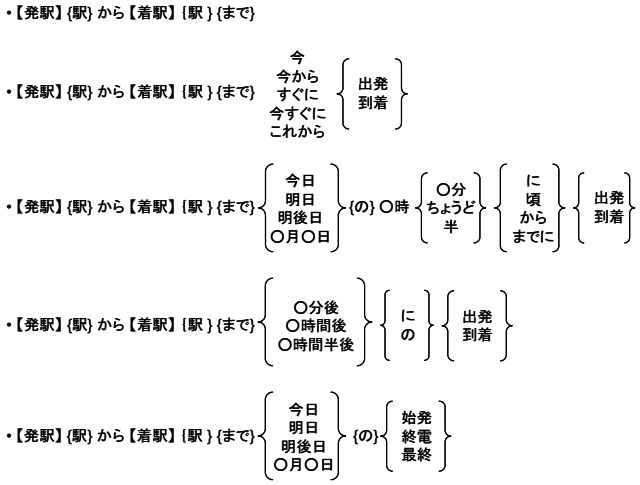

図 1　乗換検索サービスで認識可能な文型

本稿では、次節で最初のアプリケーションを紹介し、第 3 節でシステム構成、第 4 節で携帯電話機へのクライアント実装と誤認識原因の推定通知機能、第 5 節で音声認識サーバの特徴を述べる。

2. アプリケーション

2.1. 乗換検索サービス

乗換検索サービスは、指定された発駅、着駅と出発時刻もしくは到着時刻に対して電車の乗換情報を提供するサービスである。GUI ベースの条件入力画面にテキストを入力し、サーバにあるデータベースを検索するクライアント･サーバ型のアプリケーションである。

従来の 10 キー入力による乗換検索の場合、発駅、着駅、出発時刻もしくは到着時刻をそれぞれのスロットに入力する。音声入力では「【発駅】から【着駅】まで〇月〇日の〇時〇分に【出発／到着】」というように、複数の検索条件を一度の発声に含めれば、それぞれのスロットに認識結果が表示される。乗換検索で認識可能な文型を図 1 に示す。日時指定の省略に対しては「現在時刻に出発」がデフォルト検索条件として入力されるが、認識率の低下を防ぐため原則として「【発駅】から【着駅】まで【日時】に【出発／到着】」の語順は固定にしている。この発声パタンは図 2 のとおり画面表示されるので、不慣れなユーザは発声パタンを目で確認しながら発声することができる。

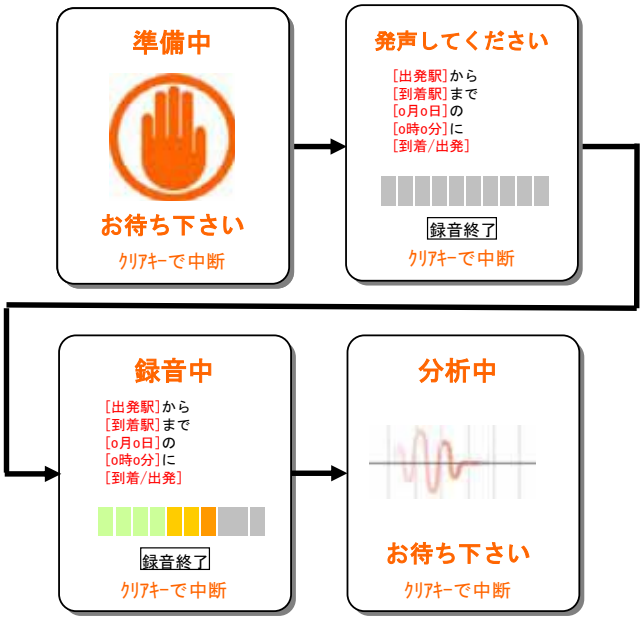

図 2　乗換検索の発声時の画面遷移

各スロットに入力された認識結果が誤っている場合は、部分的に修正することができる。修正の方法は、プルダウンメニューによる別候補の選択、10 キーによる日本語入力、音声再入力、の 3 種類である。音声再入力による修正は、各スロットの隣に置かれたマイクボタンを選択して行う。検索条件の入力後、検索ボタンを選択すると乗換情報が得られる。

2.2. 目的地検索サービス

目的地検索サービスは、住所、電話番号、店名／施設名、駅／空港名で指定された場所の地図を表示し、GPS で取得した現在地から目的地までの道順をガイドするサービスである。乗換検索サービスの場合と同様に、スロットへの入力に分散型音声認識を用いる。

目的地検索サービスにおける住所入力、電話番号入力、店名／施設名入力、駅／空港名入力はメニューが分かれている。各メニューで認識可能な文型を図 3 に示す。店名／施設名入力では、単独の店名／施設名に加えて「【地名／駅名】の【店名／施設名】」のパタンを認識できる。メニューによって辞書を切り替えるための辞書名が、音響特徴量とともにサーバに送信される。同サービスでは以下の 4 種類の辞書が切り替えて使用される。

