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

Speech Synthesis System for Call Centers with Flexibility to Handle Various Inquiries

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きている。しかし，取扱商品の多様化や，顧客に応じた応答内容のカスタマイズの必要性などにより，多種の応答ガイダンスを準備する必要があり，ガイダンスを収録する従来の方法では非常に手間を要していた。

このような状況を踏まえ東芝は，安定した品質の合成音声によって自動的に応答することができる，コールセンター向け音声合成システムを開発した。高音質な音声合成エンジンを搭載するとともに周辺の業務系システムと連携することで，顧客に応じて柔軟に応答用テキストを生成し，そのテキスト情報に基づいて肉声感のある合成音声の提供を実現した。

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.　However, as a wide range of responses for guidance are required to handle various products and to customize the responses to each customer, the conventional method of preparing audio guidance recordings takes a great deal of time.

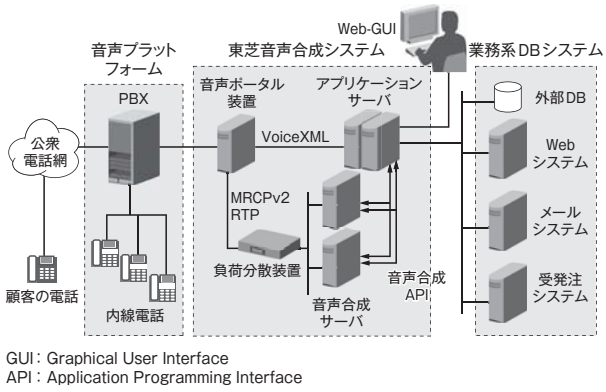
With these trends as a background, Toshiba has now developed a speech synthesis system for call centers that incorporates a high-quality speech synthesis engine and can function cooperatively with peripheral business systems.　This system automatically generates stable response guidance texts based on information provided by customers, and responds using a synthesized voice with realism close to that of a real operator's voice.

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，自動音声応答（IVR：Interactive Voice Response）システムが広く利用されている。

このシステムでは，電話の自動応答を行うための音声ガイダンスの準備に，プロのアナウンサーの声を収録して編集する方法や，特定のオペレーターの声を録音する方法が一般的に用いられている。一方コールセンターでは，取扱商品の多様化や，顧客満足度向上のための応対内容のカスタマイズ化が進み，必然的に音声ガイダンスの種類が増大する傾向にある。このため，アナウンサーによる音声収録費用の増大や，自身の声を公開するオペレーターの精神的な負担が指摘されていた。

そこで東芝は，音声合成技術を活用することでコールセンターの音声ガイダンス生成のための負担を軽減し，顧客に応じて柔軟な応答ガイダンスを生成し，自動的に電話の応対を行うことができるコールセンター向け音声合成システムを開発した。ここでは，このシステムの概要と特長的な機能について述べる。



GUI：Graphical User Interface
API：Application Programming Interface

図1．コールセンター向け音声合成システム ― 音声合成機能を装備し，柔軟に応答ガイダンスを生成できる自動音声応答システムである。
Configuration of speech synthesis system for call centers

2　音声合成システムの概要

この音声合成システムは，従来のIVR機能に加えて，電話システムや業務系システムとの間をIP（Internet Protocol）ネットワークで通信することで，着信呼に関するより詳細な付属情報を収集できる。これらの情報から，着信呼に応答するための処理シーケンスの選択と，音声合成するためのテキスト情報の自動生成を柔軟に行うことができるので，顧客からの様々な

表1．音声合成サーバの主な仕様

Main specifications of speech synthesis 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
言語辞書	14万語程度
波形辞書	数十Mバイト／音声種類
入力テキスト形式	漢字かな交じり文，SSML，JEITA TT-6004*
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	G.711μ-law，G.711A-law
音声ファイル形式	WAVEフォーマット 8 kHz：G.711μ-law，G.711A-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)

SSML：Speech Synthesis Markup Language　　OS：基本ソフトウェア
PCM　：パルス符号変調
＊（社）電子情報技術産業協会規格TT-6004

着信呼に対してきめ細かに応対できる。1台の音声合成サーバで，同時に最大50呼の着信呼に対して音声合成によるIVRを行うことができる。

コールセンター向け音声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，音声合成サーバの主な仕様を**表1**に示す。

PBXは，顧客からの着信呼に応答し，音声ポータル装置にその着信呼に関わる付属情報を伝達する。音声ポータル装置は着信呼の情報に従い，応答ガイダンスを含んだ応答シーケンスをアプリケーションサーバに対して要求する。

