

# 音声ポータルソリューション

## Voice Portal Solutions

### あらまし

人間とコンピュータとのインタフェースを考えると、デジタルデバイド、アクセシビリティの観点では、音声によるインタフェースは自然なコミュニケーション手段を実現させる重要な役割を持つ。

本稿では、音声インタフェースの応用分野の中でも、とくに電話を用いた情報アクセスシステムについて概観し、音声認識・音声合成・対話制御技術を用いて各種情報提供サービスを行う音声ポータルソリューションを紹介する。また、そこで使われる各音声処理技術について解説する。音声ポータルでは、とくに音声の特性を考慮した開発が重要である。そこで、音声認識では、不要語への対応、背景雑音や電話回線による音質劣化への対策を強化した。また、音声合成では、コーパスベース音声合成方式と、波形辞書自動生成技術により、自然で多様な声質の音声応答を実現した。音声ポータルは、自治体の情報提供サービスや、株価・交通情報といったリアルタイム情報提供サービスなど、幅広い応用が期待される。

### Abstract

The voice user interface is an important tool for realizing natural, easy-to-use human-computer interfaces that can bridge the digital-divide. This paper gives an overview of voice interface applications and information access systems that operate over a telephone network. It also introduces some voice portal solutions that use speech recognition, text-to-speech technology, and dialogue control and outlines the speech processing technologies used in voice portal solutions. The developers of a voice portal system must understand the characteristics of speech and ensure that the speech recognition part can cope with irrelevant words, background noise, and the distortions caused by telephone lines. Text-to-speech technology produces natural speech and other speech styles by using a corpus-based method and automatically constructing a waveform database. The information services of municipal offices, traffic information services, stock price services, and so on are widely expected to incorporate voice portal systems in the near future.



岩見田均(いわみだ ひとし)  
ITメディア研究所メディアソリューション研究部 所属  
現在、音声認識・音声対話の研究に従事。



渡辺一宏(わたなべ かずひろ)  
ITメディア研究所メディアソリューション研究部 所属  
現在、音声合成の研究に従事。



伊藤 映(いとう はゆる)  
ITメディア研究所メディアソリューション研究部 所属  
現在、音声対話の研究に従事。

## まえがき

音声は、人間にとって最も自然なコミュニケーション手段である。人間とコンピュータとのインタフェースを考えると、現在主流のキーボード・マウスとGUIによるインタフェースでは、デジタルデバイス、アクセシビリティの観点で課題がある。音声によるインタフェースは、これらの課題を解決する上で重要な役割を持つ。

音声をコンピュータで処理する音声認識・音声合成の研究は数十年前から始まっている。音声認識・音声合成の研究の進展と、マイクロプロセッサの発展により、今日、パソコン（ディクテーション、コマンド）、携帯電話、カーナビ、CTI（Computer Telephony Integration）などの分野で多くの製品が現れるようになってきた<sup>(1)</sup>

本稿では、音声認識・音声合成の応用分野の中でも、とくに電話を用いた情報アクセスシステムについて概観し、音声認識・音声合成・対話制御技術を用いて各種情報提供サービスを行う音声ポータルソリューションと、そこで使われる各音声処理技術について述べる。

## 音声情報アクセス

音声を用いた情報アクセスのシステムは、古くから検討が行われており、ユーザ側の入力手段として、電話のプッシュトーンを利用したものが古くから用いられている。富士通は、音声合成技術を用いた、音声FAX応答システムVoiceScript<sup>(2)</sup>を開発した。このシステムは、音声情報アクセスシステムとして様々な分野で利用されている。

音声認識を用いたシステムは、1990年代半ばごろから製品が発表され始めた。その後、携帯電話の急激な普及の後押しもあり、2000年以降、音声ポータルと称する各種情報提供サービスを行うシステムの試行・実運用が拡大の一途をたどっている。音声ポータルサービスの世界市場は2005年には123億ドルになるとの予測もある<sup>(3)</sup>

音声ポータルに用いられる音声対話のシナリオ作成については、これまでは個別の仕様に基づくものが多かったが、最近になり、音声対話制御の標準仕様としてVoiceXML（Voice Extensible Markup Language）<sup>(4)</sup>が登場し、さらに、マルチモーダル

