身近になった音声処理技術と東芝の取組み

Speech Processing Technologies Becoming Common in Daily Life, and Toshiba's Approach

赤嶺　政巳
■AKAMINE Masami

キータッチの代わりに自分の声で機器を操作する音声インタフェースの利用が身近なものとなった。

東芝は，1980年代から音声処理技術の研究開発を行っており，音声インタフェースを支える各種の基盤技術を長年，研究し開発してきた。その成果は，カーナビなど組込み機器向けの音声ミドルウェアや，パソコン（PC）のソフトウェア，Web上での音声コンテンツサービス，文書情報の機械翻訳システムなどの製品に活用されている。更に，単に音声インタフェースにとどまらず気の利く個人秘書のようなコグニティブアシスタントの実現を目指し，音声処理技術の高度化及び高精度化だけではなく関連技術の開発と利用者に新たな価値をもたらす製品及びサービスの開拓と開発に努めている。

Speech interfaces have recently become increasingly widespread for interacting with digital devices such as smartphones instead of touch keyboards.

Since the 1980s, Toshiba has been developing various core technologies supporting speech interfaces, such as automatic speech recognition, speech synthesis, and so on. These technologies have been applied to a variety of products including speech middleware for in-car navigation systems, dictation software, content-creation services on websites, and machine translation systems. Aiming at the realization of the so-called cognitive assistant, we have been continuously engaged in the development of not only speech technologies but also technologies related to multimodal interfaces and various new products and services.

身近になった音声インタフェース

音声インタフェースは，ボタンを押したりキーボードをたたいたりする代わりに，自分の声で機器を操作するインタフェースの総称であり，安全性の観点からハンズフリーやアイズフリーが必須のカーナビなどの組込み応用で利用されてきた。最近では，ネットワークの高速化を含めたクラウドシステム技術の発展に伴い，サーバと連携してスマートフォンで利用できる音声インタフェースが利用者の関心を集め，より身近なものとなった。

例えば，カーナビに音声で「東京駅」と行き先を入力すると，運転中はカーナビが「100ｍ先，XX交差点を右方向です」と音声で道案内する。スマートフォンに「あすの朝10時から会議」と話しかけると，スケジューラが起動され，「あすの朝10時，会議をセットしました」と音声で知らせてくれる。また，自分の声で知りたい情報を簡単に検索することも，海外旅行先で日本語で話しかけた内容を英語や中国語など他の言語に翻訳し，合成音声で提示することもできるようになった。

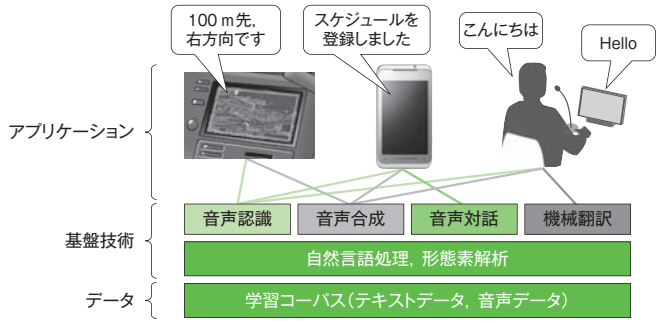


図1．音声インタフェースを支える基盤技術 ― 音声インタフェースは，基盤技術，アプリケーション，及びデータの階層的な構成で実現されている。
Core technologies constituting speech interfaces

音声インタフェースが身近になったのは，CPUやメモリなどのハードウェア技術の進歩，ワイヤレスネットワークの高速化，クラウドシステム技術の進展を背景に，いつでも，どこでも簡単かつ楽に機器を利用したいという利用者の要求をある程度満足できるレベルにまで音声処理技術が進歩し，高度化したためと考えることができる。

東芝は，音声処理技術の研究開発を1980年代に開始し，音声インタフェースを支える各種の基盤技術を長年，研究，開発し，蓄積してきた。その成果は，カーナビなど組込み機器向けの音声ミドルウェアや，PCのソフトウェア，Web上での音声コンテンツサービス，文書情報の機械翻訳システムなどの製品

