音声認識 スマートフォン ユーザインタフェース

Technology Reports

端末機能やサービスの利便性向上のための音声認識技術とアプリケーション開発

移動端末のような小型デバイス上で，複雑な操作をより直感的かつ低負担に行う手段として，音声入力を用いたユーザインタフェースが注目されている．ドコモにおいても，音声による文字入力や端末機能呼出しを提供しており，将来的には，「話せば何でも応えてくれる」ユーザインタフェースの実現を目指している．今回，その実現に向け，音声認識性能を向上させるために，大量データに基づく大語彙化を実現し，有効性を確認した．また，音声認識結果に言語処理を適用し，ユーザの操作をサポートするアプリケーション開発の事例を紹介する．

先進技術研究所 飯塚 真也 辻野 孝輔
サービス＆ソリューション開発部 小栗 伸
移動機開発部 古川 博崇

1. まえがき

近年，移動端末では，電話やメールのようなコミュニケーションツールとしての基本機能以外にも，付加機能やアプリケーションによる多様なサービスが利用可能である．例えば，旅行の宿泊先や交通機関の予約・決済，観光地での地図の確認や記念写真のweb上での共有まで，携帯電話1台で行うことができる．

一方，機能やサービスが拡充・高度化されるほど，ユーザにとってはこれらを効果的に利用するためのスキルや複雑な操作が必要となる．具体的には，自分の知りたい情報や受けたいサービスにアクセスするための方法を把握し，状況に応じた適切な使い分けや詳細な設定操作が要求される．

このような複雑化する操作に対し，移動端末のような小型デバイス上で低負担なユーザインタフェースの提供を目的とし，筆者らは音声入力について研究開発を実施している．音声入力は，複雑な階層構造や複合的な条件についても，ダイレクトに指定することが可能なため，操作をより低負担で実現する手段として注目されている．

近年普及の進むスマートフォンでは，音声入力を利用したアプリケーションが数多く提供されている．ドコモにおいても，音声による文字入力や端末機能呼出しを提供しており，将来的には**図1**に示すような「やりたいことを話せば何でも応えてくれる」ユーザインタフェースの実現を目指している．想定する実現イメージは，移動端末が発話の意図を理解し，ニーズに適合したソリューションをワンストップで提供するもので，ユーザは複雑な操作を必要としない．この実現に必要な技術領域は，ユーザの発話した音声をテキストに変換する音声認識技術と，テキストの意味を解釈し，アプリケーションとしての応答を決定する言語処理技術からなる．

本稿では，音声認識技術の概要，および音声認識技術と言語処理技術を利用した開発事例について解説する．

2. 音声認識技術

2.1 音声認識の方式と特長

音声認識の方式には，端末内音声認識とサーバ型音声認識が存在す

る．端末内音声認識は，端末内で音声認識エンジンを動作させる方式であり，サーバ型音声認識は音声信号または音声特徴量[*1]をサーバに送信し，サーバ内の音声認識エンジンを動作させた後，認識結果の文章を端末へ返送する方式である．

端末内音声認識は，処理装置や消費電力の制約により比較的小語彙での音声認識に限られるが，通信状況による遅延や圏外の影響を受けないため，端末操作のような限定用途だが常時利用を要求されるアプリケーションに適している．一方，サーバ型音声認識は通信状況の影響を受けるが，演算量を比較的必要とする音声認識に対応可能である．このため，検索や文字入力のような大語彙対応が必要とされるアプリケーションに適している．

2.2 音声認識の高度化

音声認識システムの典型的な構成を**図2**に示す．入力音声信号に対し周波数特性分析による特徴量抽出処理を適用し，特徴量を音声認識エンジンに入力する．音声認識エンジンは，蓄積済みデータからあらかじめ学習された音響モデルおよび言語モデルと，入力された特徴量とを比較照合し，最も尤度の高い形態素[*2]列を認識結果として決定する．音響モデルは音声の特徴量と音素[*3]（個々の母音・子音）との対応関係を示すモデルであり，言語モデルは形態素の前後間のつながりやすさを示すモデルである．

音声認識の性能は，これらのモデルをいかに実利用環境での入力と近い条件で学習できるかの再現精度によって決定される部分が大きい．すなわち，音響モデルの学習は，実際のユーザの音声の特徴を反映することが重要である．一方，言語モデルの学習においては，幅広い発話内容の認識に対応するため大語彙を網羅する必要がある．このため，大語彙言語モデルを構築するためには，大量のテキストデータを学習対象とする必要がある．

