音声言語情報処理 33−12
（2000．10．27）

VoiceXMLインタプリタと連続単語認識エンジンの開発
−音声ポータル向け音声認識技術の開発−

鯨井 俊宏*、高橋 久**、天野 明雄*、畑岡 信夫*
*（株）日立製作所中央研究所、**（株）日立超LSIシステムズ
e-mail：kujira@crl.hitachi.co.jp

音声によるWebアクセスを可能とする音声ポータルが実用化され始め、音声アクセス専用のコンテンツ記述言語の標準化が活発に行われている。本報告では、音声ポータルの概念を明確にし、コンテンツ記述言語の標準化の必要性と、具体的な標準化活動について述べている。さらに、VoiceXMLに基づいて音声ポータルシステムの試作を行った。本報告では、システムを構成するVoiceXMLインタプリタと連続型電話音声認識エンジンについて詳細に報告している。

キーワード：音声ポータル、VoiceXML、ボイスブラウザ、標準化、連続単語認識

Development of VoiceXML Interpreter and Continuous Words Recognition Engine
—Development of Speech Recognition Technologies for Voice Portal−

Toshihiro Kujirai*, Hisashi Takahashi**, Akio Amano*, Nobuo Hataoka*
*Central Research Laboratory,Hitachi,Ltd., **Hitachi ULSI Systems Co., Ltd.

Voice Portal systems have been put into practical use and the standardization of languages for web access via voice are actively made. First, this paper describes the need for the standardization and activities for that. Secondly, this paper gives a description of our Voice Portal system besed on VoiceXML and its components, i.e. VoiceXML interpreter and Telephony Speech Recognition Engine.

Keywords：Voice Portal, VoiceXML, Voice Browser, Standardization, Continuous Words Recognition

1. 音声ポータルと音声コンテンツ記述言語の標準化

1.1 音声ポータル

最近、米国において音声ポータルサービスの実用化が始まっている。音声ポータルとは、Webコンテンツへ音声を用いてアクセスするための入り口を果たすものであり、実用化されているサービスは、主に電話からのWebへのアクセスを可能にしている。

コンテンツとしては、初めから音声でのアクセスを考慮して作成されたものと、従来のコンテンツを音声でアクセスできるような形式に変換したものの両者が考えられる。ここでは、これらをまとめて音声コンテンツと呼ぶことにする。

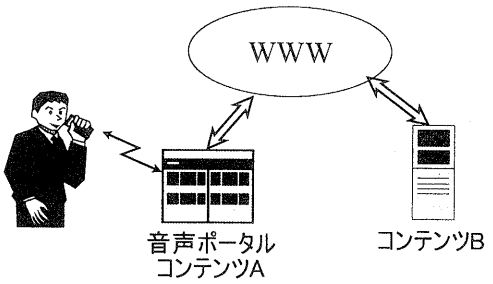

図1　音声ポータルシステム

図1は、音声ポータルシステムの一例である。音声ポータルは、音声認識システムと音声合成システムを備えており、コンテンツAに記述されているサービスをユーザに対して提供する。この例のように、音声ポータルとコンテンツAが同一のサブシステム上に存在し管理されている場合、音声ポータルとコンテンツAの間のインタフェース仕様はサービス業者が任意に設定することができる。実際、従来のCTI(Computer Telephony Integration)システムでは、各社が独自のインタフェースを設定していた。これらのインタフェースはAPI(Application Program Interface)のレベルではいくつかの標準化候補が提示されている([1][2])が、未だ統一されていないのが現状である。音声コンテンツ提供者もしくは音声コンテンツ作成の代行者は、利用する音声ポータルに合わせて、これらのAPIやコンテンツ記述言語を習得する必要があった。

これに対して、現在のHTMLで記述されているWebコンテンツのように、コンテンツBが任意のサーバに置かれており、複数の音声ポータルからのアクセスが可能な場合、コンテンツ記述言語のレベルで、なんらかの標準規格が必要なことは明らかである。また、このような標準化によって、コンテンツ提供者は音声ポータル独自のAPIや、コンテンツ記述言語を複数習得する必要がなくなり、音声コンテンツの作成がより容易になると考えられる。

