顧客の発話を認識して応答できるコールセンター向け自動音声応答システム

Automated Interactive Voice Response System Using Speech Synthesis and Recognition Technology

酒井　静磨　　安田　博和
■SAKAI Shizuma　■YASUDA Hirokazu

コールセンターでは，運用コストの削減やオペレーターの業務負荷軽減のために，オペレーターによる電話受付に代わり，自動音声応答（IVR：Interactive Voice Response）システムによる自動受付を利用する機会が増えている。このシステムでは，顧客からの電話を受け付け，要望を自動判別するために音声ガイダンスの中から所望の選択肢をプッシュトーンで指定してもらう方法を取っていた。しかし，コールセンターで扱う業務の多様化に伴い，個々の要望を詳細に把握するためにはプッシュトーンによる指定だけでは不十分なケースが増え，オペレーターが改めて対応してその不足部分を補う運用が一般的になっている。

このような状況を踏まえて東芝は，発話内容をオペレーターに代わり音声認識することで，その結果に応じて自動的に応答できる，コールセンター向けIVRシステムを開発した。

Call centers are experiencing an increasing need for automated interactive voice response (IVR) systems as an alternative to operators, in order to reduce both operating costs and the burden on operators. In general, an IVR system provides customers with certain options using voice guidance, allowing them to select their requirements by pushing touch-tone buttons on their telephone handset. However, as it is difficult to understand the requirements of each customer in detail due to the diversification of operations in call centers, operators provide assistance to customers in many cases.

With these trends as a background, Toshiba has developed an automated IVR system for call centers that makes it possible to recognize a customer's voice and respond to it automatically, applying its proprietary speech synthesis and recognition technology.

1　まえがき

コールセンターでは顧客向けの製品販売やサービスなどへの注文，苦情，及び各種問合せを受け付ける業務を行っている。従来，これらの受付業務はオペレーターによって運用されるケースが主流であったが，近年，コールセンターの運用コスト削減とオペレーターの業務負荷軽減を目的として，IVRシステムが広く使用されている。

東芝は，このシステムの可用性を高めるために，顧客に応じて柔軟に自動応答できるコールセンター向け音声合成システムを2011年に開発し，コールセンターにおける音声ガイダンス生成の負荷軽減に寄与してきた。しかしその一方で，コールセンターに電話をかけた顧客が，IVRの提供する音声ガイダンスの中から，所望の選択肢をプッシュトーンで指定して回答する操作は従来のままである。このため，氏名や住所などプッシュトーンでは表現しきれない情報を確認する必要がある場合は，依然としてオペレーターが介在して確認しなければならなかった。

そこで当社は，音声認識技術を活用して，プッシュトーンでは表現しきれない発話内容も自動判別して，より柔軟に応対できるコールセンター向けIVRシステムを開発した。

ここでは，今回開発したIVRシステムの概要と特長，及び動作を検証した結果について述べる。

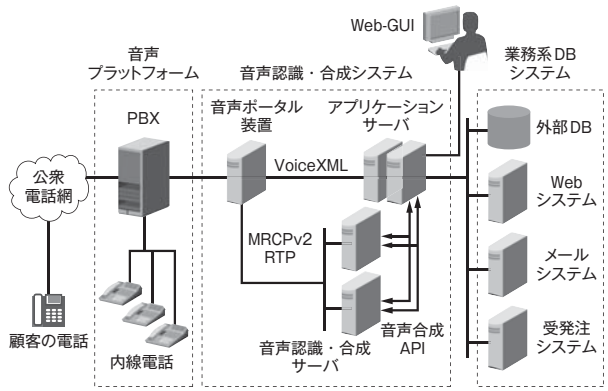
2　音声認識・合成によるIVRシステムの概要

このIVRシステムは，当社が2011年に開発した音声合成システムに音声認識技術を追加実装したもので，従来と同様に各種装置との間をIP（Internet Protocol）ネットワークで通信する構成にしている。これにより着信呼に付随する情報を取得して，応答処理シーケンスの選択，音声合成に用いるテキスト情報の生成，及び音声認識した結果の活用ができ，様々な着信呼に対して，きめ細かく応対できる。

