音声認識サービス基盤に関する取り組み

吉村 貴博・長友 健太郎・北出 祐
谷 真宏・服部 浩明

要　旨

これまで音声認識を業務アプリケーションで利用するためには、ソフトウェアの改造や高速なPCの導入など、さまざまな準備が必要でした。ここでは、音声認識エンジンをデータセンターに配置し、音声認識をサービスとしてユーザに提供する音声認識サービス基盤を紹介します。音声認識サービス基盤を実現する技術について解説するとともにその活用例について紹介します。

キーワード

●サービス型アーキテクチャ ●リアルタイム音声認識 ●電話音声認識 ●SFA
●伝言音声メール ●音声認識辞書

1. まえがき

本稿では、音声認識をサービス型アーキテクチャで提供する音声認識サービス基盤について紹介します。

従来、業務アプリケーションやコンシューマサービスに音声認識機能を組み込む際には、既存ソフトウェアを改造し、用途に応じた音声認識辞書を整備した上で、お客様のサーバやPCにこれらをインストールする形態が一般的でした。しかし、この提供形態では次のような課題がありました。

・ 既存アプリケーションの改造費や音声認識辞書の構築費など初期導入コストが高い。
・ 音声認識エンジンを稼働させるために高スペックなサーバやPCが必要。
・ 新語などに対応するための音声認識辞書のメンテナンスコストが高い。
・ 音声認識に適したマイクを用意する必要がある。

これらの課題を解決するために、弊社では音声認識サービス基盤の開発を進めています。音声認識サービス基盤とは、弊社のデータセンターに配置した音声認識エンジン・音声認識辞書をネットワーク経由のサービス型アーキテクチャで提供することで、より簡単に音声認識を活用できることを目指したプラットフォーム技術です。その特長は次のとおりです。

・ ウェブ技術をベースとしたAPIを備え、既存アプリケーションへの組み込みが容易。
・ サービス型アーキテクチャのため、サーバやPCの初期導入コストや運用コストを削減できる。
・ 弊社データセンター上で音声認識辞書を集中管理するので、各種メンテナンスコストが抑えられる。
・ 電話機など、マイク以外の多様なデバイスを利用できる。

本稿では音声認識サービス基盤を実現する技術について解説するとともに、その活用例について紹介します。

2. リアルタイム音声認識サービス基盤

2.1 はじめに

弊社が提供している「議事録作成支援サービス」はバッチ型音声認識システムを用いています。これは、あらかじめICレコーダなどで録音した音声ファイルをサーバにアップロードし、音声認識によって得られたテキストをまとめてダウンロードするというものです。

その一方、発言したそばからその内容をリアルタイムに入力・編集したいという要望も少なくありません。そこで我々は、リアルタイム動作が可能な新たなサービス型音声認識システムを開発しました。

2.2 システムの概要

今回開発したリアルタイム音声認識サービス基盤システムは**図1**に示す4つの要素から構成されています。

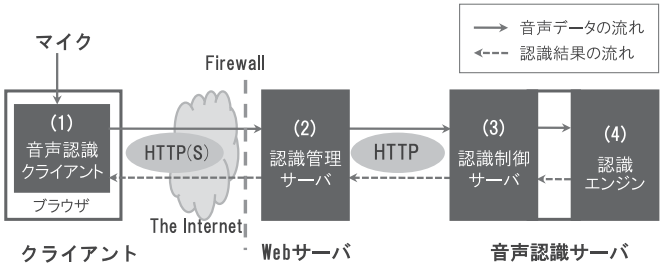

図1　リアルタイム音声認識サービス基盤システム概要

(1) 音声認識クライアント

ユーザ音声をマイクから入力し、0.5秒単位で分割してHTTP（S）リクエストを用いて認識管理サーバに順次送信します。同時に、リクエストに対するレスポンスを用いて認識結果を適宜受信し、ブラウザに表示します。

(2) 認識管理サーバ

音声認識クライアントから受信した音声データを認識制御サーバに転送します。併せて、認識制御サーバから受信した認識結果を音声認識クライアントに適宜転送します。クライアントや音声認識サーバが複数存在する場合は、クライアントごとに音声データを振り分けます。