に活用されている。

音声インタフェースを支える音声処理の基盤技術

図1に示すように，音声インタフェースは音声認識や音声合成など音声処理の複数の基盤技術で支えられている。

この中で形態素解析技術は，かな漢字交じりの日本語文章を単語ごとに分割し，品詞情報を特定する処理である。形態素解析は，音声処理におけるもっとも基本的な処理であり，音声認識や音声合成，機械翻訳において音響モデルや言語モデルの作成，読み情報の付加や構文の解析のために用いられる。当社は，世界で初めて(注1)実用化した日本語ワープロの開発初期から長年にわたって形態素解析の高度化に取り組み，音声認識や音声合成，機械翻訳の処理に応用してきた。

この特集では，音声認識（p.6－9）や音声合成（p.10－13）だけではなく，所望の信号を確実に取り入れるためのマイクロホンアレイ及びノイズキャンセル技術（p.22－25）や，音源分離技術（p.26－29），対面業務向けの同時通訳技術（p.18－21），音声理解や対話の知識処理技術（p.14－17）など音声処理の幅広い技術と，製品応用（p.30－33，p.34－37）について当社の取組みを述べる。

ここでは，音声インタフェースを構成する音声認識と音声合成の基盤技術について，実用化の課題とその解決に向けた取組み，及び今後の展望を述べる。

実用的な音声認識技術への取組み

カーナビやスマートフォンを利用する環境下で心地よく音声入力機能を利用するためには，街角や車中などの高雑音環境下でも90％以上の単語認識精度が必要である。また，「あすの天気を教えて」のように対話的に音声を利用できるようにするためには，連続音声認識技術が不可欠であり，更に友人に話しかけるような砕けた表現に対応するためには話しことばの音声認識が必要である。

音声認識の主な応用分野における使用環境の雑音レベルと音声認識の対象を**図2**に模式的に示す。

音声認識の対象が孤立的に発声される単語から連続的に発声されるフレーズや文，話しことばになるに従って，音声認識機能を利用する利用者にとって，認識対象の単語を正確に発声する必要がない，文章を読み上げるような話し方をする必要がない，ふだんどおりに話しかけるように話せばよいなど，制約が少なくなり使いやすくなる。しかし，音声認識にとって，認識対象が孤立単語から連続発声された文や話しことばになるに従い難しいタスクとなり，認識の精度も低下する。

一方，音声認識機能の利用環境が静かなオフィスから街角や車中などへ雑音レベルが高くなるに従って，音声認識の精度は低下する。音声認識機能をどんな環境でも利用できるようにするためには雑音に頑健な高精度の音声認識技術が必要である。

当社は，研究開発を図2の右上方向に進めており，これまでに雑音に頑健な音声認識技術と，連続音声認識の技術を開発してきた。

まず，カーナビなどの組込み機器向けに，雑音環境下でも発話音声区間を高精度に抽出できる音声区間検出方式の開発(1)，音素レベルの識別能力を最大化する競合学習方式の開発，及び雑音環境下で高い認識性能が得られる少ない計算量の認識方式を開発した。これらの方式は音声認識ミドルウェアとして大手自動車メーカーのカーナビ向けに2005年にライセンス供給を開始し，翌年からは当社のBluetooth®(注2)チップに組み込み，北米市場向けの車載ハンズフリー通話用として，自動車メーカーに出荷している。このチップにより運転中の音声ダイヤル機能を実現し，米語，カナダフランス語，及び米スペイン語の3言語に対応している。

カーナビなどの車載応用のほか，テレビなど情報家電向けの次世代インタフェースにも取り組んでおり，音声リモコンによるテレビ番組検索システムを開発している（この特集のp.30－33参照）。