例えば，IVRからの「お電話ありがとうございます。まず，お客さまの生年月日をお伺いします。発信音の後にお客さまの生年月日をおっしゃってください。ピッ。」というアナウンスに対して，電話の発信者が「1990年1月1日」と応答すると，この音声を基に生年月日を認識して，データベース（DB）に登録されている生年月日情報との整合を自動的にチェックし，次の応答シーケンスを選択するなどの動作が可能である。

音声認識・合成システムは主に，アプリケーションサーバ，企業内電話交換システム（PBX：Private Branch Exchange），音声ポータル装置，及び音声認識・合成サーバから構成される。システムの構成を**図1**に，音声認識・合成サーバの主な仕様を**表1**に示す。

前述の例を用いて，IVRが「…お客さまの生年月日を…」のフレーズを音声合成して再生する動作の概要を以下に述べる[1]。



GUI：Graphical User Interface
API：Application Programming Interface

図1．コールセンター向け音声認識・合成システム ― 音声認識・合成機能を装備し，ダイナミックに応答ガイダンスを生成し，顧客の発話を認識判別可能なIVRシステムである。

Configuration of speech synthesis and recognition system for call centers

表1．音声認識・合成サーバの主な仕様

Main specifications of speech synthesis and recognition server

項　目	仕　様
音声合成原理	コーパスベース 複数素片選択融合方式
音声合成辞書	言語辞書 ：14万語程度 波形辞書 ：音声種類当たり数十Mバイト
音声認識辞書	単語発音辞書 認識グラマ：SRGS
音声合成用入力テキスト形式	漢字仮名交じり文，SSML，JEITA TT-6004*1
通信インタフェース	MRCPv2，音声合成API（東芝独自仕様）
音声コーデック	ITU-T G.711*2 μ-law，A-law
合成音声ファイル形式	WAVEフォーマット 8 kHz：μ-law，A-law 16 kHz：リニアPCM
対応言語	日本語（外国語は今後対応予定）
対応OS	Linux(注1)

SRGS：Speech Recognition Grammar Specification
SSML：Speech Synthesis Markup Language
PCM　：パルス符号変調
OS　：基本ソフトウェア

＊1：一般社団法人 電子情報技術産業協会規格TT-6004
＊2：国際電気通信連合 電気通信標準化部門勧告 G.711

PBXは，顧客からの着信呼を音声ポータル装置に転送する。音声ポータル装置は，その着信呼に付随してPBXから取得した着信呼情報に従い，アプリケーションサーバに対して応答シーケンスを要求する。アプリケーションサーバは，応答に必要なテキスト情報を生成し，そのテキスト情報を格納した所望の応答メッセージをVoiceXML（Voice Extensible Markup Language）形式で音声ポータル装置に応答する。その後，音声ポータル装置はアプリケーションサーバから受信した応答メッセージ及びテキスト情報に従い，音声合成要求を音声合成サーバへMRCPv2（Media Resource Control Protocol version 2）を用いて送信する。音声合成サーバはこの要求に基づいてアナウンスを生成して，合成音声を音声ポータル装置へRTP（Real-time Transport Protocol）でストリーム送信し，IVRによる受付を開始する。

この例では音声合成によるアナウンスで顧客に対して生年月日を発話するように促している。このアナウンスに続いて，音声認識処理をより円滑に行うため，アプリケーションサーバは，生年月日を認識するための情報として，生年月日の発話を単語の並びとして記述した“認識グラマ”を生成し，更にこの情報を格納したメッセージをVoiceXMLの形式で音声ポータル装置に送信する。音声ポータル装置は認識グラマを含むメッセージを音声認識サーバにMRCPv2を用いて送信する。この認識グラマは最終的に音声認識エンジンに提供され，生年月日の認識処理に活用される。