(3) 認識制御サーバ

認識管理サーバから受信した分割された音声データを連結して認識エンジンに渡します。また、認識エンジンから取得した認識結果を認識管理サーバへ送信します。

(4) 認識エンジン

弊社の音声認識ミドルウェア「WebOTX Speech Recognition 7.1.0」です。

2.3 システムの特長

このシステムの最大の特長は、ユーザが話し始めた時点から音声認識エンジンが処理を開始することができる点です。これによってユーザが話し終えた直後に認識結果を得ることができるようになります。実際にインターネットを介して評価したところ、発話終了から認識結果表示までにかかる遅延は、発話の長さに関わらず、ほぼ1秒以内に収まりました。

ほかにも、以下のような特長があります。

・ 音声認識クライアントはActiveXコントロールとして提供されるため、事前のインストール作業など面倒な手間をかけずに音声認識システムを利用できます。

・ JavaScriptなどで簡単に音声入力の開始・終了を制御可能であり、さまざまなシステムやサービスと容易に連携させることが可能です。

・ クライアントとサーバ間の通信プロトコルにHTTP（またはHTTPS）を使用しているため、Firewallなどへの対応が容易です。また、モバイル通信プロバイダのように提供サービスが限定される通信環境でも利用できます。

2.4 今後の課題

上述のように、今回開発したリアルタイム音声認識サービス基盤は汎用性の高いHTTP（S）プロトコルを用いており、さまざまなクライアントへの展開が可能です。今後はスマートフォンや携帯電話などへとクライアント側の対応プラットフォームを拡大していく予定です。

3. 電話音声認識サービス基盤

3.1 はじめに

音声認識を使う際に意外に問題となるのがマイクです。今日でもマイクが標準搭載されているPCはそれほど多くはありません。

その一方、電話機はますます身近なものになっています。最近は誰でも常に携帯電話端末を持ち歩いていますし、職場に個人用の内線電話機を持っている方も多いでしょう。これらをマイクの代わりに使えれば、音声認識の導入に当たっての大きな障壁を1つクリアできます。