- 駅名／空港名
- 全国住所
- 電話番号
- 全国の主要施設名／店名

■住所

- {【都道府県】}【市区町村】{【町名】}{【地番】}

■電話番号

- 【固定電話番号】
【携帯電話番号】
【PHS電話番号】
【IP電話番号】

■店名／施設名

- {【地名】}{の／にある}【店名／施設名】

■駅名／空港名

- 【駅名】{駅}
【空港名】

図 3　目的地検索サービスで認識可能な文型

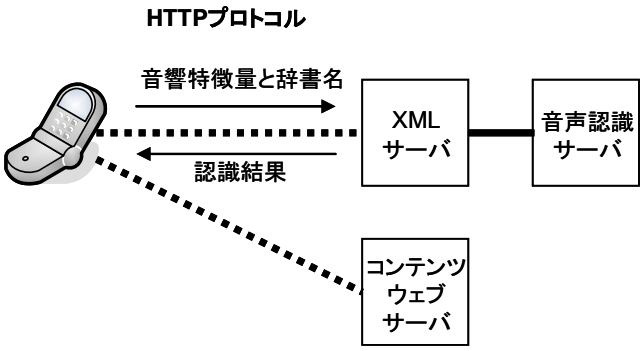

図 4　分散型音声認識システムの構成

3. 分散型音声認識システムの構成

分散型音声認識システムの構成を図 4 に示す。携帯電話、XML サーバ、音声認識サーバで構成される。分散型音声認識のサーバ群はアプリケーションのコンテンツウェブサーバとは別のサーバを使用している。携帯電話のマイクに入力された音声は携帯電話機上で音響特徴量に変換され、XML サーバをゲートにして音声認識サーバに送信される。音声認識サーバによる認識結果は XML ドキュメントもしくは端末表示用の HTML ドキュメントに加工されて携帯電話に返却される。

既存のウェブコンテンツやウェブベースのアプリケーションに音声認識を追加しやすいように、携帯電話と XML サーバ間の通信は HTTP[3] を使用している。携帯電話から XML サーバへの音声認識要求は POST メ

ソッドで、音響特徴量とアプリケーション名、辞書名、音響特徴量の種類名を含むXMLが含まれている。音声認識サーバのインタフェースもHTTPである。CGIプログラムが起動し、音声認識要求から音響特徴量を取り出し、同じく音声認識要求から取り出した辞書名を指定して認識処理を行う。音声認識サーバの応答は、アプリケーションに応じてXMLドキュメントやHTMLドキュメントのフォーマットで返却される。

XMLサーバは、双方向の通信に含まれるXMLドキュメントのフィルタとして機能する。アプリケーションに応じた音声認識サーバの振り分け、不正な音声認識要求の排除、携帯電話機種によるフォーマットの違いの吸収、音声認識要求に含まれる認識に不要な要素の除去、認識結果をキーとするデータベースの検索等を行う。

検索結果を得るには、10キーによるテキスト入力と同様にスロットに格納された検索条件をコンテンツウェブサーバに送信し、コンテンツウェブサーバの応答を受信する。

4. 携帯電話機へのクライアント実装

4.1. 特徴量の抽出

マイクから入力された音声は携帯電話機上で音響特徴量に変換され、サーバに送信される。音響特徴量はETSI (European Telecommunication Standard Institute) 勧告の標準方式ES201108[1]とES202050[2]を採用した。サンプリング周波数は8kHzである。ES201108は背景雑音の抑圧処理を含まずMFCC音響特徴量を抽出し、ベクトル量子化によりデータを圧縮する。ES202050は音響特徴量抽出の前段にWienerフィルタによる背景雑音抑圧処理を含むため、耐雑音性が強化されている半面、計算量が多く、CPUに対する負荷が大きい。そこで、背景雑音の大きさに応じて2種類の音響特徴量を使い分けることにした。

音声の取り込みが始まると、発声を促すプロンプト表示の前の数百ミリ秒間（図2の左上の画面）で背景雑音を取得する。この背景雑音の平均パワーを基準にして、相対パワーにより発話検出を行うとともに、2種類の特徴量抽出方式を切り替える。背景雑音の平均パワーが閾値以下ではES201108が選択され、閾値以上ではES202050が選択される。発話検出後の入力信号に対して順次音響特徴量が抽出される。終話検出も発話検出と同様に背景雑音パワーを基準とする相対パワーに基づいて行われる。終話が検出されると、音響分析処理は終了し、音響特徴量がサーバに送信される。