を意識した仕様としてSALT（Speech Application Language Tags）なども登場してきた。今後、これらの標準化仕様に準拠した音声対話システムが急伸することが予想される。

## 音声ポータルソリューション

富士通では、自社開発の音声認識・音声合成と対話制御技術を用いた音声ポータルソリューションのための試作システムを開発した。

### 構成

音声ポータルの試作システム（以下、本システム）の構成を図-1に示す。パソコンなどでアクセスする既存のWebシステムのコンテンツを流用して対話シナリオを生成し、対話制御エンジンの指令により、音声認識・音声合成エンジンを動作させながら、ユーザの望む情報を提供する。既存の電話網を利用するので、ユーザは、いつでもどこでも手軽に電話での情報アクセスが可能となる。

### 特長

本システムは、単に既存システムを音声で置き換えるという発想ではなく、音声というモダリティ（コミュニケーション様式）の特性を考慮した以下の特長を持つ。

#### (1) 音声認識エンジン

音声を発話するときに、特有の不要語や文末の言い回しのパリエーションへの対応に優れているワー

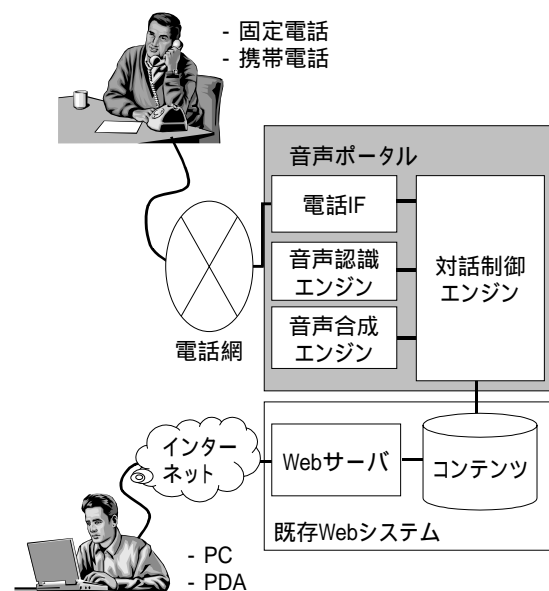


図-1 音声ポータルの試作システムの構成  
Fig.1-Trial voice portal system.

ドスポッティング技術と、背景雑音や電話回線による音質劣化への対策を強化したことにより、高い電話音声認識性能を実現した。

(2) 音声合成エンジン

大量の音声データ（コーパス）を利用して、所望の合成音声を生成するコーパスベース音声合成方式と、音声合成で用いる波形辞書を自動的に生成する波形辞書自動生成技術により、自然で声質のバリエーションに富んだ音声応答を実現した。当音声合成エンジンは国内で多くの導入実績を持つ。

(3) 対話制御エンジン

音声対話制御の標準仕様であるVoiceXML2.0に準拠したインタプリタを開発し、システム主導対話、ユーザ主導対話など、ユーザの多様な話し方に対応した対話シナリオの実現を可能にした。

応用例

本システムは、政府・自治体の情報提供サービスや、株価・交通情報といったリアルタイム情報提供サービスなど、幅広い応用が期待される。とくに自治体などの公的機関のサービスでは、デジタルレバイド対策としても大きな意味を持つ。音声ポータルサービスを応用した対話例を図-2に示す。従来は、一問一答型でユーザは単語発声のみという対話が多かったが、本システムでは、対話例に示すように、一度に複数の項目を含む自然な発声が可能となった。