そこで我々は、電話機を手軽な音声入力デバイスとして利用し、さまざまなウェブベースのサービスと連携する技術を新たに開発しました。

3.2 システムの構成

今回、内線電話機をマイクとした内線電話音声認識システムを試作しました。**図2** にシステムの構成図を示します。

(1) 内線電話機

内線電話通話に用いる電話端末です。通常のビジネスフォンのほかに、構内PHSや無線LAN電話、更にソフトフォン

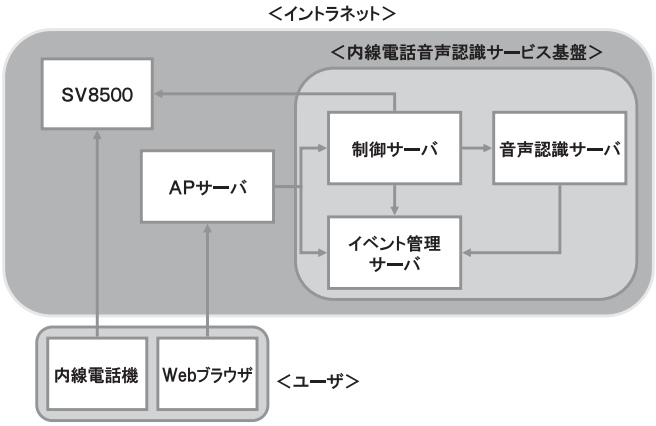

図2　内線電話音声認識システム構成図

などさまざまな端末が利用できます。また、SV8500を経由して外線から接続することも可能です。

(2) SV8500

構内PBX装置です。

(3) 音声認識サーバ

電話帯域の音声（8kHz）に最適化された音声認識を行います。その結果はイベント管理サーバに順次送られます。

(4) イベント管理サーバ

認識処理完了などの音声認識に関する各種イベントを簡易なウェブベースのAPIを通じて通知します。また、音声認識サーバが出力した認識結果を蓄積・管理します。

(5) 制御サーバ

上記の各サーバを結びつけ、電話音声認識サービス基盤を実現します。

(6) APサーバ

電話音声認識サービス基盤を利用するウェブベースのサービスです。

(7) Webブラウザ

本システムは標準的な技術のみを利用しているため、ブラウザの種類は問いません。

3.3 システムの利用例

このシステムを利用してさまざまな音声認識システムを作ることができます。ここでは、音声を使ってウェブベースのサービスを操作する音声UIついて紹介します。

音声認識を利用したいサービス（以降、ウェブサービスと略す）の提供者は、自分のウェブページに制御サーバへのリンクをあらかじめ埋め込んでおきます。これは、数行のJavaScriptを追加するだけです。

ユーザは、あらかじめ制御サーバに自分の電話番号を登録しておきます。次に上述のウェブページのリンクをクリックすると、登録した番号に向かってシステムから電話がかかってくるようになりますので、後は受話器を取って話し掛けるだけです。

ウェブサービスは、イベント管理サーバが提供する簡便なウェブベースのAPIを通じて音声認識結果を受け取ります。まず、ウェブサービスは、自分が受け取りたい認識結果の話者、回線（電話番号やIPアドレス）、時刻などをあらかじめイベント管理サーバに登録します。すると条件に見合ったイベントが現れる都度、イベント管理サーバから登録したウェブサービスにその内容が通知されるようになります。

この通知フレームワークはAJAX技術によって実装されており、JavaScriptを始めとするさまざまな既存テクノロジーと高い親和性を持っています。また、イベント通知は必要なタイミングで非同期的に行われるため、ポーリングなどによるネットワークへの過剰負荷の心配もありません。更に、非常に拡張性の高い設計となっているので、認識結果にさまざまな加工をする拡張エージェントを後から容易に組み込むことができます。

例えば、一度に複数の話者の音声のモニタリングを要求することができますので、複数人での電話会議を順次テキスト化するような使い方も可能です。ほかにも、認識結果から特定のキーワードをモニタリングする「キーワードモニターサーバ」や、認識結果を翻訳する「翻訳サーバ」などの利用が考えられます。

3.4 今後の課題

本システムによって、ユーザ導入の容易さ、ウェブサービス構築の簡便さなど、さまざまな課題を解決することができました。今後はさらなる利便性の向上のほか、後述する「伝言メール」のような社内試験を通じて、より良い音声認識サービスの開発に役立てたいと考えています。

4. SFAにおける活用例

4.1 はじめに

ここでは音声認識サービス基盤の活用例としてSFA（Sales Force Automation）での効果について説明します。SFAとは、営業活動を支援し効率化するための情報システムです。営業日報や商談の進捗のような業務情報を蓄積できるとともに、蓄積情報を共有化することでこれまで個々人に依存していた営業ノウハウを共有・標準化し、戦略的な営業活動を支援することができます。

しかしながら、いざ実際に運用してみるとさまざまな問題に直面します。

・ 情報の入力に手間がかかり、多忙な営業担当者ほどシステムを使ってくれない。

・ 古い情報や正確性に欠ける情報が蓄積されてしまい、十分に活用できない。

これらを解決する手段として音声認識技術が注目されています。音声認識を活用することで、次のような利点が見込まれます。

・ キーボードやメモ帳を使わず携帯電話から簡単に入力できる。

・ 顧客訪問の直後に電話で速報を登録できるようになるので、素早く正確な情報を残すことができる。

・ 速報を利用して、帰社後の報告書作成にかかる時間を減らすことができる。

・ 管理者はよりタイムリーな営業状況の把握が可能となり、迅速なフォローができるようになる。

4.2 システムの構成

音声認識サービス基盤と、弊社の営業が利用しているSFAと組み合わせたシステムを試作しました。**図3** にシステムの構成と利用の流れを示します。この図にならって、システムの動作フローを説明します。

