身近な機器での音声インターフェース構築に活用できる音声ミドルウェア

Speech Middleware to Incorporate Speech Interface into Smart Devices

瀬戸　重宣 SETO Shigenobu　　三上　茂 MIKAMI Shigeru

家庭内や日常的に持ち歩いて使用する身近な機器に音声インターフェース機能を付加するためには，メモリーサイズや計算量が機器のリソースによる制約を受ける場合でも，良好な性能で音声合成や音声認識を行えることが必要である。

東芝デジタルソリューションズ（株）は，リソース制約下でも効果を発揮でき，エッジ側でリアルタイム動作が可能な音声ミドルウェアを提供している。東芝コミュニケーションAI“RECAIUS”の音声認識ミドルウェア ボイストリガーと音声合成ミドルウェア ToSpeakは，身近な機器上で応答性良く稼働する音声合成・音声認識の機能を提供するソフトウェア部品であり，省メモリーで高品質な音声インターフェース構築に有効である。

In order to add a speech interface function to home appliances and mobile devices, speech recognition and speech synthesizer engines are required that can operate well even in an environment with a limited amount of memory and low computational capacity.

Toshiba Digital Solutions Corporation provides speech middleware components that work in real time with the limited resources of local devices. The voice trigger middleware and the ToSpeak text-to-speech (TTS) middleware available with RECAIUS make it possible to realize smart electronic devices incorporating speech recognition and speech synthesis functions. These middleware components are effective for the development of high-quality speech interfaces with small memory consumption.

1. まえがき

2014年に米国でリリースされ，我が国でも2017年の秋からリリースされた“スマートスピーカー”は，音楽やニュースを聞くだけでなく，メッセージの送受信や，買い物，列車の運行案内などの情報を聞くことができる，音声インターフェースを備えた，身近な機器の代表例である。

音声インターフェースは，20年近く前から，カーナビでのコマンド発声・応答や，テレマティクスサービスでの情報読み上げなどの機能を果たしてきている。最近では，日常的に持ち歩くスマートフォンに搭載された音声インターフェース（SiriやGoogle音声認識など）や，家庭に入り込み始めたスマートスピーカー（アマゾン社のAlexaや，Google社のGoogle Homeなど）により，日常生活で音声インターフェースが更に身近な存在となるのに加え，“使える”という期待感も高まってきている。

その一方で，音声インターフェースは，実使用環境での性能向上がかねてからの課題であった。例えば，音声合成の声が不自然であることや，読み誤りが発生すること，音声認識性能が十分でないこと，どのような発声をすれば音声認識性能を改善できるのかがエンドユーザーに分かりにくいことなどの問題があった。現実には，機器の計算能力による制約を考慮する必要があり，クラウドサービスを活用することでこれを緩和できる。しかし，ネットワークの良好な接続が確約できない利用シーンや，ネットワークを介してデータを授受する場合の遅延を許容できない応用分野では，クラウドサービスの活用は，しばしば解決策とはなり難かった。

東芝デジタルソリューションズ（株）は，雑音環境下での音声区間検出[1]や，少ない計算量での認識方式，少量のデータから元の話者の声に似せる技術[2]など，リソース制約下でも効果を発揮できる基盤技術を開発し，エッジ側でリアルタイム動作が可能な音声ミドルウェアを提供している。これらは，ユーザーの手元の機器側で，音声インターフェースを構築するための基幹部品ソフトウェアとして活用できる。

ここでは，音声ミドルウェアの中から，音声認識ミドルウェアのボイストリガー及び音声合成ミドルウェアのToSpeakについて述べる。更にこれらを，クラウドサービスで動作する自由文音声認識などの，より幅広い語彙を扱う音声認識技術と組み合わせた活用例について述べる（**図1**）。