図1　想定イメージ

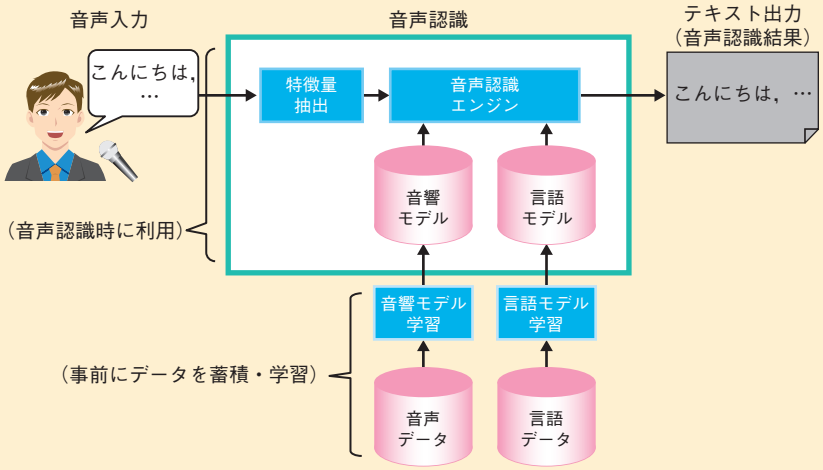

図2　音声認識システムの構成

＊1 **特徴量**：音声波形から音声認識に必要な情報のみを取り出したもの．特徴量として，音声波形を高速フーリエ変換（FFT：Fast Fourier Transform），聴覚特性フィルタバンク等で処理して得られるメル周波数ケプストラム係数（MFCC：Mel-Frequency Cepstrum Coefficient）を用いることが多い．

＊2 **形態素**：ある言語において，それ以上分割できない，意味をもつ最小の単位．単語に加え，丁寧語の「お」などの接頭辞や接尾辞などを含む．文を自動的に形態素に分割する処理を形態素解析と呼ぶ．

＊3 **音素**：言語における意味の弁別に用いられる最小の音の単位．

筆者らは，数十万語彙の言語モデルの構築を行い，大語彙化による認識性能の向上を確認した．構築にあたっては，多様な表現が混在する大量のテキストデータの，言語としての構造化の精度が課題であった．そのため，テキストデータのふるい分け，形態素解析境界の最適化，形態素への読み仮名付与の自動化等を行うことで精度を向上させた．

3. アプリケーション

3.1 音声機能呼出アプリケーション

2011年夏モデルでは，2010年秋冬モデルのiモード携帯電話で実現した音声機能呼出アプリケーション[1]の技術をAndroid™＊4 OS搭載スマートフォンに展開するとともに，主に，ユーザがダウンロードしたアプリケーションも起動対象にすること，ユーザの意図に則した機能呼出ができることの2点に焦点を当てて，機能拡張を行った．端末機能呼出アプリケーション「しゃべってカンタン操作」の動作フローを**図3**に示す．

音声機能呼出アプリケーションは，端末内に存在するメニュー名，アプリケーション名にあらかじめ発話キーワードという読み仮名を付与することで，端末内音声認識の認識結果と合致する発話キーワードに紐付く機能IDに置換して，該当機能を起動する．

スマートフォンでは，ユーザが任意のアプリケーションを端末内に取り込み，利用することが前提となるため，従来のように一意に決められたメニュー名，アプリケーション名を事前に登録しておくだけでは，端末内にあるすべての機能を起動するには不十分となる．

そこで，スマートフォン向けの拡張として，音声機能呼出アプリケーションを起動する際に，端末内に存在するアプリケーションの情報（アプリケーション名，パッケージ名）を確認し，随時追加・削除をすることで，端末内の最新の状態に合ったアプリケーションリストを都度生成し，後からダウンロードしたアプリケーションも起動できる仕組みを開発した．

ただし，アプリケーション名については，取得したアプリケーション名称の文字列に対して形態素解析を行うことで，ある程度の読み仮名をシステムが自動的に付与することは可能だが，形態素解析の特性上，英数字や略語，語呂合わせのような文字列に対して正しい読み仮名を付与することはできない．また，ユーザが呼び慣れている任意の名称での呼び出しができないという課題があった．