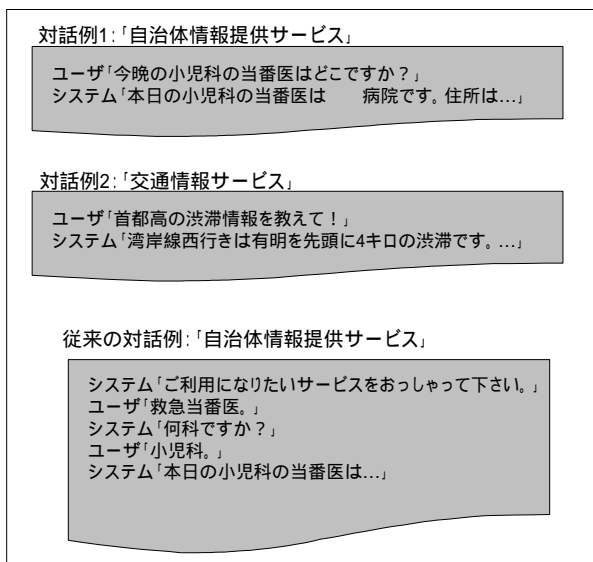


図-2 音声ポータルサービスを応用した対話例  
Fig.2-Dialogue examples of voice portal applications.

音声認識技術

本システムの高い電話音声認識性能を実現させた音声認識技術を以下に示す。

音声認識の構成

音声認識システムの一般的な構成を図-3に示す。入力された音声は、音声区間検出や雑音抑圧などの音響処理が施された後に、特徴量が抽出される。その特徴量をもとに、音響モデルと言語モデルを用いて照合処理を行い、認識結果を得る。音響モデルは、母音・子音などの単位で音の特徴をモデル化したもので、隠れマルコフモデル（Hidden Markov Model：HMM）<sup>(6)</sup>が一般的に用いられている。言語モデルとしては、認識すべき語彙を記述した単語辞書と、単語のつながりを有限状態オートマトンで表現したネットワーク文法、または単語間のつながり度合いを確率的に表現したNグラム文法のような統計的文法が用いられる。

電話品質音声認識

音声ポータル向けの音声認識として、電話回線により品質の劣化した音声への対処や、様々な人が事前学習なしに音声認識を使えるようにすることなどが必須である。これらへの対処のために、富士通では、電話回線品質の大量の音声データを用いた音響モデルを作成し、各種環境での評価・改良を繰り返すことにより、電話品質の音声に対しても高い音声認識精度を実現することが可能となった。

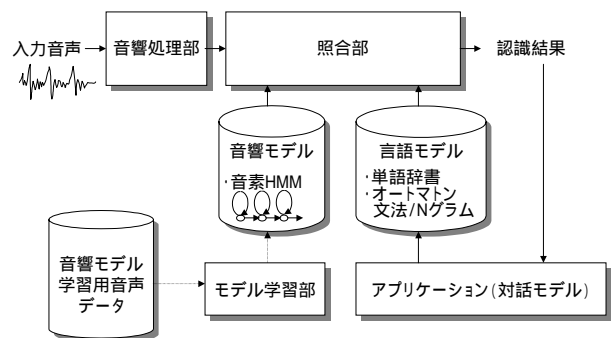


図-3 音声認識システムの構成  
Fig.3-Speech recognition system.

ワードスポッティング

音声認識技術において、自然な発声になればなるほど、その認識は難しくなる。例えば、ユーザの自然な発声には、「えー」「あー」といった不要語が多く含まれたり、文末の言い回しのバリエーション（例：「教えて」「教えて下さい」「教えてよ」）があったりするが、それらの単語は認識の対象となっていない場合があるため誤認識が生じ、意図どおりに対話を進められず、短時間で目的を達成できないことがある。そこで、富士通では、不要語などを無視できるワードスポッティング技術をもとに、高い認識性能を実現するための文法（単語並び）制約を加える、文法制約付ワードスポッティングを開発した。

ワードスポッティングとは、あらかじめ定めた単語（単語辞書）のみを音声から抽出する技術である。文法制約付ワードスポッティングの例を図-4に示す。ユーザが「浅田さんの、えーと、明日の予定を」と発声した場合、入力音声から単語辞書に含まれる「朝」、「浅田」、「明日」、「予定」と四つの単語を抽出する。ワードスポッティングだけでは、この「朝」のように、音響的に類似した単語が誤って抽出されることもあり十分な認識精度が得られない。そこで更に、文法の制約を加える。つまり、あらかじめ定義したオートマトン文法に従って、適合する単語並びを検索する。例では、文法制約により文頭の「朝」が削除され、「浅田 - 明日 - 予定」が認識結果となる。