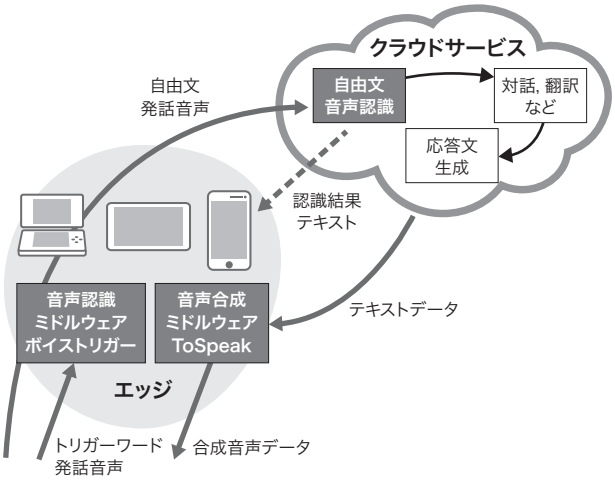

図1. エッジとクラウドサービスで連携動作する音声インターフェース

ボイストリガーをトークスイッチにして，それに続くユーザーの発話をクラウドサービスの自由文音声認識によりテキスト化した後，対話処理などで，例えばスケジュールデータを参照して，エッジ側のToSpeakで読み上げることができる。

Speech interface using middleware for edge computing and server services

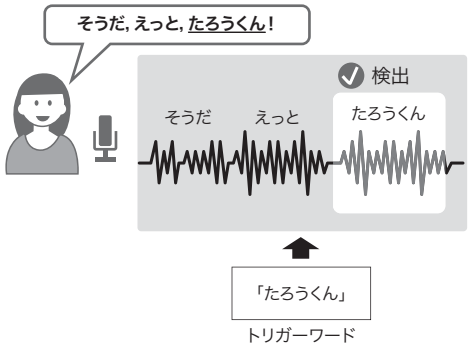

図2. 発話からのトリガーワードの検出

トリガーワードを設定すると，その言葉に反応するので，ハンズフリー機能などを実現できる。

Detection of trigger word in user utterance

図3. ボイストリガー機能を使った音声操作

機能をトリガーワードにすれば，特定の語彙の検出をトークスイッチとして活用できる。

Operation using voice trigger function

2. 音声ミドルウェア

音声ミドルウェアは，音声合成技術や音声認識技術が提供する音声処理機能を手軽に活用でき，機器内で稼働するアプリケーションから呼び出せる，ライブラリー形式の機能ソフトウェア部品である。機器メーカーが提供する端末機器製品や，サービスプロバイダーのサーバー稼働サービス，アプリケーションメーカーのアプリケーションソフトウェアなどに組み入れて，音声合成や音声認識機能を活用できる。

2.1 音声認識ミドルウェア ボイストリガー

ボイストリガーは，あらかじめ設定した特定の語彙を検出するもので（**図2**），要求メモリーサイズも計算量も小さい特長がある。常時稼働させることでハンズフリー機能を実現でき，トークスイッチとして利用することもできる。ユーザーは，トークスイッチを押してから発話するような使い方を意識する必要がなく，音声認識機能を気軽に利用できる。

ハンズフリー機能は，運転中に操作する車載機器や，キッチンで家事をしながら操作するスマートホーム端末などとの親和性が高く，今後，更に適用の広がりが期待できる（**図3**）。また，検出対象としたい語の音声データが収録されていなくても，テキストの発音表記を基にトリガーワード辞書を用意できる点も手軽である。

トリガーワードは，事前にトリガーワード辞書に設定しておく必要があるので，通常は少数の語彙を扱うのに向いている。短過ぎる語彙や，音響的特徴が類似する語彙を避けることにより数十語規模を扱うこともできるので，家電をはじめとする日常機器に導入したいというニーズが高まっている。

2.2 音声合成ミドルウェア ToSpeak

ToSpeakは，手本にするナレーターの収録音声から，声の音色や，抑揚，リズムなどの特徴を最適に反映した合成音声を任意の入力テキストに対して生成できる，当社の基盤技術を搭載した機能部品ソフトウェアである。数世代にわたる技術進化を経てきているとともに，いずれも極力少ない必要リソース（**図4**）の中で，良好で自然な音声合成を可能とする特長があるため，機器搭載に好適である。もちろん，ハードウェアスペックの制約は時代とともに緩和されてくるが，音声合成辞書のサイズが小さいことから，声の種類を増やしたり，後から声を加えたりしても，必要なハードウェアスペックは大きく変わらないので，搭載する機器の設計やコストへのインパクトは小さく，車載機器や，ゲーム機，電