アプリケーションサーバは認識グラマを送信した後，音声認識要求のメッセージを音声ポータル装置にVoiceXML形式で送信する。音声ポータル装置はこのメッセージに従って，音声認識要求を音声認識サーバへMRCPv2を用いて送信する。

一方で，顧客は音声合成で生成されたアナウンス「…お客さまの生年月日を…」を聴いたうえで自身の生年月日を発話する。顧客の音声は音声ポータルからRTPストリームで音声認識サーバに入力される。音声認識サーバは，入力された音声データ及び認識グラマにより，顧客の生年月日をリアルタイムで認識する。認識された生年月日の情報はテキスト化され，音声ポータル装置を経由してアプリケーションサーバに返信される。

テキスト化された情報はアプリケーションサーバで，種々の用途に用いられる。例えば生年月日のテキスト情報を，音声合成により復唱して内容の確認を顧客に促したり，DBと照合して一致しているかどうかを自動的に判定したりできるようになる。

3　システムの特長

システムの特長は，以下のとおりである。

3.1　音声認識エンジン

音声認識には，短い単語単位での発話を認識するためのコマンド音声認識と，会話などの比較的長い発話を認識するためのディクテーション音声認識の2方式がある。今回開発した音声認識・合成システムはIVR機能として活用するため，単語単位で認識ができれば十分である。そのため，このシステムではコマンド音声認識を採用している。

コマンド音声認識エンジンの基本構成を**図2**に示す。入力された音声からその音響的特徴を表現した特徴量を抽出し，認識辞書と照合することで音声認識結果を出力する。認識辞書は，音声として受理可能な発話内容を単語の並びとして記述した認識グラマ，各単語の発音を記した“単語発音辞書”，

（注1）　Linuxは，Linus Torvalds氏の米国及びその他の国における登録商標。

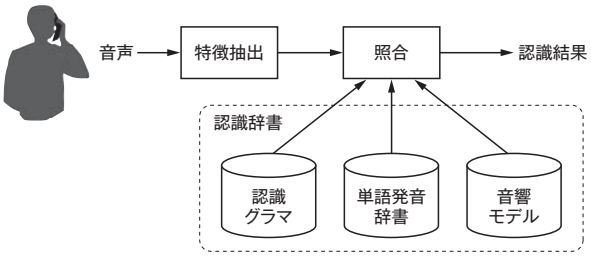


図2．音声認識エンジン方式の基本構成 ― 音声の特徴量を抽出して，その特徴量を，認識グラマ，単語発音辞書，及び音響モデルから成る認識辞書と照合することで音声認識を行う。

Flow of processes of speech recognition engine

及び発音の単位である音素ごとにその音響的特徴を記した“音響モデル”から成る。

音声認識の性能に大きな影響を及ぼすのは特徴量である。特徴量は，誰がどのような状況で話した場合でも音声の本質的な特徴を安定して表現できることが望ましい。当社の音声認識エンジンは，特徴量として，当社が独自に開発した帯域別平均時間ケプストラム（Subband Average Time Cepstrum：SATC）特徴[2]を用いることで，高精度な認識を実現している。

3.2　通信プロトコルを使用したリアルタイム音声認識

音声認識サーバと音声ポータル装置の間の通信プロトコルには，音声合成サーバと同じくMRCPv2を使用している。音声合成サーバと音声認識サーバが同一の通信プロトコルで音声ポータル装置と接続されることにより，一つの通話の中で音声合成と音声認識が互いに連携して動作する。音声ポータル装置は，アプリケーションサーバから音声合成用テキスト情報を含んだVoiceXMLメッセージや，認識グラマを含んだVoiceXMLメッセージなどを受信することに応じて，音声合成サーバに音声合成要求を送信したり，音声認識サーバに音声認識要求を送信したりする。

MRCPv2を使用して音声認識する場合，まずSIP（Session Initiation Protocol）で，MRCPセッションとRTPセッションの二つのセッションを確立する。次にMRCPセッションで音声認識するための認識グラマを通知する。