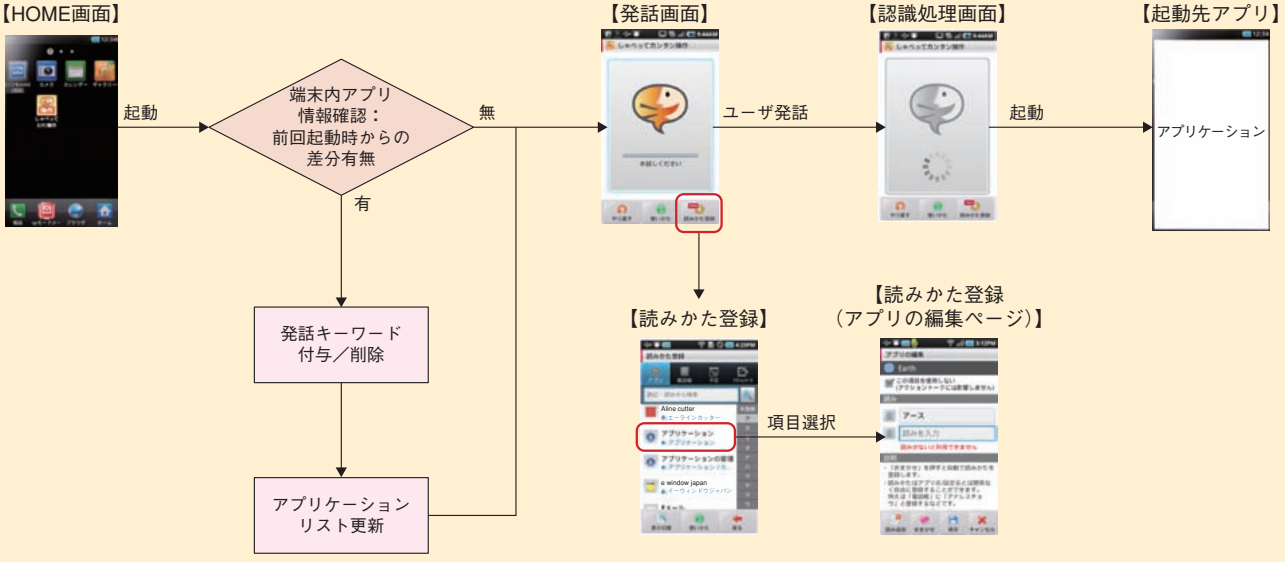

図3　端末機能呼出アプリケーションの動作フロー

＊4　Android™：米国Google, Inc.が提唱する携帯端末を主なターゲットとしたオープンソースプラットフォーム．Android™は米国Google, Inc.の商標または登録商標．

この課題を解決するため，アプリケーション名と付与されている発話キーワードとを対応付けた一覧画面「読みかた登録」を音声機能呼出アプリケーション内にもたせ，読み仮名をユーザが編集できる仕組みを設けた．読み仮名欄は，付与されている発話キーワードを編集できるだけでなく，1つのアプリケーションに対して，複数の読み仮名を追加登録することも可能とすることで，より柔軟で広範な発話キーワードを受付けられるようにした．

本開発により，スマートフォンの利用形態に則して，ユーザが自由に端末内アプリケーションを音声で呼び出す機能を実現した．

3.2 統合型ユーザインタフェースアプリケーション

移動端末の操作に加え，各種のwebサービスに横断的かつシームレスにアクセスを可能とする統合型ユーザインタフェース実現に向け，発話内容と関連する端末機能やサービスのカテゴリを推定する言語処理技術を開発した．また，開発した技術を利用しAndroid OS搭載スマートフォンで動作するアプリケーション「VOICE IT!」というソフトウェアの試作品（プロトタイプ）を作成し，2011年5月にトライアル提供した．

一般的なweb検索はインターネット上に存在するあらゆる情報やサービスにアクセスできる一方，ユーザは所望の結果を一覧の中から探し出す必要がある．このため，移動端末のように比較的小さな画面のデバイスでは，操作に負担を感じさせるケースがある．

そこで開発したVOICE IT!では，大語彙に対応したサーバ型音声認識を利用するとともに，アプリケーションとして次のような設計を実施した．

①端末機能やwebサービスのカテゴリごとに特化した個別アプリケーションを呼び出し可能とする．

②言語処理技術により，ユーザの発話がどのカテゴリを所望したものかをランキング形式で自動判定し，該当する個別アプリケーションを提示する．

③発話内容に関連する別カテゴリの個別アプリケーションにも簡単にアクセスできる画面構成とする．

上記により，ユーザはどのような端末機能やwebサービスを起動すべきかを探すことなく，やりたいことや知りたいことを発話するだけで，所望の個別アプリケーションに素早くアクセスすることができる．

