



2 携帯電話における分散型音声認識システムの実用化

加藤 恒夫
KDDI 研究所

携帯電話向け音声認識開発の背景

スマートフォンの普及が進む現在、携帯端末向けのアプリケーションが爆発的に増えている。音声認識関連のアプリケーションも、音声技術の進歩に加えて演算能力やメモリ量に対する制約が小さくなったことで、高度な技術に基づくものが次々と現れている。「Google 音声検索」はサーバを用いて検索エンジンの莫大なキーワード認識を実現している。旅行会話を対象とする音声翻訳も実用化されている。「NICT VoiceTra」,「東芝 ポケット通訳™ JE ver.0.1」はスマートフォンで動作する音声認識・翻訳・音声合成を組み合わせて、音声翻訳を実現している。NEC は携帯電話機上で動作する旅行会話自動通訳ソフトを発表している。「ATR-trek シャベって翻訳」はこれらより前の 2007 年からサーバを用いて旅行対話の音声翻訳を提供している。また、「NTT ドコモ 音声入力メール」,「Advance Media 音声認識メール」はサーバ処理によりメール本文の入力を可能にしている。

携帯電話は一般コンシューマ市場への登場から 20 年足らずの間に、移動時用・個人用の電話から個人の携帯情報端末へと急速に発展した。まずメール機能加わり、1999 年頃から携帯インターネットの端末となった。2001 年頃から登場した第 3 世代携帯電話においては、CPU の処理能力が上がり、アプリケーションの実行環境が整備された。メールやアプリケーションの機能拡充に伴い、携帯電話のテンキーは日本語テキストの入力に多く用いられるようになった。テンキーによる日本語テキスト入力は効率化が進んでいるが、たとえば歩行中の使用は危険であり、不慣れなユーザもいる。このような状況を背景に携帯電話向け音声入力の実用化が検討され始めた。

携帯電話によるインターネット接続の普及が始まった 2000 年代前半、携帯電話向けの音声認識はボイスダイヤリングのための組込み型音声認識と、古くからの電話音声認識であった。組込み型音声認識は携帯電話の演算能力とメモリ量の制約から、あらかじめ録音したテンプレートとのマッチングによるもの、携帯電話のアドレス帳から抽出した百種類程度の単語認識によるものに限られていた。電話音声認識はコールセンタの自動案内などに利用されていたが、誤認識の可能性のある認識結果を確認するために、

システム：「〇〇〇でよろしいですか？」

ユーザ：「はい」

というやりとりをする必要があり、ユーザにとって必ずしも快適なインタフェースではなかった。

携帯インターネットの日本語テキスト入力のために音声認識を利用できれば、ユーザは視認により瞬時に認識結果を確認することができ、誤認識が含まれる場合にはキー入力を使って必要な箇所だけ修正することができる。実現方法は、辞書の保守のしやすさの点からも、処理能力の点からもサーバで認識処理を行うのが最良の形態と考えられる。ただし、単純なシステム構成で、即座に応答が返るようにするには、従来の電話音声認識ではなく、音声の情報も認識結果もパケット通信を用いて携帯電話・サーバ間で交換する方式が望ましい。ちょうど 2000 年に欧州の標準化組織 ETSI (European Telecommunications Standards Institute) において、サーバクライアント型の音声認識を実現するための音響特徴量のフォーマット ES 201 108 が標準化された¹⁾。

KDDI は、この標準に基づく汎用的な分散型音声認識のクライアントモジュールを第 3 世代携帯電話に実装し、2006 年 1 月に「声 de 入力」という機能

2 携帯電話における分散型音声認識システムの実用化

名称でコンシューマ向け音声認識サービスを開始した²⁾。実用化にあたっては、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、ユーザフレンドリーなインタフェースと高速な応答性の開発に重点をおいた。

本稿では、携帯電話における分散型音声認識の開発に至った経緯とシステム構成、アプリケーション、ユーザインタフェースの工夫について紹介する。

携帯インターネットのための分散型音声認識システム

●電話音声認識との組合せによる音声入力プロトタイプ

2000年、国内の携帯電話台数は6,000万台を超え、携帯電話によるインターネット接続は爆発的に増加した。当時の代表的なWebサイトは、着信音サイト、ナビゲーションサイト、サイト検索のためのポータルサイトなどである。テンキーによる日本語入力が阻害要因となって携帯インターネットを利用しないユーザ向けに、まず従来の電話音声認識装置を応用する音声入力システムを開発した³⁾。電話音声認識を応用したのは、携帯ブラウザが持つ通話発信機能を活用し、音声入力サイト上に電話音声認識装