この認識グラマの記述フォーマットはW3C（World Wide Web Consortium）で規定されているSRGS（Speech Recognition Grammar Specification）を使用している。認識グラマには，ユーザーがこれから発話するであろうと予測される単語の候補列や，発話順序，それらの単語の発音に対応する正解テキスト，発声繰返しの有無と繰返し回数などを指定できる。この認識グラマは音声入力が始まる前に認識エンジン内に格納され，その後の音声認識処理で活用される。音声入力が開始されると音声認識エンジンは，有音検知を契機に，音声特徴量を抽出して，その特徴量を認識辞書と照合し，結果として

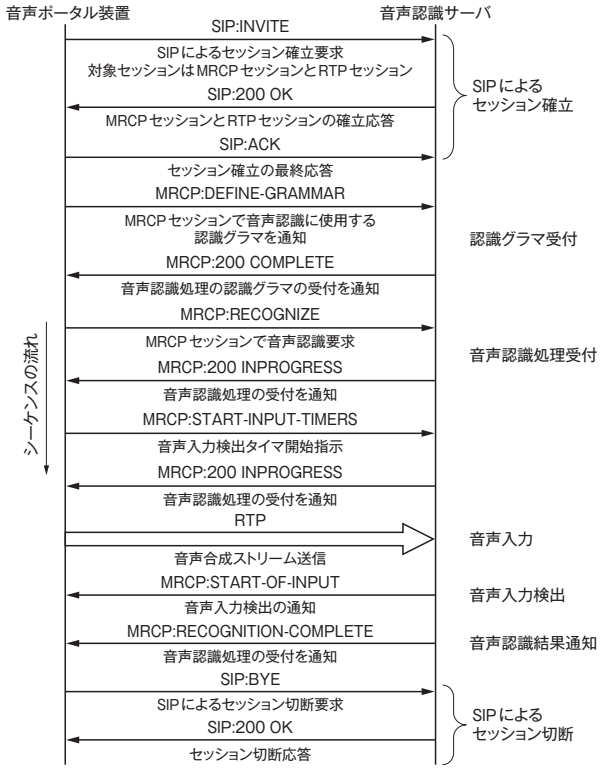


図3．MRCPプロトコル基本シーケンス ― SIPを使ってセッションを確立し，MRCPとRTPを使用して音声合成した結果をストリーミング送信する。

Basic sequence of Media Resource Control Protocol (MRCP)

認識結果のテキスト情報を生成する。認識結果のテキスト情報はW3Cで規定されるNLSML（Nature Language Semantics Markup Language for the Speech Interface Framework）形式を使用してMRCPv2プロトコルで音声認識サーバから音声ポータル装置を経由し，アプリケーションサーバに対して応答される（**図3**）。

音声認識結果はテキスト情報として提供されるので，その後のアプリケーション動作に活用することが容易である。例えば，音声認識結果のテキスト情報を再び音声合成処理することで，顧客の発話内容を復唱して認識結果の確認を促したり，認識結果のテキスト情報が注文された商品名を示している場合は，商品DBと照合して，その商品名に対応する商品コードを自動的に特定し，更にその商品の在庫状況を確認するなどの処理と連動させたりすることも可能である。

音声認識機能と音声合成機能はそれぞれ個別のサーバとして構築することも，単一のサーバ上に両機能を実装することも可能である。

3.3　単語発音辞書登録機能

音声認識技術をコールセンター業務に適用してシステムの運用を行うにあたり，今後生じる造語や，新語，取扱商品名などの特殊な固有名詞も認識の対象となる。したがってこれらの

新しい単語を受理可能にするためにコールセンター個々の運用の中で単語発音辞書を充実していく必要がある。

この対応として，コールセンターの運用者が新たな単語を単語発音辞書に追加登録できる機能を用意した。この機能を使用して追加登録することで，音声認識処理において新たな単語が受理可能になる。単語の登録にあたっては，単語の発音とこれに対応する正解テキストを登録する。