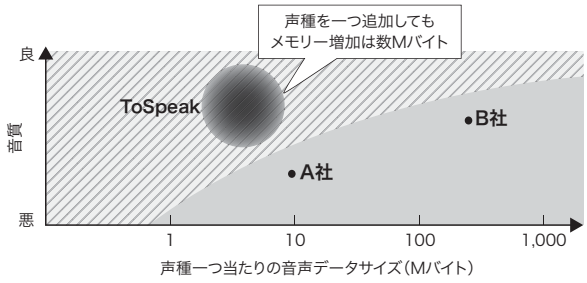

図4. 音声合成辞書の必要メモリーサイズ

ToSpeakは，他社に比べて音質を落とさずに小さいサイズで実装でき，車載機器や，ゲーム機，電子辞書などへの搭載実績がある。

Small memory footprint

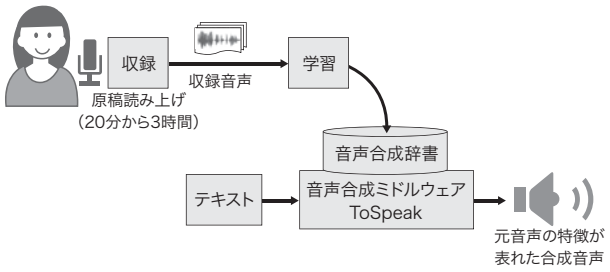

図5. 音声合成辞書の作成工程

元音声を収録して学習することで，元音声の音色，抑揚，リズムなどの特徴を反映した合成音声を作成できる。

Creation of speech synthesis corpus

子辞書などに搭載実績がある。併せて，読み上げ性能を向上させるために，クラウドソーシングを使った語彙拡充[3]などに取り組むとともに，楽曲用など特定のジャンル向けに，語彙辞書の整備も行っている。

ToSpeakの音声合成辞書の作成工程を**図5**に示す。通常のケースで数百文を，最小規模としては約100文を読み上げた収録音声を基に，元のナレーター音声の声色，抑揚，リズムの特徴を反映した合成音声を，任意のテキストから作成する[4]。更に，ToSpeakが標準として取りそろえている音声合成辞書の代わりに，特定のナレーターの収録音声からカスタムで音声合成辞書を開発することもできる。ラジオのDJ（ディスクジョッキー）やナレーターとして著名な話者の声を用いた作成事例[5]や，ゲームソフトウェアのメインキャラクターの声に活用した例[6]など，エンターテインメント用途での活用例が増えている[7]。

また，“コエステーション”[8]では，更に少ない10文の収録音声からでもカスタムの音声合成辞書を作成できる。ユーザー自身で声を登録してカスタムボイスを作成して自ら楽しめるだけでなく，ユーザー同士の間で音声合成辞書の相互利用や流通を促すプラットフォームとして活用してもらうことを目指している（**図6**）。

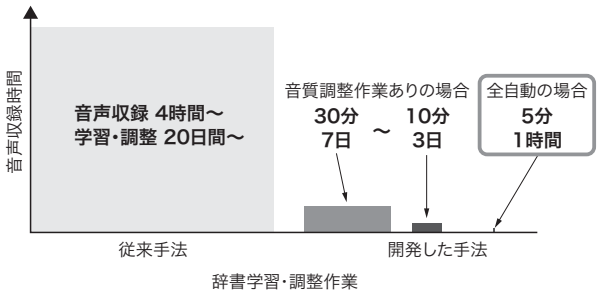

図6. 音声合成辞書の作成に要する収録データ量と辞書学習時間

従来に比べて少ない音声収録時間と学習時間での合成音声の作成を可能とした。

Required recorded voice data and TTS voice database training time

ToSpeakは，進化を続けている。“声のデザイン”技術を活用することによって，カスタムの合成音声を基準に，“明るい／暗い”や“はっきりした／くぐもった”といった声質パラメーターや，“喜・怒・哀・恐・優”などの感情パラメーターを指定することによって，後からせりふごとに声色を調整できるToSpeakG4も利用可能になっている。

更には，合成音声の自然さを飛躍的に高める新技術を搭載した，ToSpeakGx NEOも利用可能になった。現時点での音声合成辞書のメモリーサイズは大きく，省メモリーという従来のToSpeakの特長は引き継がれていないが，肉声感は極めて高く，しかも，カスタムの音成合成辞書の開発も可能で，ナレーターの声にこだわりを持つ用途に好適である。

