論文 / 著書情報
Article / Book Information

論題(和文)	音声コンテンツとVoiceXML
Title(English)	
著者(和文)	古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	Advanced Image Seminar 2002, Vol. , No. , pp. 59-66
Citation(English)	, Vol. , No. , pp. 59-66
発行日 / Pub. date	2002, 4

音声コンテンツと VoiceXML

古井　貞熙[*1]

Abstract - これからの高速ネットワーク環境において、音声ドキュメントを音声認識技術によって自動的に文字化することによって音声コンテンツを作成し、さらに音声を用いた対話システムによって、音声コンテンツを含む種々の Web ベースの情報コンテンツにアクセスする技術が重要になると予想される。本稿では、音声コンテンツを作成するための音声文字変換（音声認識）技術と、音声およびマルチモーダル対話システム構築のための記述言語 VoiceXML および SALT の動向について紹介する。

Keywords : 音声ドキュメント、音声認識、音声文字変換、情報抽出、音声対話、VoiceXML

1．まえがき

統計的パターン認識理論、ケプストラムとデルタケプストラム、隠れマルコフモデル、統計的言語モデルなどを基本とする大語彙連続音声認識の技術が近年大きく進展し、音声ワープロ、放送への自動字幕付与、音声対話システム、情報検索への適用など、種々の応用が展開しつつある[1-3]。音声認識技術の主たる応用は、比較的長い音声ドキュメントを対象として音声から文字への変換を目的とするシステム（トランスクリプション、音声文字変換）と、音声によるコンピュータとの対話システムに分類できる。図1に、これからの高速ネットワーク環境での、音声認識による音声コンテンツ作成と、音声対話による情報検索のプロセスを示す。

音声は、人と人との最も自然で、基本的、容易かつ効率的な情報交換手段である。多数の人を対象とした講演、講義、解説などでは、これからも音声が基本に用いられていくであろう。コンピュータへの入力手段としても、音声が使えれば極めて効率的になると期待されている。特に日本語に関しては、使いやすい優れたワープロ（ソフト）が普及しているとはいえ、仮名・漢字変換を必要とするため、英語のようにしゃべる速さに近い速度でタイプ入力するのは極めて難しい。音声認識技術を用いた日本語の音声文字変換ができれば、音声ワープロ、放送への自動字幕付与の他、会議の議事録、講演録、講義録、放送ニュースの音声コンテンツ化への利用など、数多くの応用が考えられる。音声文字変換の技術は、音声対話システムの技術の基礎としても重要である。

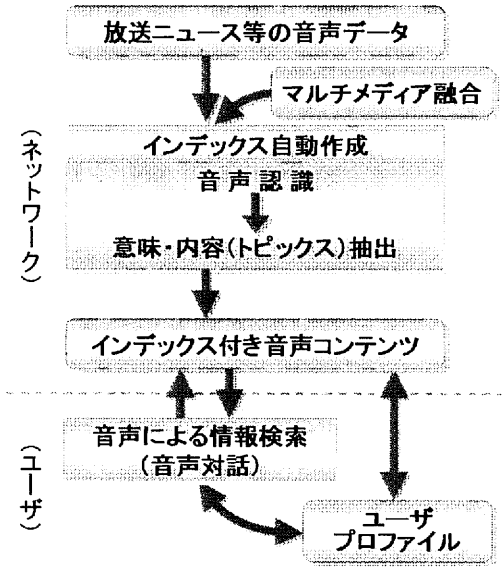

図1　高速ネットワーク環境での音声情報検索

音声対話技術は、自然なインタフェースを提供するということと、音声認識誤りが対話のプロセスによってカバーできるという、二つの面から重要な研究対象になっている。応用場面と

*1：東京工業大学大学院情報理工学研究科計算工学専攻 教授 (furui@cs.titech.ac.jp)