アプリケーションサーバは，その後段に接続される業務系システムのデータベース（DB）から応答ガイダンスに必要な情報を受信して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シーケンス内容をVoiceXML（Voice Extensible Markup Language）の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シーケンス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応答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。

3　システムの特長

システムの主な特長は，以下のとおりである。

3.1　音声合成エンジン

音声合成エンジンは，テキスト解析部，韻律生成部，及び音声信号生成部の三つのモジュールで構成されている（**図2**）。

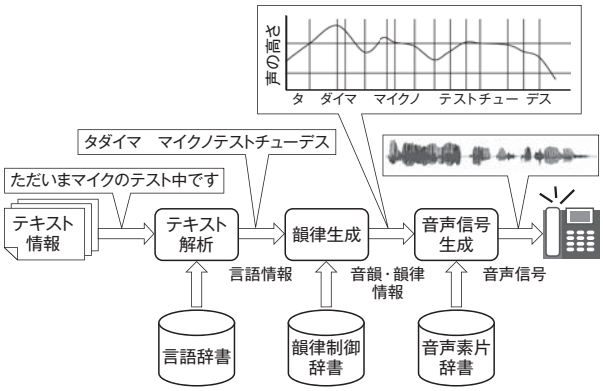


図2．テキスト音声合成 ― 任意のテキストに対して言語解析，韻律生成，及び音声信号生成の3段階の処理によって合成音声が生成される。

Mechanism of text to speech

テキスト解析部では，言語辞書を参照して入力された漢字かな交じり文を解析し，適切な読みやアクセントの位置などの言語情報を生成する。韻律生成部では，韻律制御辞書を参照して話す速度（話速）や抑揚などの韻律情報を生成する。音声信号生成部では，音声素片辞書と韻律情報に従って音声素片（合成単位ごとに切り分けられた音声波形)を接続し，合成音を生成する。

従来の音声信号生成部では，音声データ（コーパス）から目標となる音韻に近い音声素片を合成単位ごとに1個選択し，この素片を変形して接続する方法が用いられていた。しかし，適切な素片がない場合には，部分的に音質が劣化するという問題が生じていた。

当社では目標となる音韻に近い複数の素片を選択し，それらを融合することで，適切な素片を生成する独自の複数素片選択融合方式を用いている。この処理により，実際には適切な素片がコーパスに存在しない場合でも，目標に近い素片を生成し，この素片の韻律を変形して接続することで，音質劣化を防ぎ，肉声感のある合成音声を安定した品質で提供することができる（**図3**）。

3.2　通信プロトコルを使用したリアルタイム音声合成

音声ポータル装置と音声合成サーバの間の通信プロトコルとして，業界標準のMRCPv2に対応している。音声ポータル装置はアプリケーションサーバからVoiceXMLメッセージとしてガイダンスのテキスト情報を含んだ応答シーケンスを受信した後，このメッセージに対応する音声合成要求をMRCPv2で音声合成サーバに送信する。

ここでMRCPv2は，まずSIP（Session Initiation Protocol）で，音声合成するためのMRCPセッションと，合成音をストリーミングにより接続するためのRTPセッションの二つのセッションを確立する。次にMRCPセッションは音声合成す

（注1）　Linuxは，Linus Torvalds氏の日本及びその他の国における登録商標又は商標。

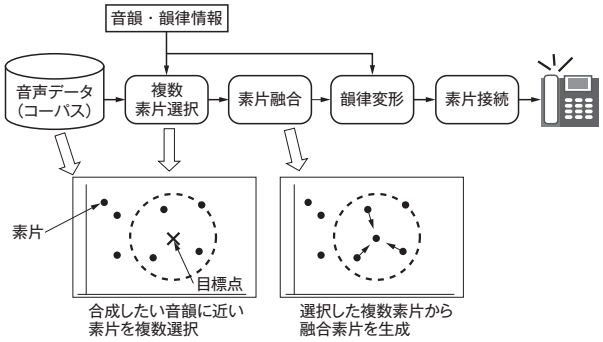


図3. 複数素片選択融合方式 ― 合成したい音韻に近い複数の音声素片をコーパスから選択及び融合し，目標となる素片を生成する。
Plural unit selection and fusion method