1.2 音声コンテンツ記述言語に要求される仕様

このような音声コンテンツ記述標準言語としては、以下のような特徴を持つことが望ましいと考えられる。

- ・プラットフォーム独立性、可搬性
- ・再利用性
- ・汎用性、拡張性
- ・現存技術による実現性
- ・容易性
- ・従来技術との整合性
- ・国際性

プラットフォーム独立性、可搬性

すでに述べたように、音声コンテンツ記述言語の標準化の第一の目的は、これまでプラットフォーム依存だったコンテンツの記述方法を統一することにある。このため標準化される記述言語はプラットフォームのインプリメンテーションとは独立したレベルで仕様が決定される必要がある。また、携帯端末では、サーバ型のシステムが提供するすべての機能を実現することは困難であるなど、プラットフォームの機能には差があるため、プラットフォームの機能に応じて、コンテンツの提示方法を変更する手段も含まれるべきである。

再利用性

音声コンテンツ記述言語を統一するメリットとして、よく利用される対話シーケンスを部品化することがある。これによって、コンテンツ提供者はコンテンツ作成の手間を大きく省くことができるとともに、ユーザはどのコンテンツに関しても、定型的な対話シーケンスにおいて、同様の操作性が保証される。

汎用性、拡張性

標準化される言語は、できるだけ汎用性を持つべきである。言語が記述できるサービスがあまりに限定される場合、独自の言語拡張が横行し、標準化の意義が薄れる危険性が考えられる。逆に、言語にサービス記述上の不備や不足があることが判明した場合、必要に応じて言語の拡張ができるようにしておくことも重要である。

現存技術による実現性

標準化される言語は、現存する技術によって実現可能な機能のみを記述できるように決められるべきである。

容易性

良い音声コンテンツを作成するためには、現在の音声処理技術について知識を持っていることが望ましいのは事実であるが、標準化される言語は、出来る限りコンテンツ作成者に処理プロセスを意識させないような記述が可能であることが望ましい。

従来技術との整合性

既存のWebインフラやツールを利用するために、言語はXMLベースで記述されることが望ましい。また、すでに存在する記述言語を用いて記述されたコンテンツの、標準言語への変換が可能であることが望ましい。さらに、言語の拡張が将来的に行われる場合、拡張された言語はベースとなった言語との整合性を持つべきである。

国際性

標準化される言語は、複数の文字コードセットや国語を扱えるべきである。

1.3 音声コンテンツ記述言語の標準化活動

次に、実際の音声コンテンツ記述言語の標準化活動について述べる。現在、音声コンテンツ記述言語の標準化活動は、VoiceXML Forum[3]と、W3C(World Wide Web Consortium)の Voice Browser Working Group(以下VBWG)[4]の2つがある。

VoiceXML Forum

VoiceXML Forum は AT&T, IBM, Motorola, Lucent が中心となって設立された団体である。VoiceXML Forumでは、各社がすでに持っていた音声コンテンツ記述言語を融合する形で、標準言語 VoiceXML の規格化を進めてきた。現在は、VoiceXML 1.0 が一般に公開されている。VoiceXML Forum では、言語の標準化とともに、その言語の一般への浸透を目的としている。

VoiceXMLの特徴としては、

- 主に電話を介しての利用を想定している(マルチモーダルは考慮されていない)
- 主に対話シーケンスについての記述を目的としている
- 自然言語理解処理についての記述ができない
- 音声合成に関しては標準仕様が決められているが、音声認識で用いる文法については決められていない。

などが挙げられる。

Voice Browser Working Group

VBWG は W3C の標準化ワーキンググループの一つである。VBWG が提唱する Voice Browser は、VoiceXML よりも広い概念を扱っている。以下に、VBWG を構成するサブグループを示す。