また，連続音声認識において，耐雑音性と高精度を実現するため，雑音抑圧及び特徴量抽出の各技術と言語モデルの高精度化，並びにクラウドソーシング

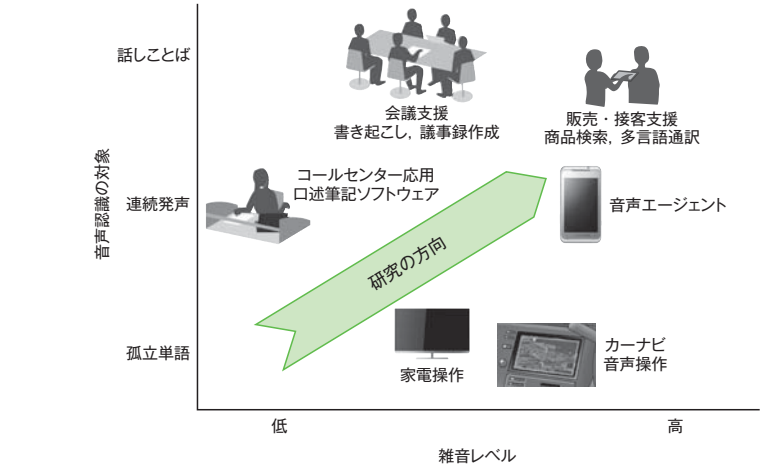


図2．音声認識の応用分野における雑音レベルと認識対象 ― 音声認識の応用分野によって利用される環境の雑音レベル及び音声認識の対象は異なり，研究は雑音に頑健な話しことばへと進んでいる。
Relationships of automatic speech recognition applications, environmental noise levels, and recognition tasks

（注1）　1978年9月時点，当社調べ。

（注2）　Bluetooth®ワードマーク及びロゴは，Bluetooth SIG, Inc.の登録商標。

による語彙収集に取り組んでいる。更に，クラウドシステム上で音声認識，音声合成，及び機械翻訳の各エンジンを個別に，又は統合して提供するWebサービスソリューションの開発も進めている。

音声合成の高音質化，多様性実現への取組み

音声合成は，人がことばを発するように音声を人工的に生成するための技術であり，音声認識技術とともに音声インタフェースを実現するうえで不可欠な基盤技術である。音声合成技術開発の歴史は古く，米国マサチューセッツ工科大学（MIT）で開発されたフォルマント合成に基づく“Speak&Spell”が米国テキサスインスツルメンツ社で商品化されたのは1978年のことである。Speak&Spellは英単語のスペルを正しくキー入力するとその単語を発声する教育玩具である。人が聞いて何を言っているかがわかるという観点では，教育玩具として必要な要件は備えていた。しかし，合成音声の音質は鼻にかかったこもった音で人の音声には程遠く，抑揚のないロボット的な話し方で，合成音の品質は肉声感及び韻律の観点から十分なものではなかった。その後，高音質化と自然性向上の取組みが長年行われてきた。

高音質化の取組みとして，1990年代から，波形編集方式とコーパスベース方式の検討が盛んに行われるようになった。波形編集方式は，“あ”，“い”のような音節単位の音声素片と呼ばれる短い音声波形の辞書をあらかじめ作成し，音声合成時に，入力されたテキスト中の音節に対応する音声素片を音声素片辞書から選択し，その韻律を合成したい抑揚に合わせて変更処理した後，接続して音声を生成する。コーパスベース方式では，大量の音声データをあらかじめ収録し，そのスペクトルや韻律情報，音韻情報とともにコーパスと呼ばれるデータベースに蓄積しておく。合成時に，入力されたテキストに対して最適な音声素片を，ある尺度に基づいてコーパスから選択し，そのまま接続して音声波形を生成する。

このように，合成音声の波形生成は音声素片の処理で行われるため，その音質は音声素片辞書の作成法に大きく左右される。従来の波形編集方式では，音質のよしあしを技術者が評価しながら音声素片を試行錯誤的に作成していたため，ノウハウを熟知した熟練者による長時間の調整が必要で，音質に限界があるとの問題があった。一方コーパスベース方式では，コーパス中の音声波形を単純に接続することで韻律変更のための信号処理を避けて音質を改善している。しかし，この方式は大規模なメモリと，大規模なデータベースを探索処理するための高性能なプロセッサが必要であり，メモリ容量に制約のある機器や処理能力の低い機器では実現が困難だった。