音声合成技術

本システムにおいて、自然で声質のバリエーションに富んだ音声応答を実現した音声合成技術を以下

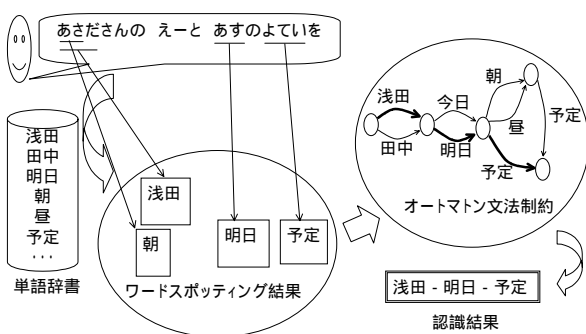


図-4 文法制約付ワードスポッティングの例  
Fig.4-Example of word spotting with grammar.

に示す。

技術開発のポイント

今般、音声ポータルやコールセンタの分野で、音声認識・音声合成技術を利用したシステムを導入して、オペレータを増員することなく、情報提供やチケット予約の処理件数を増やし、増収を図る動きが活発化している。このような動向に対応するためには、ユーザに不自然さを感じさせない合成音声の品質の実現と、様々なサービスに対応できるバリエーション豊かな合成音声の提供がポイントとなる。ここでは、著者らが開発した、自然性の高い合成音声を生成する「コーパスベース音声合成方式」と、バリエーション増強技術として「波形辞書自動生成技術」を紹介する。

コーパスベース音声合成方式

大量の音声データ（コーパス）を利用して、所望の合成音声を生成する方式を総じてコーパスベース音声合成方式（以下、本方式）と呼ぶ。音声合成システムは、言語処理部、韻律生成部、波形生成部から成り、その波形生成部に本方式を導入した（図-5）。入力テキストは、言語処理部および韻律生成部によって、合成する音素列と韻律（各音素の時間長、ピッチ、振幅のパターン）に変換される。波形生成部では、大量の音声データが蓄積された波形辞書から、合成する音素列や韻律に対して、より長く、接続点の不連続が小さくなる音声データを選択する。図-5の例では、「山梨県のJR中央線...」という入力テキストに対し、波形辞書から「山梨の高校から...」の「山梨」（音素列：y-a-m-a-n-a-sh-i）、「別

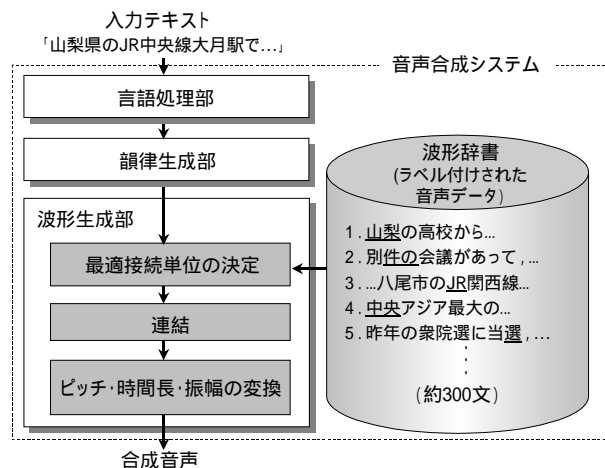


図-5 コーパスベース音声合成方式の概要  
Fig.5-Outline of corpus-based text-to-speech.

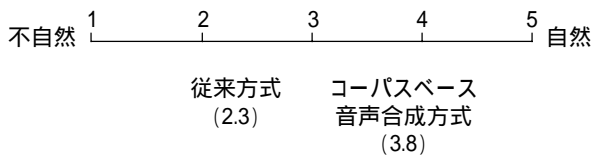


図-6 コーパスベース音声合成方式の評価結果  
Fig.6-Evaluation result of text-to-speech systems.