置とWebサーバ間で情報の受け渡しをする機能を開発するだけで、既存の携帯電話に対するファームウェアの改修や新たなソフトウェアの導入を行うことなく、音声入力機能を追加できるからである。この音声入力システムは、着信音サイト、ナビゲーションサイト、ポータルサイトなどに試験的に応用された。その一例が大阪梅田地下街のナビゲーションサイトである。地下街にある店舗を検索し、地下街地図中の位置を確認したり、現在地からの道順をガイドさせたりすることができる。同サイトは歩行中の利用が多いため、音声入力のメリットは大きい。

システム構成を図-1に示す。図中左のコンテンツサイトに対して音声入力機能を提供する音声入力サイト(図中右)を追加する。音声入力サイトは、コンテンツサイトとのインタフェースとなるWebページと、音声回線上の電話音声認識装置から構成される。コンテンツサイトでは、音声入力を必要とするWebページに音声入力サイトへのリンクを記述するとともに、音声認識結果に基づくコンテンツを表示するWebページを用意する。

音声入力サイトでは以下の手順で音声入力を実現する。

(1)ユーザが音声入力サイトにアクセスする。

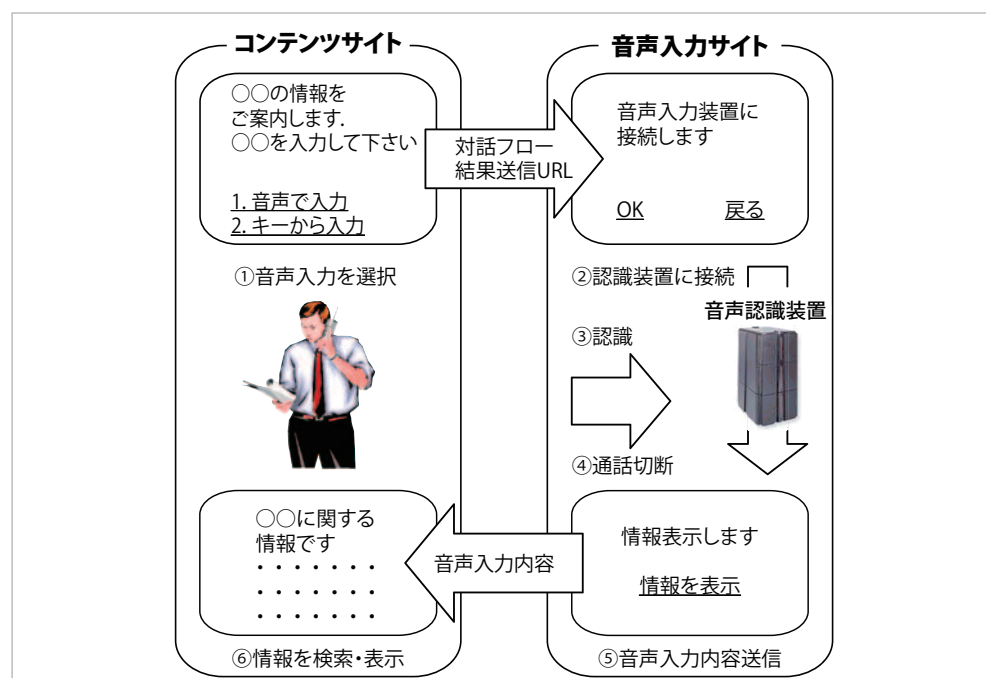


図-1 電話音声認識を用いた携帯インターネット用音声入力システム



- (2) 音声入力サイトから携帯ブラウザに対して電話音声認識装置への発呼を指示し、音声回線を通じてユーザと電話音声認識装置を接続する。このときパケット通信は切断される。
- (3) 電話音声認識装置とユーザ間の音声のやりとりにより音声入力の内容を認識する。
- (4) 認識終了後、音声回線を切断し、再度携帯インターネットに接続し、認識結果に基づくページ(コンテンツ表示ページ)へのリンクを記した音声入力内容送信ページを表示する。リンクには音声入力内容を表す文字列を追加する。
- (5) 利用者のキー操作によりコンテンツサイトに音声入力内容を送信する。
- (6) コンテンツサイトにおいて、音声入力結果に基づくコンテンツの表示などの処理を行う。