図4にVOICE IT!アプリケーショ

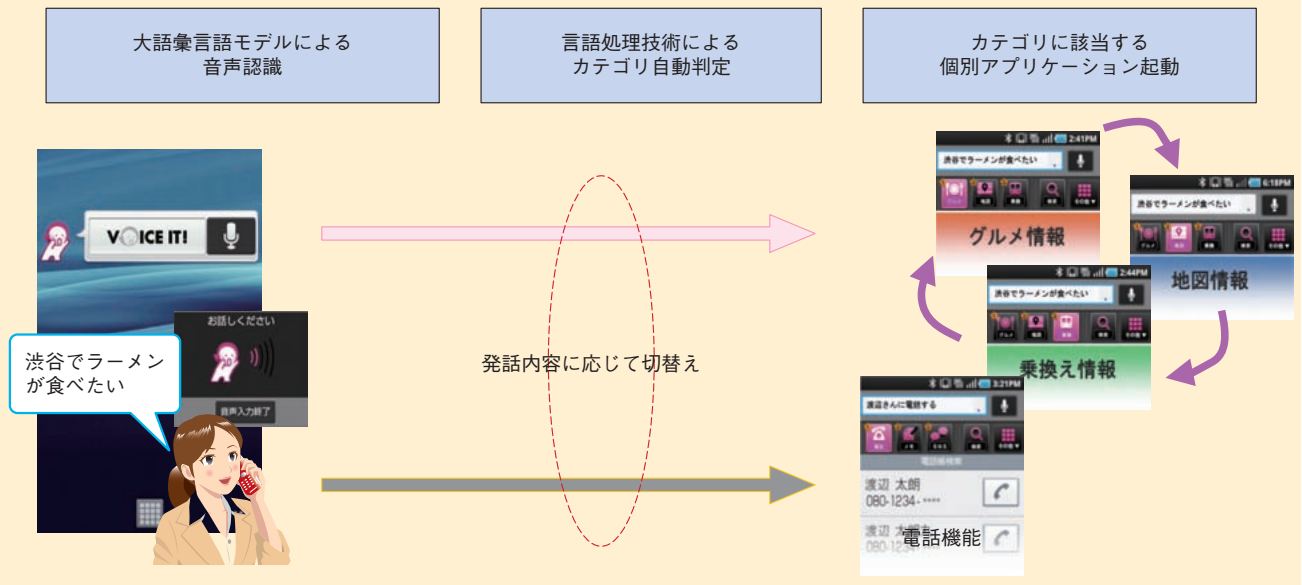

図4　VOICE IT! アプリケーションの画面遷移

ンの画面遷移イメージを示す．例えば，「渋谷でラーメンが食べたい」と話した場合，グルメ情報サービスでの，渋谷周辺のラーメン店情報が提示される．さらに，渋谷周辺の地図情報や渋谷までの乗換え情報を知りたい場合には，該当するアイコンを押下することで切替え可能である．プロトタイプでは，**表1**のジャンルの，個別アプリケーションを具備した．

VOICE IT!のトライアル提供による知見をもとに，今後，より利便性の高いユーザインタフェースの実現に向けて取り組む予定である．

4. あとがき

本稿では，「話せば何でも応えてくれる」ユーザインタフェースの実現を目指した音声認識技術およびアプリケーションの開発について解説した．今後は，提供サービスに応じた音声認識性能のさらなる向上，およびテキスト化された発話内容をより柔軟かつ高度に理解可能な言語処理技術を検討する予定である．

表1　VOICE IT! カテゴリー覧

分類	カテゴリ
webサービス	web検索
	レストラン検索
	トラベル検索
	レシピ検索
	動画検索
	音楽検索
	書籍検索
	ショッピング検索
	アプリケーション検索
	ブログ検索
	乗換え検索
	地図
	天気
	ニュース記事検索
	辞書
	百科事典
	Q＆A検索
端末機能	電話
	メール
	メモ
	カメラ

文　献

[1] 古川，ほか：“2010-2011冬春モデル搭載アプリケーション機能―進化しつづけるケータイアプリケーション，” 本誌，Vol.18, No.4, pp.17-24, Jan. 2011.