(1) 営業が携帯端末から指定の電話番号に発信し、速報を口頭で入力します。その内容は音声録音サーバによって音声ファイルとして記録されます。

(2) 記録された音声ファイルが音声認識サービス基盤に送信され、音声認識エンジンによってテキストに変換されます。

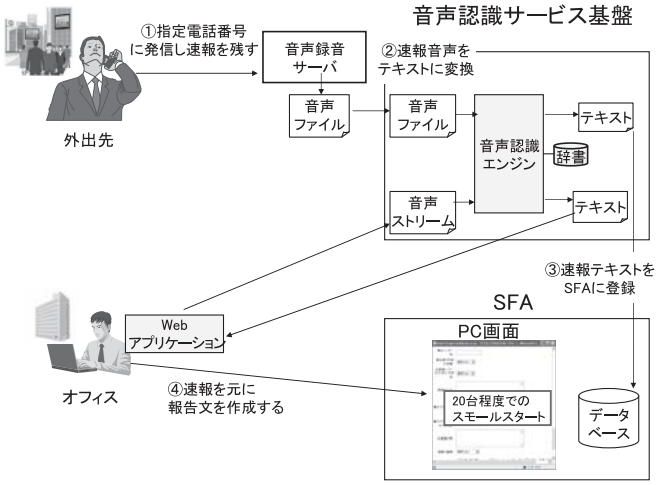

図3 SFA連携システムの構成と利用の流れ

(3) 変換された速報テキストがSFAのデータベースに登録されます。

(4) 営業がオフィスのPC端末から速報の内容を確認し、報告文を作成します。

本システムはリアルタイム音声認識サービス基盤の「音声認識クライアント」も組み込んでいるため、報告文の作成に音声入力を利用することも可能です。

なお、本システムの構築に当たってSFA側で必要になった改造はごくわずかで、非常に短時間に稼働させることができました。

4.3 有効性の検証

このシステムの有効性を検証するため、実際の営業部門と連携して実証実験を行っています。実験は現在も継続していますが、これまでに以下の知見が得られています。

・ 報告文の作成など、事前に文章の推敲をしてから話すのは難しく、ある程度の慣れが必要である。

・ 速報のようにメモやキーワード中心の内容であれば推敲の必要がなく、使いやすい。

これらは、従来の業務フローとは異なる新たな操作にユーザが戸惑った面も影響していると考えられます。今後はユーザにシステムに慣れてもらうだけでなく、システム側もユーザの戸惑いを少なくする工夫が必要となります。

4.4 今後の課題

実証実験の結果を勘案し、まずはメモやキーワードなどの個人速報用途をターゲットに使い勝手の向上を進めていきます。また、個人速報から効率的に報告文を作成するフローを作り上げることで、実用レベルに引き上げる活動を行っていきます。

5. 伝言メール送信における活用例

5.1 はじめに

ここでは、電話音声認識サービス基盤を用いた伝言音声メール送信サービスについて紹介します。

オフィスで頻繁に目にする風景に電話の取り次ぎがあります。取り次ぎ先の人間が不在の場合、電話をかけてこられた方からの伝言をメモとして残すことも日常的です。最近では伝言メモをわざわざメールに書き写して送ることも多いのではないでしょうか。このサービスは、音声認識サービス基盤を用いてこの作業を支援するためのものです。

5.2 サービスの概要

伝言を頼まれたユーザは、まず所定のウェブページにアクセスし、あらかじめ登録された自分宛の電話番号を確認したら、「電話をかける」ボタンを押します。すると、指定した端末に電話がかかってくるので、受話器を取って伝言内容を音声で入力します。音声をテキスト化した結果がテキストボックスに順次表示されるので、内容を確認します。その後、送信するメールアドレスを選択して、送信ボタンを押します。これで、伝言メールが取り次ぎ相手に届けられます。**図4** にサービスのイメージを示します。

音声認識サーバの認識精度は利用する音声認識辞書に大きく依存しますが、事前に登録しておける言葉にはどうしても限界がありますので、適宜言葉を補充する必要があります。

そこで、このサービスでは、よく使う文言や電話番号など最小限の言葉のみからなる音声認識辞書をあらかじめ用意しておき、足りない部分についてはユーザが適宜追加できる仕組みを用意しました。具体的には、複数の単語列を持つ定型文をいくつか用意し、その単語列部分にユーザが音声認識さ