2.3 幅広い語彙を扱う音声認識との組み合わせ活用例

当社は，自由文音声認識技術を使った音声認識サービスをクラウドサービス上で提供している。ディープラーニング手法を用いた高精度な音声認識，かつ，少ない手間での辞書カスタマイズが可能な音声認識が利用できる[9]。

例えば，この音声認識サービスと，エッジ側のボイストリガーを組み合わせて，アプリケーションサービスを構築する。ボイストリガーを特定のトリガーワードに反応する音声スイッチとして稼働させ，トリガーワードを検出したら，それに続く自由文の発話はクラウドサービス側の音声認識サービスで扱うように分担させることで，ハンズフリー機能を持つ自由文音声認識を実現できる。

また，自由文音声認識の代わりに，グラマー音声認識をボイストリガーと組み合わせることで，ハンズフリーでデータ

入力をすることも可能である。グラマー認識は，あらかじめ聞き取る単語とその単語の並び方をグラマー（語彙辞書）に具体的に記載しておくことで，ボイストリガーよりも規模の大きい語彙を扱うことができるタイプの音声認識である。ボイストリガーほど省リソースではないが，ボイストリガーとともにエッジ側で閉じて稼働させることが可能である。音声認識機能をネットワーク越しに利用するのではなく，エッジ側で完結させて処理できるので，ネットワークの使えないフィールド環境でのデータ入力に活用できる。

3. あとがき

音声ミドルウェアは，身近な機器で音声インターフェースを構築するための機能部品として活用できる。当社は，あらかじめ設定したキーワードを検出するボイストリガーのほか，省メモリーで高品質な音声合成ミドルウェア ToSpeak などの組み込み機器用のミドルウェアを提供している。また，高度な音声認識・合成・翻訳・対話・知識処理・画像認識が可能なクラウドサービス（WebAPI（Application Programming Interface）など）と組み合わせて活用することにより，ユーザーの手元の機器上で稼働させる音声ミドルウェアは，応答速度に優れ，軽快な音声インターフェースを実現できる。また，これに加え，省リソースという特長から，ユーザーごとのカスタマイズやパーソナライズを実現するのにも活用できる。

文　献

⑴ 山本幸一，赤嶺政巳．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．2009，64，12，p.41–44.
⑵ 橘健太郎，ほか．“個人声の合成音作成フレームワークの開発”．2011年春季研究発表会講演論文集．東京，2011-03，日本音響学会．2011，1-Q-34（c）.
⑶ 森田眞弘，ほか．多様な声や感情を豊かに表現できる音声合成技術．東芝レビュー．2013，68，9，p.10–13.
⑷ 芦川将之，ほか．“CrowdSourcing を用いた単語への読み付け，及びアクセント付け手法の提案”，電子情報通信学会技術研究報告．2012，111，447，p.11–16.
⑸ 東芝デジタルソリューションズ（株）．“RECAIUSのカスタムボイス機能から生まれた小林克也さんの似声キャラクター「コバカツ君」紹介ページ”．あなたを想うAI RECAIUS．<http://www.toshiba.co.jp/recaius/special-kobakatsu.html>，（参照 2018-05-21）.
⑹ 東芝デジタルソリューションズ（株）．“めがみスピークエンジン特設ページ”．あなたを想うAI RECAIUS．<http://www.toshiba.co.jp/recaius/2016_12_08.html>，（参照 2018-05-21）.
⑺ 赤嶺政巳．身近になった音声処理技術と東芝の取組み．東芝レビュー．2013，68，9，p.2–5.
⑻ 東芝デジタルソリューションズ（株）．“あなたのコエがしゃべり出す”．コエステーション．<https://coestation.jp>，（参照 2018-05-21）.
⑼ 許　海天，益子貴史．雑音に強い音声認識システム．東芝レビュー．2010，65，11，p.58–59.

瀬戸　重宣　SETO Shigenobu
東芝デジタルソリューションズ（株）
RECAIUS 事業推進部 営業部
日本音響学会・電子情報通信学会会員
Toshiba Digital Solutions Corp.

三上　　茂　MIKAMI Shigeru
東芝デジタルソリューションズ（株）
RECAIUS 事業推進部 営業部
Toshiba Digital Solutions Corp.