表1 VBWG のサブグループ

Grammar	文法表現を記述する
Reusable Dialog	日付、金額など再利用可能な対話シーケンスを記述する
NL Semantics	自然言語理解のためのセマンティクスを記述する
Speech Synthesis	音声合成に関する言語
DialogML	対話管理言語 VoiceXML に相当
Multimodal	携帯電話の画面等、音声以外の入出力の扱い

DialogML(Markup Language)サブグループでは、VoiceXML 1.0 をベースとして、音声コンテンツ記述言語の標準化を進めていく方針であり、今後 VoiceXML Forum との協力体制を作りあげていくことになっている。

2. VoiceXML インタプリタの開発

以上に述べたような状況の中で、我々は VoiceXML 1.0 に基づいた音声ポータルシステムの試作を行っている。VoiceXML では、システムを

図2のように3つの層に分類している。第一層は、呼処理や入出力インタフェースなどを管理するプラットフォーム、第二層は音声認識・合成エンジン、そして第三層が VoiceXML で記述されたコンテンツを解釈し実行するインタプリタである。

インタプリタの実装に関しては規定されておらず、各プラットフォームや音声認識・合成エンジンに特化した実装が許されている。

図3は、我々の試作した VoiceXML インタプリタの実行画面である。インタプリタは Visual Basic®によって実装されており、プラットフォームと音声認識・合成エンジンを含んだ機能を ActiveX®の形式で取り込んでいる。これによって実際に使用する ActiveX®を切り替えることで、電話回線経由での実稼働モードと PC 上でのテストモードを使い分けることができる。テストモードでは、音声認識エンジンの代わりにキーボードによるテキスト入力を行い、インタプリタと VoiceXML で記述されたコンテンツのテストを行うことができる。現時点では、エラーハンドリングなどのイベント処理と、ECMAScript を用いる必要のある変数処理に関しては未実装である。