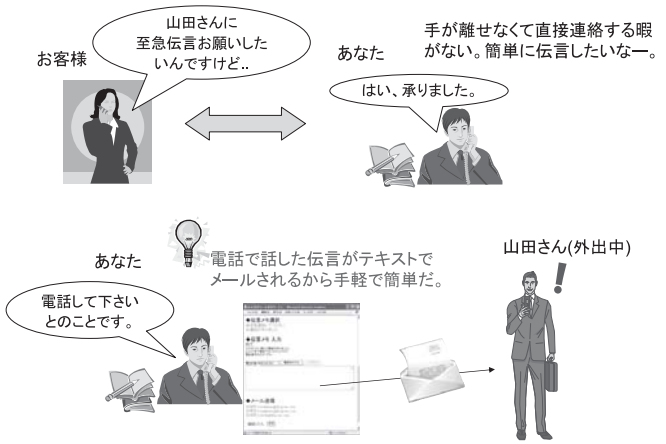

図4　伝言音声メール送信サービスのイメージ

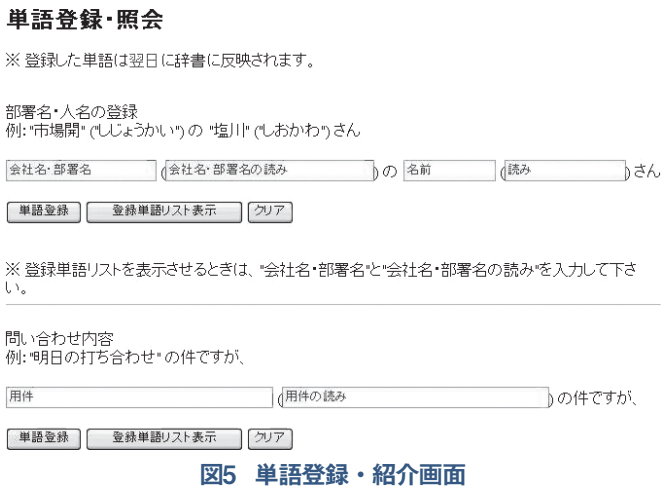

図5　単語登録・紹介画面

せたい言葉を登録できるようにしました。**図5** に実際の登録画面を示します。

5.3 音声認識辞書の共有

上述した定型文登録機構は、音声認識辞書をユーザの集合知で強化する仕組みであるといえます。

あるユーザが登録した言葉を、別のユーザが音声入力する際にも利用できるようにすることで、個々のユーザが自分のために言葉を追加すると、すべてのユーザがその恩恵を受けられるようになります。このようにして、ユーザが集まって音声認識辞書を“育てる”ことにより、実際の利用に即した

さまざまな文章を認識できるようになります。

従来は音声認識辞書は各ユーザのPC上に置かれており、こうしたユーザ間の連携を実現するのは容易ではありませんでした。それに対して、音声認識サービス基盤を用いれば、音声認識辞書をネットワーク上で共有することが可能なので、このような仕組みを極めて容易に実現できるのです。

5.4 今後の課題

開発したシステムは、弊社内で実際に運用し、その有効性を検証しています。現時点では、利用できる定型文が少ないため効果は限定的です。定型文の数が増えてくれば利便性向上が予想されるため、引き続き定型文登録機構の改良を続けていきたいと考えています。

6. むすび

今回開発した音声認識サービス基盤を用いることで、音声認識を利用したシステムをより簡単に低コストで構築・運用することが可能となりました。更には音声認識辞書を集合知によって”育てる”ような、新しい運用形態も実現できるようになりました。

今後は更なる利便性の向上と対応環境の拡大を進め、より幅広いシステムやサービスから利用できる基盤を目指して開発を進めて参ります。

執筆者プロフィール

吉村 貴博
市場開発推進本部
主任

長友 健太郎
共通基盤ソフトウェア研究所
主任

北出 祐
共通基盤ソフトウェア研究所

谷 真宏
NEC情報システムズ
先端技術ソリューション事業部
電子情報通信学会会員

服部 浩明
NEC情報システムズ
先端技術ソリューション事業部
シニアエキスパート
日本音響学会
電子情報通信学会各会員