件の会議が...」の「件の」(音素列:k-e-N-n-o)などの音声データが選択される。このとき、入力テキストと波形辞書から選択するデータは、必ずしも同一の単語である必要はなく、音素列が一致していれば選択できるようにしている。波形辞書は、約300文の文章発声データと日本語の音声を網羅する約180個の音節発声データ(a, i, u, ...など)を含んでいる。選択した音声データを連結し、入力テキストに応じたピッチ・時間長・振幅に変換し、最終的な合成音声とする。固定長の音素波形を短い単位で接続する従来方式<sup>6)</sup>と比べ、本方式では可変長のより長い単位の音声を接続できるため、音質劣化を引き起こす接続箇所を大幅に減らすことができる。また、より長い区間で音素列が一致する音声データを使用することから、原音声の音質が保存されやすくなり、声質の自然性が向上する。声質の自然性をプリファレンススコア<sup>(注)</sup>により評価した結果を図-6に示す。本方式の評価は3.8であり、従来方式の評価2.3に比べると声質が大幅に改善されたことが分かる。

また、本方式の枠組みを用い、感情別の波形辞書と韻律生成モデルを用意することによって、「喜び」「怒り」「悲しみ」といった単純な感情を表現する合成音声生成できることを確認した<sup>7)</sup>

現在、本方式に基づく音声合成エンジンFineSpeechは、Microsoft Speech API 5.0に対応した音声合成ライブラリを用意しており、多くのIVR(Interactive Voice Response)システム、音声ポータルで採用されている。

波形辞書自動生成技術

音声合成技術が多様な場面で使われるにつれて、声質のバリエーションに対する要求が高まってきている。新しい声種に対応した波形辞書を作成するためには、新たに収録した音声データに音素ラベリン

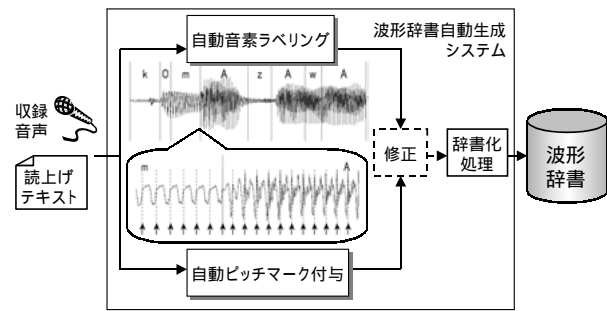


図-7 波形辞書自動生成技術  
Fig.7-Automatic construction of waveform database.

グ作業とピッチマーク付与作業を行う必要がある(図-7)。音素ラベリングとは、音声波形上に各音素の位置を設定することであり、ピッチマーク付与とは、周期性を持つ波形のピッチ位置をマークすることである。従来、これらの作業を人間の目と耳を頼りに手動により実施していたために、熟練者でも数か月の工数を必要としていた。この問題を解決するため、従来の手動作業を自動で行う波形辞書自動生成技術を開発し、ユーザの要望に迅速に応えられるようにした。

(1) 自動音素ラベリング

隠れマルコフモデルによる音素認識技術を用いて音声波形と音素の対応付けを自動で行う技術を開発した。また、無音部に音素ラベルが付与される誤りや、極端に短い音素ラベルが付与される誤りなどを修正する後処理部を開発した。

(2) 自動ピッチマーク付与

ピッチマークを付与する位置は、生成される合成音声の音質に大きく影響する。音声データから、各フレームの基本周波成分を自動抽出し、音質が最適となるピッチマーク付与位置を自動検出する技術を開発した。

いずれの技術も、発声が不安定な箇所などにおいては、手動に比べ精度が十分ではないが、この技術を用いて数時間の自動生成処理を行うことで、一応の合成音声出力することが可能になった。また、合成音声の品質を上げるためには、自動生成後に手動による修正作業を行う必要があるが、波形辞書自動生成技術は、波形辞書をすべて手動により作成する場合と比較し、大幅に工数を削減できることを確認している。