るためのテキスト情報及び，音声合成の際のアクセント位置や読みなどを指定する各種パラメータを音声合成サーバへ通知する（**図4**）。

テキスト情報や音声合成パラメータを指定するフォーマットはW3C（World Wide Web Consortium）で規定されているSSML（Speech Synthesis Markup Language）を使用している。音声合成パラメータとして，語句の強調範囲や，強調レベル，話速，声の高さ，抑揚の大きさ，音量，ポーズの挿入，読み方などを指定できる構成としている。

このように，音声合成パラメータをテキスト情報に合わせて音声合成サーバに対して送信することで，音声合成する際の

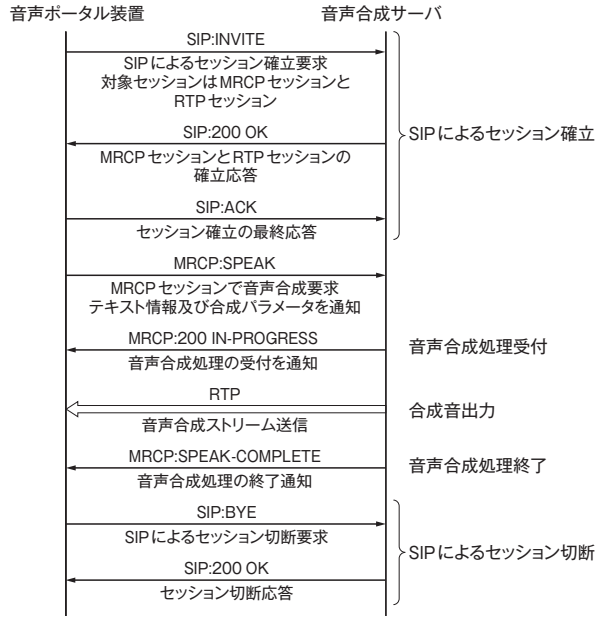


図4. MRCP基本シーケンス ― SIPを使ったセッション確立に引き続き，MRCPやRTPを使用して音声合成した結果をストリーミング送信する。
Basic sequence of Media Resource Control Protocol (MRCP)

合成音の細かな調整も可能にしている。したがって，VoiceXMLを生成するアプリケーションにより，発信者の様々な属性情報に応じたきめの細かい音声合成制御ができ，IVRを使用するケースでも場面に応じて従来よりも柔軟に自動応対することができる。

3.3 Web-GUIを使用したオフライン音声合成

アプリケーションサーバは，コールセンターの運用者に対してオフライン音声合成するためのWeb-GUI（Graphical User Interface）機能も提供する。

コールセンターの運用者は，パソコン（PC）のブラウザからこのWeb-GUIにアクセスしブラウザ画面上でテキストを入力することで，音声合成API（Application Programming Interface）を経由して音声ファイル出力を得ることができる。このファイルは，アプリケーションサーバ内部又はWeb-GUI操作を行ったPCに保存できる。

この音声ファイルを用いて，アプリケーションサーバは，VoiceXML内部にこの音声ファイルを組み込んだ応答シーケンスを生成し，音声ポータル装置に通知することもできる。定型の応答ガイダンスについては，MRCPv2による音声合成処理をつど要求するよりも，このWeb-GUIで生成された音声ファイルを繰り返し使用するほうがサーバリソースの有効活用という点で好ましい。

また，電話系業務だけでなくWeb系も含めたマルチコンタクトセンターを運営するコールセンターでは，ホームページに電話の応答ガイダンスと同じガイダンスを組み込むことで，電話系業務とWeb系業務の統一性を顧客にアピールすることができる。

Web-GUIを使用した音声合成機能では，MRCPv2で提供されない次の特長的な機能が提供される。

(1) 中間言語の生成　　中間言語は，人間が使用する自然言語のテキスト情報と合成音声との中間に位置する言語で，アクセント位置や，言語の読み，前後の言語情報に基づいた修飾情報などを明示的に示すための言語である。中間言語ではアクセント位置や，読み方，ポーズ位置などの細かな調整ができる。

Web-GUIを使用して音声合成したいテキストを入力し音声合成を実行すると，合成音声を試聴できるとともに，中間言語が生成される。試聴した音声のアクセント位置や読みなどを修正したい場合，この中間言語を用いてアクセント位置の修正などの編集作業を行う。

このWeb-GUIで調整した最終形の中間言語をSSMLのタグを使用して，前節で紹介したMRCPv2で伝達することにより，リアルタイムでの合成に活用することもできる。

(2) ユーザー辞書　　音声合成をコールセンターの業務に適用するにあたっては，今後生じる造語やコールセンターで取り扱う商品名などの新しい用語のアクセント位置や読みの情報を，辞書として追加登録する必要がある。ユー