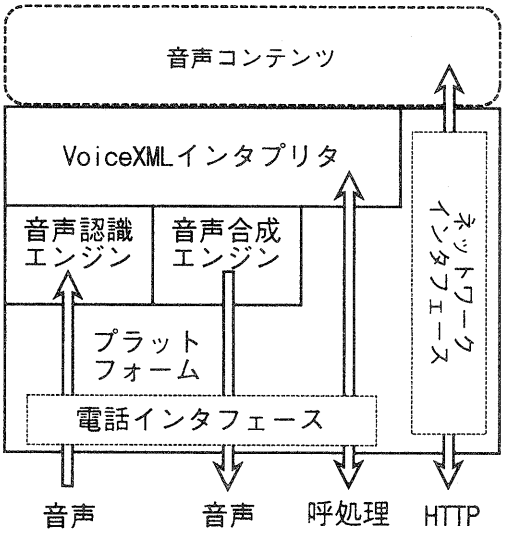

図2　VoiceXML システム概念図

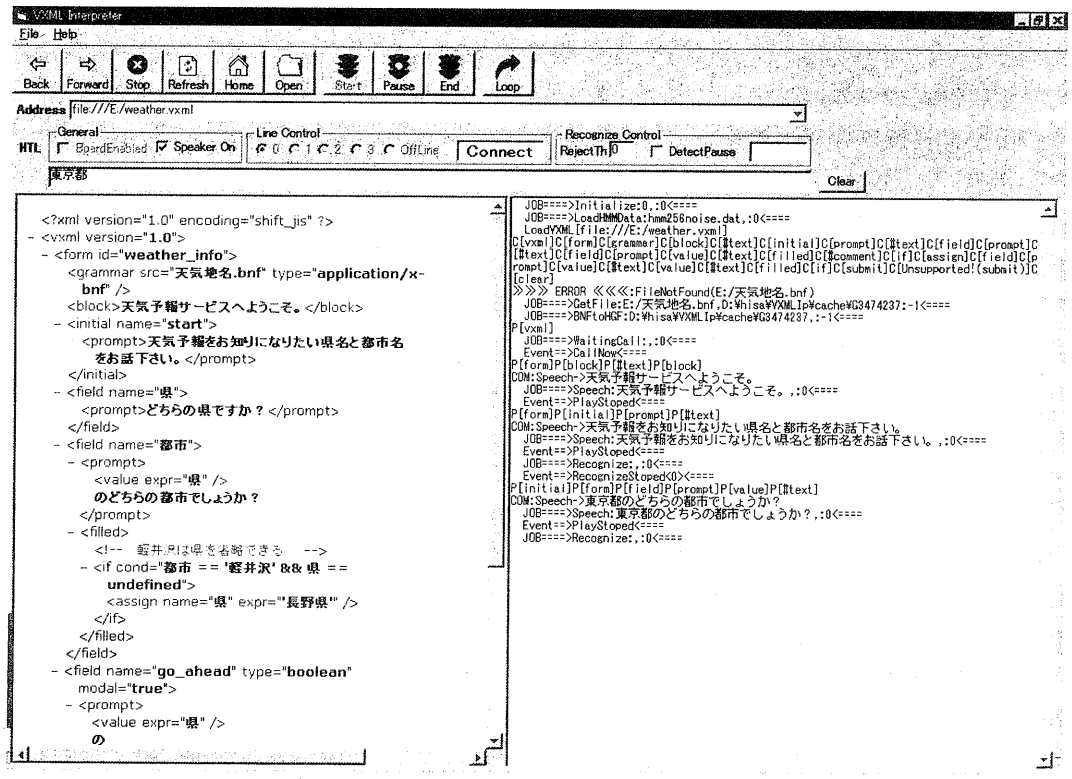

図3　VoiceXML インタプリタの実行例

図3で示しているのは、Mixed Initiative Form の例である。これは対話のシーケンスをユーザ主導にすることを可能にする機能である。この例では、県名と都市名を入力することで、その場所の天気を聞くことができる。さらに、入力が(i)県名のみの場合,(ii)都市名のみの場合,(iii)県名と都市名両者の場合,(iv)ヘルプを要求した場合など、それぞれに対して、異なった対応を行うことができる。これを実現しているのは、VoiceXML で規定されている FIA(Form Interpretation Algorithm)である。入力されるべき要素(県名、都市名等)には、すべて変数が対応しており、これらの値が空かどうかによって、実行する部分を制御し、複雑な対話シーケンスを記述することができる。

3. 連続単語認識エンジンの開発

本章では、図2で示したシステムの構成要素である、連続型電話音声認識エンジンについて述べる。

VoiceXML では認識エンジンに対する要求は、何らかの文法によって受理可能な発声を規定できること程度である。この要求を満たし、さらに実用的な対話型の音声認識システムを可能とするために、本システムの認識エンジンは以下の特徴を持っている。

- 電話機種による入力特性差の補償機能
- 文法ベースの連続単語認識機能
- ポーズを含む発声の区間検出機能

図4は、本認識エンジンの処理フローを示した図である。以下では、それぞれの機能について説明を行う。

3.1 電話機種による入力特性差の補償機能

電話応用の音声認識システムにおいては、ユーザが使用する電話機種を特定できないため、電話機種による入力特性差を補償する必要がある。この特性差は主に乗法性の歪みであり、ユーザのシステム利用期間中の変化は少ないと考えられる。

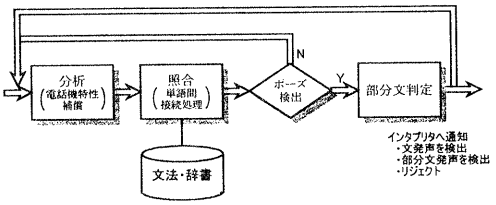

図4　電話音声認識エンジンの処理フロー

我々は、このような歪みに対して、高速かつ安定に適応する方式としてCMNをベースとした方式を開発し[5]、本システムにも適用している。

3.2 文法ベースの連続単語認識エンジン

VoiceXML では、特定の文法表現を標準形式として定めていない。しかし、何らかの正規文法をサポートして、システムが受理可能なユーザ発声を指定することが必要となる。そこで、 本システムでは BNF 形式の文法を採用した。