(注) 被験者の評価点(1~5の5段階)の平均値。

## 対話制御技術

音声対話シナリオ作成とその制御技術として、これまででは独自の仕様によるものが多かったが、1999年ごろよりXMLベースのマークアップ言語としてVoiceXMLというスクリプト言語の標準化活動が始まった。VoiceXMLは、目的達成に必要な情報を対話を通じて獲得するタスク（目的）指向言語であり、一問一答型によるシステム主導型の対話はもちろん、複数の項目を一度に発声したり、または複数の項目のうちの一部を発声したりして、不足分の情報をシステムが質問するといった目的達成の戦略をとることも可能である。また、VoiceXMLは、既存のWebシステムと高い親和性を持つという特徴がある。VoiceXMLで書かれた対話シナリオの例を図-8に示す。

富士通では、VoiceXML2.0に準拠した音声対話制御エンジンを開発した。本エンジンはインタプリタとインタプリタコンテキストの二つの処理ブロックで構成される。インタプリタは、VoiceXMLで書かれた対話シナリオを解釈する部分であり、インタプリタコンテキストは、プラットフォーム制御、音声認識・音声合成エンジンなどとのインタフェース制御を行う部分である。

音声対話システムの良否を決める上で、対話シナリオが洗練されていることが重要である。単純に既存のGUIなどでの対話を音声対話に置き換えるという発想ではうまくいかず、音声の特性を考慮した開発が必要である。例えば、ユーザに聞かせるプロンプトの表現や長さが適切であるか、対話の各場面

```
<form id= airplane_service >
  <grammar src= "案内.grxml" type = "application/grammar+xml" />
  <block>いらっしゃいませ、富士通航空時刻表案内サービスです。</block>
  <field name= "出発地" >
    <prompt> 出発はどちらの空港でしょうか? </prompt>
    <grammar src= "空港.grxml" type = "application/grammar+xml" />
  </field>
  <field name= "到着地" >
    <prompt>到着はどちらの空港でしょうか? </prompt>
    <grammar src= "空港.grxml" type = "application/grammar+xml" />
  </field>
  <field name="出発日">
    <prompt> 出発日はいつでしょうか? </prompt>
    <grammar src= "月日.grxml" type = "application/grammar+xml" />
  </field>
  <block>
    <prompt> ご利用ありがとうございました。 </prompt>
    <submit http= "http://www.vxml.com/servlet/airplane_service">
  </block>
</form>
```

図-8 VoiceXML対話シナリオ例  
Fig.8-Example of VoiceXML script.

における認識文法はユーザが発話する可能性のある文を網羅しているかといった点について、目的達成までの時間や達成率、被験者による主観評価などにより、総合的に評価する必要がある。富士通では、施設予約やスケジュール案内などの実アプリケーションを意識したプロトタイプシステムを作成し、社内で試行・改良を繰り返すことで音声対話シナリオに関するノウハウを蓄積し、音声対話を洗練させている。

## む す び

本稿では、音声インタフェースの応用分野の中でも、とくに電話を用いた情報アクセスシステムについて概観し、音声認識・音声合成・対話制御技術を用いて各種情報提供サービスを行う音声ポータルソリューションを紹介した。また、そこで使われる各音声処理技術については、音声の特性を考慮して開発した技術が音声によるコミュニケーションを実現する上ですぐれた特長を持っていることを述べた。今後、音声ポータルソリューションを様々な分野の応用に普及させることを目指し、自然なコミュニケーション手段を実現するための試行・評価を繰り返していきたい。

## 参考文献

- (1) 木村晋太：音声合成・認識技術の進展．FUJITSU，Vol.49，No.1，p.41-46（1998）．
- (2) 辻内秀敏：音声合成を用いた電話情報サービス構築ツール：VoiceScript．FUJITSU，Vol.49，No.1，p.57-60（1998）．
- (3) Global Information, Inc.ホームページ．  
[http://www.gii.co.jp/press/rd7497\\_jp.shtml](http://www.gii.co.jp/press/rd7497_jp.shtml)
- (4) The World Wide Web Consortiumホームページ．  
<http://www.w3.org/>
- (5) 中川聖一：確率モデルによる音声認識．初版，東京，電子情報通信学会，1988．
- (6) 片江伸之ほか：高品質音声合成技術．FUJITSU，Vol.49，No.1，p.47-51（1998）．
- (7) 片江伸之ほか：感情音声合成における声質と韻律の制御の効果．音響学会講論集，2-1-7，p.187-188（2000.9）．