終話検出後まもなく音響特徴量が送信されるように、特徴量抽出プログラムは整数演算化、レジスタ変数の参照効率化、コードの最適化等を行いネイティブ層に実装された。その結果、ES201108、ES202050ともにリアルタイム処理が行われるようになった。ES201108の処理負荷はES202050の約1/3である。

4.2. 誤認識原因の推定通知機能

誤認識の可能性が高い場合に、推定される原因をユーザに通知する機能として、マイク入力信号のオーバーフロー検出、過大な背景雑音の検出、話頭切断の検出の3機能を追加した。3機能はいずれも音声認識処理を必要とせず、クライアントの処理だけで判定されるものである。

なお、判定に用いた3種類のパラメータについて、ある値を境に認識率が急激に低下するといった明確な関係が得られなかったため、閾値を超えた場合には音声認識結果に加えて「声が大きかったため認識結果が正しくない可能性があります」、「雑音が多かったため認識結果が正しくない可能性があります」のように画面表示を行うこととした。

4.2.1. マイク入力信号のオーバーフロー検出

マイク入力信号がA/D変換の最大レベルを超えるとディジタル信号は最大値／最小値でクリッピングされる。オーバーフローが発生した信号の短時間周波数分析の結果には本来存在しない高周波成分が現れ、認識率の低下を招く。そこで、クリッピングしたサンプルをカウントし、閾値を超えたらアラームを発生させる。

判定に用いたパラメータは次式のとおりである。

$$N_{OFS} = \max_{1 \le t \le T} n_{OFS}(t)$$

$n_{OFS}(t)$はt番目の特徴量抽出フレームにおけるオーバーフローサンプルの数、Tは発声のフレーム数であるのでN_{OFS}は発声全体の最大値である。サンプリング周波数8kHzで、特徴量抽出フレームの大きさは25msであるため、$n_{OFS}(t)$は200以下の値である。

4.2.2. 過大な背景雑音の検出

携帯電話機に実装した音響特徴量抽出部では背景雑音が大きい場合に雑音抑圧処理を含む分析方式を使用し、音声認識サーバでは背景雑音を考慮した音響モデルを用いてパタンマッチングを行っているが、背景

雑音の増大に伴い認識率は低下してしまう。

そこで、発声プロンプトの前の数百ミリ秒で取得した背景雑音レベルに基づき、2 種類の分析方式を切り替えるとともに、さらに別の閾値より大きい場合には過大な背景雑音のアラームを発生する。即ち、

- $P_{noise} \leq \theta_{switch}$ の場合・・・ES201108
- $P_{noise} > \theta_{switch}$ の場合・・・ES202050
  - さらに $P_{noise} > \theta_{alarm}$ の場合

・・・過大背景雑音アラーム

ここで、P_{noise} は背景雑音レベル、θ_{switch} は分析方式切替用の閾値、θ_{alarm} はアラーム発生用の閾値で、大小関係は $\theta_{switch} < \theta_{alarm}$ である。

4.2.3. 話頭切断の検出

話頭切断は、音声の取り込みが始まる以前にユーザが発声を開始した場合に、発声の先頭が収録されないことを指し、発声の先頭が欠落しているために認識率の低下を招く。

正しく収録された発声は、発話検出前と終話検出後に無音声区間があり、突発的な背景雑音がなければ両者のパワーは大きく異ならない。話頭切断が起きた発声は、発話が検出された場合でも発話検出前に無音声区間がないため、音声取り込み開始直後のパワーが、終話検出後の無音声区間のパワーよりも高くなると考えられる。そこで、音声取り込み開始直後の対数パワーと終話検出後の対数パワーの差分に対して閾値処理を行い、閾値以上の場合にアラームを発生させる。即ち

- $10\log_{10}\frac{P_{micon}}{P_{endpo\,\mathrm{int}}} > \theta_{early}$ の場合

・・・話頭切断アラーム

ここで、P_{micon} は音声取り込み開始直後のパワー、$P_{endpo\,\mathrm{int}}$ は終話検出後の無音声区間のパワー、θ_{early} は閾値である。