電話音声認識装置は不特定話者 HMM を用いる連続音声認識エンジンである。本装置は、都市部の街頭や屋外での移動中の使用が多いため、雑音対策として携帯電話音声に含まれる非定常な雑音を収集、分類し、雑音を音声と識別するための雑音 HMM を強化した。これにより実環境で収集した音声に対する誤認識率を約 40% 削減した。

本システムにより携帯インターネットへの音声入力が可能になったが、パケット通信から音声回線への切替、音声回線からパケット通信への切替にそれぞれ約 10 秒の時間を要した。そのため、音声回線を用いずパケット通信だけで音声認識装置とデータの送受信を行うことで、ユーザを待たせることのないシステムが必要との結論に至った。

●分散型音声認識システムの構築

《システム構成》

音声認識エンジンの一連の処理は、音声の入力側から①入力音声の取り込み、②音響分析による音響特徴量の抽出、③音響モデル・単語辞書・文法(もしくは確率言語モデル)に基づく探索、④最も確からしい単語系列の再評価、⑤認識結果の出力となる。分散型音声認識は、この一連の処理を端末とサーバに分散する。端末では、①と②を行い、さらに符号

化により音響特徴量をデータ圧縮してサーバに送信する。サーバでは、符号化された音響特徴量を受信して復号し、③、④、⑤を行い、認識結果を端末に返却する。

分散型音声認識のシステム構成を図-2に示す。携帯電話、XMLサーバ、音声認識サーバで構成される。音声認識サーバはコンテンツ Web サーバとは別にしている。携帯電話マイクへの入力から抽出された音響特徴量は XML サーバを経由して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 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 Web コンテンツや Web ベース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 サーバ間の通信には HTTP 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・辞書名などを含む XML が送信される。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 ドキュメントとして返却される。

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、端末機種によるフォーマットの違いの吸収などを行う。

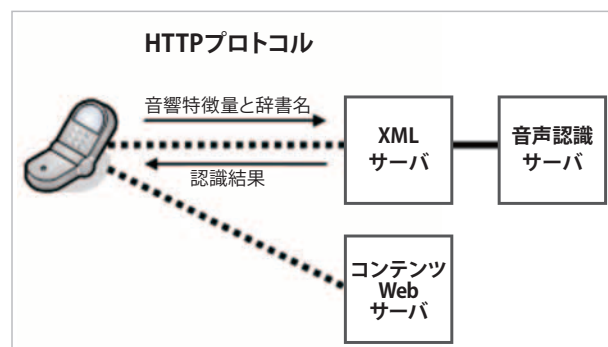


図-2 分散型音声認識システムの構成

《分散型音声認識の標準化》

ETSI のロバスト音声認識に関するワーキンググループ AURORA は 1997 年から 2002 年まで活動し、さまざまな端末から利用できる分散型音声認識プラットフォームの基盤としていくつかの音響特徴量を標準化した。1 つ目は 2000 年に標準化された ES 201 108¹⁾。2 つ目は、背景雑音や未知の周波数特性に対する頑健性を強化するために特徴抽出の前段にウィナーフィルタによる雑音抑圧、特徴抽出の後段にブラインド等価処理を加え、2002 年に標準化された ES 202 050 である⁴⁾。音響特徴量は 4.8kbps に圧縮される。認識率と低通信量を両立するため、この標準を採用することとした。

《分散型音声認識クライアントの携帯実装》

前述の①入力音声の取り込みと②ETSI 標準の音響分析および符号化処理を行う分散型音声認識クライアントモジュールを携帯電話に実装した。

2 種類の音響分析 ES 201 108 と ES 202 050 は、背景雑音の大きさによって使い分ける。さまざまな背景雑音で性能を比較した結果、背景雑音が小さい場合には ES 201 108 が ES 202 050 と同等の精度を示すこと、ES 202 050 は耐雑音性が強化された反面、計算量が多く、CPU 負荷が大きいことがその理由である。