3.4　認識グラマダウンロード機能

3.2節で述べた認識グラマは，MRCPv2メッセージで認識グラマの内容そのものが通知されるケースと，認識グラマが格納されている位置を示すURL（Uniform Resource Locator）が通知されるケースがある。必要に応じて使い分けることが可能になるように，両者の方式に対応している。URLで認識グラマの格納位置が通知された場合は，該当URLからダウンロードして，その後の認識処理に活用する。

3.5　RTP受信処理

PBXに入力された顧客の発話音声は音声ポータル装置を経由して，RTPのストリーム形式で音声認識サーバにリアルタイムに入力される。この音声ポータル装置の処理負荷あるいはネットワークの負荷状況に依存して，音声認識サーバが受信するRTPパケットの受信間隔に揺らぎが生じる可能性がある。このため音声認識サーバ内部の受信バッファでいったん格納し，この揺らぎを吸収したうえで，音声認識エンジンに入力している。また，RTPのペイロードフォーマットとして，音声合成サーバと同様にITU-T（国際電気通信連合 電気通信標準化部門）勧告のG.711μ-law（わが国や米国で採用）とG.711A-law（主に欧州で採用）の2方式をサポートした。

3.6　冗長構成

システムの信頼性を確保するための構成として，音声認識サーバを複数台並列運転させることが可能な構成としている。かりに1台の音声認識サーバに障害が発生し，そのサーバ上で音声認識サービスが提供できなくなった場合は，この音声認識サーバがネットワークから切り離され，音声ポータル装置が別の音声認識サーバに接続先を切り替えて運用を継続できる。

また，3.3節で述べた単語発音辞書登録機能を使用して登録された辞書情報は，1台の音声認識サーバに登録するだけで，サーバ間のデータ同期機能を使って直ちに他のサーバに反映される。このため，ある音声認識サーバで障害が発生しても同じレベルの音声認識サービスを継続できる。

4　音声認識・合成システムの動作検証

MRCPv2はリアルタイム性に優れ，時間的な応答性が良い。システムとしての動作を検証するため，音声合成サーバに音声認識機能を追加実装して，発話を終了してから認識結果を出力するまでの時間を測定した。この結果，上限値に設定した3,000 msを超えない良好な結果を得た。

また音声認識性能については，今回，電話回線を通した音声を認識するため，この特性に適合した音響モデルを新たに作成して検証を行った。不特定話者で，話者の周囲環境を無作為に抽出し，電話回線を通した音声で，氏名，電話番号，住所のそれぞれについて90％以上の認識率を確保できていることを確認した。

5　あとがき

音声認識・合成によるコールセンター向けIVRシステムは，2011年に開発した音声合成システムに，更に当社独自の音声認識技術をコアとして追加したシステムである。音声認識技術の活用により，これまでのIVRシステムでは，電話のプッシュトーンだけで表現されていた回答を，発話で表現できるようになり，顧客側の観点からも柔軟性の高いシステムにできた。また，コールセンターの運用面でもオペレーターに代わりIVRを活用できる場面が拡大することで，コールセンターの業務負荷を低減できる。

今回，コマンド音声認識技術の活用により，IVRシステムとしての基盤技術が完成した。今後は，コマンド音声認識のいっそうの性能向上を図るとともに，ディクテーション音声認識の技術開発も並行して進めていく。

文　献

(1) 酒井静磨 他．顧客に応じて柔軟に自動応答ができるコールセンター向け音声合成システム．東芝レビュー．66，10，2011，p.43－46.
(2) 中村匡伸 他．“群遅延に基づく音声特徴量の雑音音響下での評価”．日本音響学会2012年春季研究発表会講演論文集．横浜，2012-03，日本音響学会．2012，p.135－136.

酒井　静磨　SAKAI Shizuma
社会インフラシステム社 府中事業所 放送・ネットワークシステム部主査。IP電話システムの開発・設計に従事。
Fuchu Complex

安田　博和　YASUDA Hirokazu
東芝通信インフラシステムズ（株）エンジニアリング本部 ソフトウェア設計部。IP電話システムの開発・設計に従事。
Toshiba Communications Infrastructure Systems Corp.