連続単語認識エンジンは、この文法に応じて入力された音声との照合を行う。この際、本エンジンでは単語間の接続において以下の処理を行い、認識率を向上させている。

- 音素片による単語間接続のスムージング
- 単語間ポーズへの対応

単語間のスムージングは、特に数字列など、文が短い単語で構成されている場合に効果が見られている。

3.3 ポーズを含む発声の区間検出機能

文の発声においては、挿入されるポーズが非常に長くなることが知られている[6]。このようなポーズは、単語認識で用いられるようなポーズ継続長しきい値による区間検出を用いることはできない。ポーズによって区間検出が発声の途中で行われた場合、部分的な発声は指定された文法によって受理されず、リジェクトされてしまう。

そこで本システムでは、発声中に存在するポーズの可能性を記述した文法を用いて、ポーズを含む発声の区間検出を行っている[7]。これによって、ポーズが検出された場合に、発声が途中であるか

どうかを検出し、途中である場合には、その先の発声を待つか、部分的な発声を用いて対話シーケンスを先に進めるかをインタプリタが選択できるようにしている。

この区間検出方式によって、表2に示した認識条件において、区間検出を従来のポーズ継続長で行った場合と、本手法で行った場合、さらに人間が発声を聴取して決定した場合の3種類について、評価実験を行った。その結果、以下のような結果を得ることができた。

部署・名前のタスクにおいては、区間検出に起因する文理解の誤りを94.5%、生年月日のタスクにおいても49.2%削減することができている。

表2 評価条件

評価条件	
分析	8kHz,Hanning窓 フレーム長20ms,シフト幅10ms LPCケプストラム14次 Δケプストラム14次,Δパワー
音声モデル	音素片HMM 各2状態, 2混合 音素バランス単語216を含む268語 100人名・1000駅名 男女各40名
評価データ	男性3名, 女性2名 部署と名前 100発声 生年月日 100発声

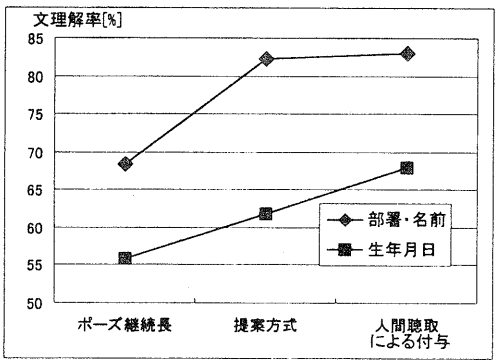

図5 音声区間検出による文理解率の差

4．まとめ

本報告では、まず音声コンテンツ記述言語標準化の必要性とその標準化動向とについて報告した。さらに、VoiceXML に基づく音声ポータルシステムの実現を目標として、VoiceXML インタプリタの概要と、インタプリタが使用する電話音声認識エンジンの詳細について報告した。

5．今後の課題

VoiceXML はバージョン 1.0 がリリースされているが、すでに機能の不足や仕様変更の要求が出されている。現在の仕様で何ができて何ができないのかは、具体的に音声コンテンツを作成した経験を通して分かってくる部分も多いであろう。音声コンテンツ記述言語の評価のために、様々な分野のアプリケーションを、標準テストアプリケーションとして定めていく必要があるのではないかと考える。

また、国際化の観点からは、日本語の文字コードでコンテンツをが記述できるといった表面的な部分だけではなく、欧米言語での対話と日本語での対話の違いなどについても考えていきたい。

参考文献

[1] Microsoft SAPI：
http://www.microsoft.com/iit/

[2] Java Speech API：
http://www.java.sun.com/products/java-media/speech/index.html

[3] VoiceXML Forum：
http://www.voicexml.org/

[4] W3C Voice Browser Activity：
http://www.w3.org/Voice/

[5] 鯨井ほか2名：パワーによるクラスタリングに基づくケプストラム平均正規化手法，音講論集,95-96,(1998.3)

[6] 高木ほか2名：模擬対話音声における各種区分の持続時間の性質，信学技報 SP92-111,63-70,(1992)

[7] 鯨井ほか1名：ポーズを明示的に表現した文法による区間検出，音講論集,45-46,(2000.9)