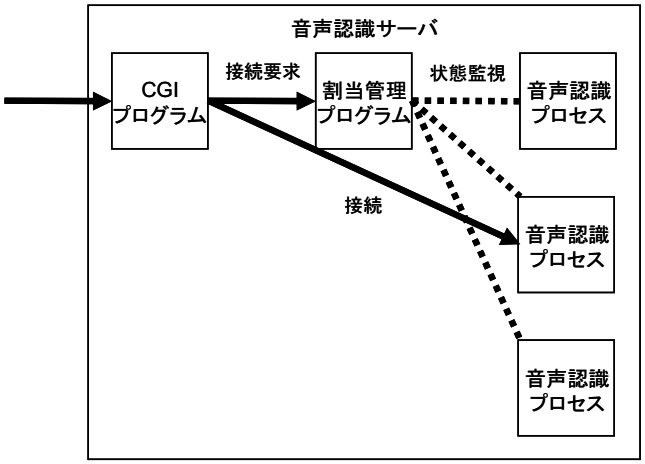

図 5 音声認識サーバの構成

5. 音声認識サーバ

5.1. 音声認識サーバソフトウエア

図 5 に音声認識サーバの構成を示す。同時に複数の認識要求を処理できるように、複数の音声認識プロセスが起動している。音声認識プロセスの状態は｛起動中／処理待ち／処理中｝の 3 状態で割当管理プログラムにより管理される。CGI プログラムからの認識要求に対して、割当管理プログラムは処理待ちのプロセスを割り当て、割り当てた音声認識プロセスの状態を処理中に変える。音声認識プロセスが CGI プログラムに認識結果を返却し処理が完了すると、そのプロセスの状態は処理待ちに戻る。また、割当管理プログラムは音声認識プロセスが不意に終了した場合に、音声認識プロセスを自動的に再起動する。

音声認識プロセスは、CFG文法と木構造単語辞書に基づくワンパスの時間同期ビーム探索器[4]である。男女別のコンテキスト依存の連続確率密度分布音素HMMと雑音モデル[5]を使用している。音響ベクトルは、ES20110 とES202050 の音響特徴量に基づき、MFCC、ΔMFCC、ΔΔMFCC、Δpower、ΔΔpowerの 38 次元を用いている。発声単位でCMS[6]を行っている。

応答時間は平均で約 1 秒である。

5.2. サービス無停止辞書更新機能

HTTP を用いた分散型音声認識は、ウェブコンテンツやウェブベースアプリケーションにおける検索条件入力に適している。ネットワークのデータベースに日々追加される新しいキーワードを認識できるようにするには、音声認識の辞書・文法を頻繁に更新するこ

とが必要になる。更新のたびにサービスを停止するのでは頻繁な辞書更新が不可能になるため、サービスを停止せずに辞書を更新する機能を開発した。

1 台のサーバには複数の音声認識プロセスが起動し、割当管理プログラムが各プロセスの状態を管理している。この割当管理プログラムから、処理待ちの音声認識プロセスの一つを起動中に変えて、該当する音声認識プロセスを新しい単語辞書と文法で再起動する。これを起動している音声認識プロセス分だけ順番に繰り返すことで、サービスを停止せずに辞書を更新することができる。新しい単語辞書と文法は予め検証用サーバで動作確認され、メモリ上の単語辞書と文法がバイナリ辞書ファイルとして出力される。商用サーバはそのバイナリ辞書ファイルを読み込むことで、辞書更新時のエラーを回避し、辞書更新の時間を数十秒に抑えることができる。

6. おわりに

携帯電話アプリにおける日本語入力を支援するために携帯電話に実装した分散型音声認識システムを紹介した。ウェブコンテンツやウェブベースアプリケーションとの親和性を高めるため、携帯電話・サーバ間の通信には HTTP プロトコルを用い、携帯電話からサーバへは音響特徴量に加えて、アプリケーション、辞書の識別子を送信する。

音声認識に対するストレスや不安を軽減するため、認識性能に加えて、高速な応答性とユーザフレンドリーなインタフェースの開発に重点をおいた。携帯電話機上でのリアルタイム特徴量抽出、音声認識サーバにおける平均 1 秒間の認識処理により、発声終了から認識結果の取得までの時間を数秒に抑えた。誤認識の場合、認識結果がない場合にその理由がユーザに提供されない状況を改善するため「声が大きすぎます」、「雑音が大きすぎます」、「発声が早すぎます」の 3 種類のアラームを発生させる仕組みをクライアントソフトウエアに実装した。

また、ウェブコンテンツやウェブアプリケーションのデータベースに日々登録される新しいキーワードを認識できるようにサービスを停止せずに単語辞書と文法を更新する機能を開発した。

文　献

[1] ETSI ES201 108 v.1.1.2 distributed speech recognition; front-end feature extraction algorithm; compression algorithm. 2000.

[2] ETSI ES202 050 v1.1.1 STQ; distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms. 2002.

[3] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, T. Berners-Lee, "Hypertext Transfer Protocol -- HTTP/1.1, " IETF RFC 2616, June 1999.

[4] 安藤彰男, リアルタイム音声認識, (社)電子情報通信学会, 2003.

[5] T. Kato and T. Shimizu, "Noise-Robust Cellular Phone Speech Recognition Using CODEC-Adapted Speech and Noise Models," Proc. ICASSP, May 2002.

[6] 黒岩, 加藤, 樋口, "最ゆう状態系列を用いた実時間ケプストラム平均値正規化の検討," 電子情報通信学会論文誌, Vol. J82-D2, No.3, pp.332-339, 1999.