音声入力終了後まもなく音響特徴量をサーバに送信できるように、入力音声の取り込み、音響分析、符号化をパイプライン処理化して携帯電話 OS のネイティブ層に実装した。さらに、音響分析部は整数演算化、レジスタ変数の参照効率化を行った結果、ES 201 108 の 3 倍の処理負荷を要する ES 202 050 もリアルタイム処理された。

この分散型音声認識クライアントモジュールは、BREW 携帯電話を提供する全メーカーの機種に実装された。そのため「声 de 入力」の機能は、サービス開始以降のほとんどの機種に搭載されている。

《音声認識サーバ》

音声認識サーバは、木構造単語辞書と文脈自由文法 (CFG) もしくは確率言語モデルに基づくワンパスの時刻同期ビーム探索器である。音響モデルとし

て男女別の不特定話者音響モデルと雑音モデルを使用している。

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、認識プロセスの割当管理プログラムが各プロセスの状態を { 起動中 / 処理待ち / 処理中 } の 3 状態で管理する。日々更新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辞書を更新する機能を備えている。

分散型音声認識のアプリケーション

●乗換検索サービス

分散型音声認識の最初のアプリケーションは、その利便性を多くのユーザに訴え普及促進を図れるものとして、300 万ユーザを抱えるナビゲーションアプリになった。その主要な機能である乗換検索サービスは、発駅、着駅と出発時刻もしくは到着時刻の指定に対して、適切な経路情報を提供する。GUI ベースの条件入力画面にテキストを入力し、サーバのデータベースを検索するサーバ・クライアント型のアプリケーションである。

このアプリケーションはユーザ数が多いだけでなく、以下の特徴を有している。

- ・テキスト入力の分量が多く、音声入力の利便性を実感できる。
- ・移動中に利用されることが多く、音声入力の利便性を実感できる。

そこで、このナビゲーションアプリのメニューに「声で乗換検索」の選択肢を加えた。「声で乗換検索」を選択すると、図-3 の画面に切り替わり、音声入力を受け付ける。乗換検索の発声パターンは、「【出発駅】から【到着駅】まで」を基本形として、「【出発駅】から【到着駅】まで { 今日 / 明日 / ○月○日 } の ○時 ○分に { 到着 / 出発 }」, 「【出発駅】から【到着駅】まで、{ いますぐ / ○分後に } { 到着 / 出発 }」, 「【出発駅】から【到着駅】までの { 始発 / 終電 / 最終 }」などの表現を受理する。これらの表現により複数のスロットに一度に検索条件が入力できる。日時指定の

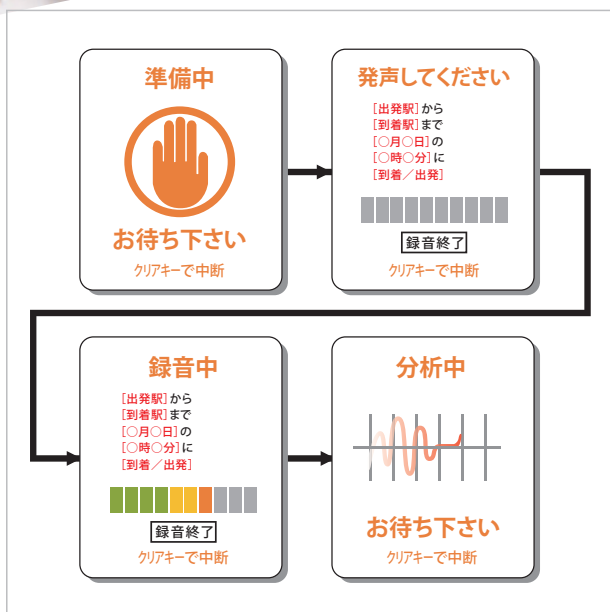


図-3 乗換検索における発声時の画面遷移

ない場合には「現在時刻に出発」がデフォルト検索条件として指定される。駅名は日本全国の9千数百種類が登録されている。

●目的地検索サービス

ナビゲーションアプリに目的地設定は欠かせない機能である。分散型音声認識は、歩行者用ナビゲーションアプリと自動車の同乗者用のナビゲーションアプリの2つの目的地設定に応用された。4種類の目的地設定によりメニューが分かれており、住所入力、電話番号入力、店名/施設名入力、駅/空港名入力ができるようになっている。