当社は，これらの課題に対して“音声素片の閉ループ学習方式”とこれを韻律学習に拡張した“自然韻律パターンの閉ループ学習方式”という二つの新技術[2),3)]を開発し，これらの学習技術によって得られた韻律と音声素片を用いて音声を合成する音声合成技術を開発した。閉ループ学習の根本原理は，合成音声のひずみの定式化とそのひずみを最小化する音声素片及び韻律パターンの自動生成である。閉ループ学習方式により，長年問題となっていた“鼻声”や“ロボット声”を解消するとともに，技術者のノウハウに頼らず音声素片辞書の自動作成ができるようになった。この技術は，カーナビやビデオゲーム，電子辞書などに広く採用されている。

音質の向上に伴い，インターネット上での音声コンテンツ作成や情報提供，ビデオゲームなどでは，親しみのある口語的かつ対話的な話し方や，キャラクター性のある声質の合成音が求められるようになっている。また，楽しい，悲しい，腹立たしい，いらいらするなどの感情に応じて抑揚や話し方，声質を自由に制御できる音声合成が望まれている。

当社はこれらの要求に対して，新しい音声合成エンジンToSpeak V2を開発した。このエンジンは，録音された音声から話者の声質や抑揚，発話スタイルの特徴を精度よく学習することができ，性別だけでなく落ち着いた声や若い声，子どもやキャラクターの声，有名人の声など様々な声の辞書をそろえている。

更に，音声合成の利用者からは自分の声でホームビデオにナレーションを付けたいという要望もある。この要求に対して，30分の音声を録音すれば，自分の声の音声合成辞書が作成できる全自動カスタム音声合成作成システムを開発し（この特集のp.10－13参照），Web上で一般に公開している。

多様な合成音声を実現するため，隠れマルコフモデルに基づく音声合成（**囲み記事参照**）の開発にも取り組んでおり，感情に応じて抑揚や声質を自在に制御できる音声合成方式を開発している[4)]。

今後の展望

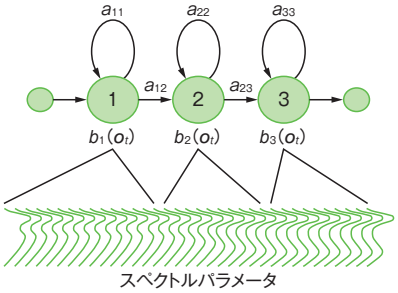
音声インタフェースのスマートフォンへの搭載は，同インタフェースが“便利で使える”との認識を一般の利用者に与えたと思われる。しかし，現在の音声インタフェースは極論するとボタンやキー操作を音声操作に置き換えているにすぎない。すなわち，あすの天気を知りたい，スケジュールを登録したいなど，利用者の明確な意思のある発話をシステムが認識し，アプリケーションを起動したりWebを検索したりしている。現状では，無線LANが接続できなくなり「インターネットに接続できない，困った」とスマートフォンに話しかけても，問題解決に向けたアドバイスは期待できない。また，午後3時の休憩時間に近くのカフェへの案内を期待して「ちょっと腹が減った」と話しかけると，「フレンチのフルコースを予約しました」と，かってなことをされるかもしれない。

隠れマルコフモデルによる音声合成

隠れマルコフモデル（HMM：Hidden Markov Model）は**図A**に示すように，任意の確率を持って遷移する複数の状態と各状態から出力される観測ベクトルの確率分布によって定義される。

これらの状態遷移確率と各状態の確率分布（モデルパラメータ）は，あらかじめ収集された大量の観測ベクトルから学習される。例えば“あ”に対応する音声波形が大量に与えられた場合，10 ms程度の一定フレームごとに音声波形から計算されたスペクトルパラメータを観測ベクトルと設定し，この観測ベクトルがHMMのモデルから生成される確率（尤度（ゆうど）と呼ぶ）を最大とするよう，モデルパラメータが求められる。