ザー辞書では対象となる語句，その読み，及びアクセント位置を指定することができる。

また，あるアカウントで登録したユーザー辞書は，そのアカウントに関連して生成されるMRCPv2の音声合成要求でも適用される。

3.4 RTP生成

音声合成エンジンが生成した音声データをRTPパケットに格納して，音声ポータル装置に対して送出する。長文のテキスト情報の場合でも顧客に合成音声が聞こえ始めるまでの応答性を良くするため，音声合成エンジンが生成した音声波形データの出力開始直後から，そのデータを適宜RTPパケット化して送出している。またRTPのペイロードフォーマットとして，電話システムで広く用いられているITU-T（国際電気通信連合電気通信標準化部門）勧告のG.711 μ-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。

3.5 冗長構成

MRCPv2セッションの接続経路と，Web-GUIで使用する音声合成APIの接続経路について冗長化できる構成にした。

MRCPv2セッションの接続経路では，音声ポータル装置に対してActiveな音声合成サーバが複数台配置され，その間をネットワーク上の負荷分散装置で接続する。音声合成サーバで障害が発生して音声合成サービスを提供できなくなった場合，ネットワーク上からそのサーバが切り離され，残った正常な音声合成サーバで音声合成サービスを継続的に提供する。

また，音声合成APIの接続経路ではアプリケーションサーバを二重化できる構成にしており，各々のアプリケーションサーバから各音声合成サーバにAPI接続できる。これにより，アプリケーションサーバで障害が発生した場合にはアプリケーションサーバの切替えにより，音声合成サーバで障害が発生した場合には音声合成サーバの切替えにより，Web-GUIを使用した音声合成サービスが維持される。

4 音声合成システムの性能評価

MRCPv2による音声合成方式の優れたリアルタイム応答性に対して性能評価を行った。評価は，3種類のテキストサンプルを用いて音声合成要求の開始からRTPストリーム出力が開始されるまでの応答時間を測定し，検証した（**図5**）。

500文字，2,000文字規模のテキストの音声合成要求に対して，同時要求セッション数の増加にほぼ比例して，応答ガイダンスの応答開始時間が増加しているが，上限最悪値と想定している3,000 msをクリアする良好な結果を得た。50文字の短文テキストに対しては，ほぼ均一な応答時間となり，応答性を早めるための処理が有効に機能している。

以上の結果から，短文から長文に至るまで，リアルタイムの音声合成に適用可能であることが確認できた。

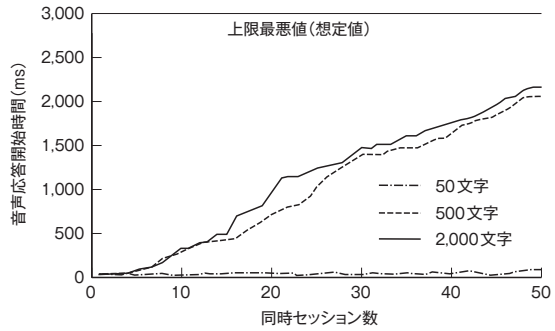


図5．音声合成の応答性評価 — 音声合成開始指示から音声応答までの時間は，同時セッション数（コール数）が増えても，上限最悪値と想定した3,000 ms以内であり，短文から長文まで有効に機能している。

Evaluation of speech synthesis responsivity

5 あとがき

コールセンター向け音声合成システムは，当社で長年培ってきた音声合成技術をコアとし，MRCPv2及び，Web-GUIのための音声合成APIの通信インタフェースを付加したシステムである。このシステムにより，コールセンターの電話受付業務で顧客に応じた応答シーケンスや応答ガイダンスのカスタマイズが可能になり，従来のIVRシステムに関わるコールセンターの業務負荷を低減することができる。

MRCPv2標準仕様への対応ができたことで，今後は市販の汎用音声ポータル製品との接続検証を進めていくとともに，日本語の音声種類の増加及び多国語への対応にも取り組んでいきたい。

文 献

(1) 平林 剛 他．次世代音声合成システムToSpeak$_{TM}$ V2を支える多様性向上技術．東芝レビュー．**65**，4，2010，p.43－47.
(2) 籠嶋岳彦．高音質で聞きやすい音声合成システム ToSpeak$_{TM}$．東芝レビュー．**62**，12，2007，p.34－37.

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
東芝放送ネットワークエンジニアリング（株）ソフトウェアエンジニアリング部。IP電話システムの開発・設計に従事。
Toshiba Broadcasting and Network Engineering Corp.