住所入力では全国の住所、店名/施設名では単独の店名/施設名の指定に加えて、「【地名/駅名】の【店名/施設名】」のパターンを受け付ける。

●社内電話帳

分散型音声認識を用いて、KDDI 研究所では3年以上、社内電話帳を運用している。「〇〇グループの〇〇さん」と発声すれば、認識結果とともにサーバに登録された電話番号とメールアドレスが携帯電話に通知され、そのまま電話をかけたり、メールを編集したりすることができる。図-4に画面遷移を示す。

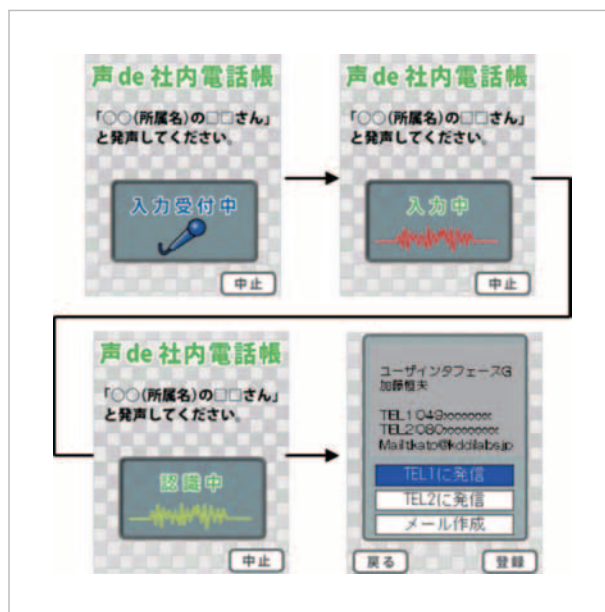


図-4 社内電話帳の画面遷移

ユーザは携帯電話のアドレス帳に社員の情報を登録する必要がなくなり、外出先であっても名前の発声だけでサーバのアドレス帳を検索することができる。ほかにも組織変更や社員の異動に対して、サーバ管理者が登録情報を変更するだけで、すべてのユーザが常に新しい連絡先を知ることができる。たとえ携帯電話を紛失したとしても、社内の連絡先リストを漏洩してしまうおそれがない、といったメリットがある。

ユーザインターフェースの工夫

分散型音声認識を用いた携帯インターネットやサーバ・クライアント型のアプリケーションでは、従来の電話音声認識と異なり画面表示を利用できるため、発声の仕方を伝えたり、認識結果を瞬時に確認させたりすることができる。ここではユーザ、特に初めて音声認識を利用するユーザにとっても使いやすさを感じさせるための工夫を紹介する。

●発声パターンの表示

初めて利用するユーザは、音声入力の開始を合図するプロンプトトーンに対して何を発声してよいか分からないことも多い。前述の乗換検索の場合、「発

2 携帯電話における分散型音声認識システムの実用化

声してください」の表示の下に基本パターンを表示した。これにより、不慣れなユーザは発声パターンを目で確認しながら発声することができる。

●誤認識原因の推定・表示機能

音声認識は、背景雑音、語彙外の発声内容、言い誤りや言い淀み、音節単位に区切った発声、音響モデルとのミスマッチなどさまざまな原因により、誤認識を生じたり、認識結果を得られないことがある。従来の音声認識では認識結果が得られない場合に「認識できません。もう一度発声してください」のような応答を返すものが多かった。ユーザはなぜ認識結果が得られなかったのか理由が分からないために、同様の発声を繰り返し、同じエラーメッセージを繰り返し確認する状況に陥り、最終的に音声入力を諦めてしまうことも多いと考えられる。

こうした状況を改善するために、認識結果が得られなかった場合や誤認識の可能性が高い場合に、その原因を可能な範囲で推定し、表示によってユーザに伝えるようにした。推定できる現象は、(1) マイク入力のオーバーフロー、(2) 過大な背景雑音、(3) 話頭切断、の3種類である。

(1) マイク入力のオーバーフローは、デジタル信号の最大値／最小値でクリッピングされたサンプルをカウントし、閾値処理を行うことで検出する。オーバーフローを検出すると、「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」と表示する。(2) 背景雑音はプロンプトトーンの前の数百ミリ秒で取得している。背景雑音レベルが閾値よりも高い場合に、「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」と表示する。(3) 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した場合に発声の先頭が収録されないことを指し、発声先頭の欠落によって認識率の低下を招く。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顕著に高くなりやすい。これを利用して、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って話頭切断を検出する。検出された場合、「発声が早すぎます」と表示する。