音声合成は，入力されたテキストに対応するスペクトルパラメータと音声波形を，テキストに対応する音素系列のHMMモデルから生成する。まず，HMMモデルから生成される観測ベクトルの尤度が最大となるようスペクトルパラメータを生成する。尤度を最大化する観測ベクトルは確率分布の平均ベクトルとなることから，生成されるパラメータは分布の平均ベクトルとなる。ただし，状態遷移ごとに分布は変わるので，生成されるパラメータは状態間で不連続となり，音質が劣化しする。この問題を避けるため，観測ベクトルのフレーム差分を動的特徴としてモデル化し，動的特徴も考慮してパラメータを生成する。



1，2，3：状態
a_{ij}：状態iからjへの遷移確率
$b_i(o_t)$：状態iから出力される観測ベクトルo_tの確率分布

図A．隠れマルコフモデル

今後は，単純な音声インタフェースから更に進め，気の利く個人秘書のようなコグニティブアシスタントの実現を目指したい（**図3**）。高級ホテルのコンシェルジュのように，時間と場所を含む利用者の状況や，時に利用者自身が自覚していない意図をも理解して，適切にアドバイスしたり情報提供したりするなど，サービスを24時間，自動的に提供できれば，利用者の便はより高まり，関連のビジネスも拡大するものと期待される。

コグニティブアシスタントの実現のためには，現在の音声処理技術の高度化や高精度化だけではなく，利用者の発話認識を超えた意図の理解及び，画像や位置など種々のセンサ情報も利用した状況理解（認識力），状況認識や判断，アドバイスのための常識や分野知識の獲得と表現（知識），対話フィードバックや情報提供のための合成音声や自分の分身となるキャラクターであるアバターなどの表現モデル（表現力）が必要である。当社は，これらの技術の開発に取り組むとともに，これらの技術を融合し統合して，利用者に新たな価値をもたらす製品及びサービスの開拓と開発に努めていく。

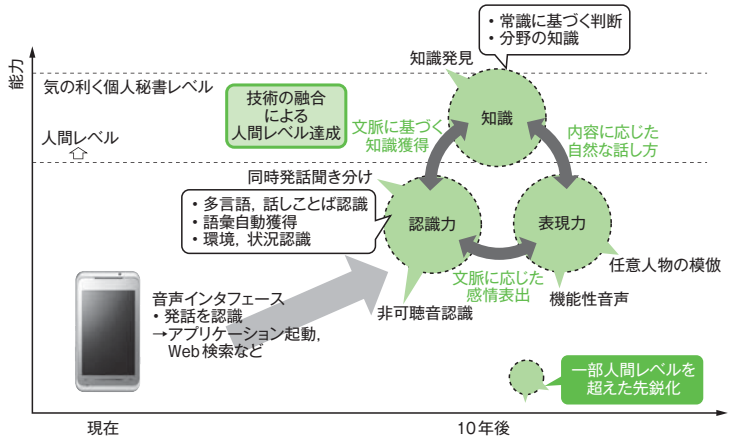


図3．コグニティブアシスタントに向けた音声インタフェースの展望 ― 今後の音声インタフェースは，音声出力技術の高度化や高精度化に加えて，気の利く個人秘書のような認識力や知識，表現力を持った役割が求められるようになる。

Trends in speech interfaces toward realization of cognitive assistant

文　献

(1) 山本幸一 他．雑音にロバストな音声と非音声の判別技術，東芝レビュー．**64**，12，2009，p.41－44.

(2) Akamine, M.；Kagoshima, T. "Analytic Generation of Synthesis Units by Closed Loop Training for Totally Speaker Driven Text to Speech System (TOS Drive TTS)". Proc. ICSLP98. Sydney, Australia, 1998-12, ASSTA. p.1927－1930.

(3) Kagoshima, T. et al. "An F0 Contour Control Model for Totally Speaker Driven Text to Speech System". Proc. ICSLP '98. Sydney, Australia, 1998-12, ASSTA. p.1975－1978.

(4) Latorre, J. et al. "Speech factorization for HMM-TTS Based on Cluster Adaptive Training". Proc. INTERSPEECH2012. Portland, OR, USA, 2012-09, ISCA. p.971－974.

赤嶺　政巳
AKAMINE Masami, D.Eng.
研究開発センター技監，工博。
音声処理技術の研究開発に従事。電子情報通信学会，日本音響学会，IEEE会員。
Corporate Research & Development Center