●認識処理の短時間化

応答の速さも、ユーザが感じる使いやすさに強く関係している。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ために、音声認識サーバの認識処理は平均1秒で行っている。

音声認識では、各時刻の入力音声に対する音の種類が確定的でないために、探索途中では同時に多数の仮説を保持することによってさまざまな認識結果に至る可能性を残している。認識精度と高速処理の両立を図る手法として、各仮説から到達し得る単語の数により仮説の保持／棄却の判定基準を変化させることで必要な仮説数を大幅に削減する手法を開発した。これにより認識精度の劣化なく、従来に比べて3倍以上の高速化を果たした。

●修正しやすい確認画面

分散型音声認識を用いたアプリケーションでは、画面表示により認識結果をユーザに瞬時に確認させ、必要に応じて部分的に修正することが容易にできる。

前述の乗換検索の場合、出発駅、到着駅、日付、時刻、到着／出発といったスロットごとに認識結果を表示している。認識結果に含まれる2位以下の候補(N-best出力)はプルダウンメニュー式に選択できるようにし、同時にプルダウンメニューの1つとしてキーによるテキスト入力を加え、部分的な修正を可能にしている。また、スロットごとに音声入力のボタンを配置し、音声入力でも部分的な修正ができる。

携帯電話向け音声認識のこれから

分散型音声認識の仕組みによって携帯インターネットやネットワークのサーバと連携した携帯アプリケーションの日本語テキスト入力に音声入力ができるようになった。冒頭に述べたように、現在で



特集 音声認識技術の実用化への取り組み

はサーバ処理によりメール本文の入力を行うサービス、インターネット検索の膨大なキーワードを認識するサービスも現れている。個別のアプリケーションでは音声入力ができるようになったが、先に紹介したアプリケーション例では、音声入力に至るまでにキー操作で起動するアプリケーションを選択したり、アプリケーションの中で音声入力のメニューを選択したりする必要がある。

こうした携帯電話の機能呼び出しや、携帯電話のアドレス帳検索を可能にするため、携帯電話単体で動作するローカル音声認識エンジンをすでに開発している⁵⁾。ローカル音声認識エンジンによって、たとえば「乗換検索」、「社内電話帳」などの発声でアプリケーションを起動したり、「アラーム1を午前7時にセット」のように携帯電話機能を設定したり、「〇〇さんにメール」のようにアドレス帳の人物宛のメール編集画面を表示したりすることができる。アドレス帳の登録変更などに即座に対応できるように、携帯電話上で単語辞書と文法を更新することができる。このように携帯電話の機能呼び出しやアドレス帳検索の場合、音声認識辞書のもとになるデータが端末にあるので、認識処理も端末で行われるのが最適な形態である。ローカル音声認識エンジンは、音

響モデル・単語辞書・文法を用いた探索処理を携帯電話の限られたメモリで高速に実行するため、省メモリ・高速処理を徹底した設計となっている。

ローカル音声認識と分散型音声認識を組み合わせれば、携帯電話自体の機能はローカル音声認識で呼び出し、ネットワークにある情報は分散型音声認識で検索することが可能になる。従来のキー操作に加えて、音声入力による一貫した操作インタフェースが実現できると考えている。

参考文献

- 1) ETSI ES 201 108, ETSI ES 201 108 Distributed Speech Recognition Encoding (Feb. 2000).
- 2) 加藤他:分散型音声認識の商用システム構築, 情報処理学会研究報告, 2006-SLP-63, pp.39-44 (2006).
- 3) 内藤他:携帯インターネット用音声入力システムにおける雑音対策, 情報処理学会第66回全国大会講演論文集, pp.217-220 (Mar. 2004).
- 4) ETSI ES 202 050, ETSI ES 202 050 v1.1.1 Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms, ETSI (Apr. 2002).
- 5) KDDI 声 de 入力 ローカル音声認識タイプ, http://www.kddi.com/corporate/news_release/2008/0925c/ (Sep. 2008). (平成22年9月7日受付)

加藤 恒夫 (正会員) tkato@kddilabs.jp

1996年東京大学大学院工学系研究科修士課程修了。同年、国際電信電話入社。音声認識、話者認証の研究開発に従事。現在、ユーザインタフェースグループグループリーダー。