しては、両手がふさがっている車の中での車載情報サービス、すなわちカーナビからネットワーク型システムへの展開が、安全性、操作性、利便性の観点から、現在最も注目されている。携帯電話や車載機器を用い、利用者の位置情報を取り込んだ新しいサービスも検討されている。このような背景の下に、近年、電話音声を用いて Web アクセスをインタラクティブに行う音声ポータルシステム（図2）を容易に作成できるようにすることを目的に、WWW コンソーシアム（World Wide Web Consortium; W3C）で、インタラクション（ユーザとシステムのやりとり）を明記する言語として、VoiceXML (Voice eXtensible Markup Language)が検討されてきている。最近では、Microsoft を中心として、マルチモーダルの要素も取り込んだ SALT (Speech Application Language Tags)を提唱するフォーラムも活動を開始している。

以下では、音声コンテンツを作成するための音声文字変換技術と、対話システム構築のための VoiceXML および SALT の動向について紹介する。

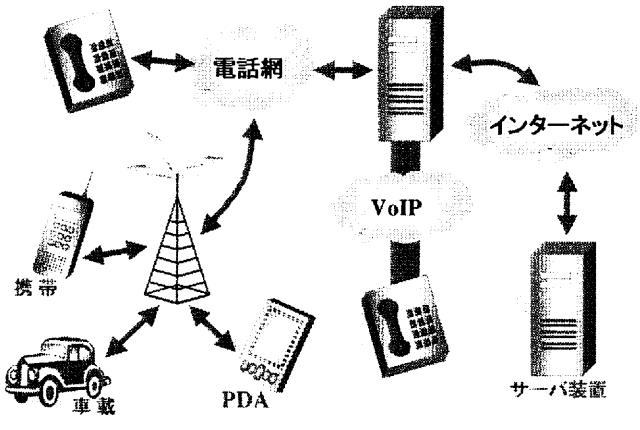

図2　音声ポータル環境

2. 音声文字変換研究のこれまで

2.1 読み上げ音声を対象とした研究

近年の音声文字変換研究の牽引車は、Wall Street Journal を始めとする北米の種々の新聞を読み上げた音声を対象として行なわれた DARPA (Defense Advanced Research Projects Agency) のプロジェクトである。このプロジェクトは 1991 年に開始され、文脈依存音素 HMM と、単語を単位とする統計的言語モデルを基本とする、数万語規模の大語彙連続音声認識の研究が活発に行なわれた。メルケプストラム、状態共有 HMM、言語モデルのバックオフ平滑化など、現在の音声文字変換の基本技術の多くは、このプロジェクトによって確立したと言える。それまでの人手に頼ることを前提にしていた文法を用いるよりも、認識性能が大きく向上し、65,000 語の語彙を対象とした場合に、93%の単語正解精度が報告されている[3]。

我が国においては、単語ではなく、音素、音節、仮名・漢字などを単位とする統計的言語モデルを用いた研究が、まず行なわれた。筆者らが NTT の研究室で、単語（形態素）を単位とする日本語の音声文字変換の研究に着手したのは、1995 年の３月である。NTT の研究所で優れた形態素解析プログラムが開発されると同時に、その開発に使われた日経新聞の５年間にわたるテキストが、電子的に使えるようになったことが、研究開始のきっかけとなった。直ちに、研究用データベースの構築にとりかかり、これを用いた音声認識実験を始めた。言語モデルとして単語トライグラムを用い、英語の場合にほぼ匹敵する認識結果を得ることができた[4]。

その後我が国では、新聞の読み上げ音声を対象とした音声文字変換研究に関して、国内のボランティアグループによるデータベースとデコーダの開発が進められ、「日本語ディクテーション基本ソフトウェア」の形で成果がとりまとめられている[5]。これは、認識エンジン（Julius）、日本語音響モデル、言語モデル、及び形態素解析・読み付与ツールから構成され、優れた性能を示している。無償で一般に公開されており、大語彙連続音声認識研究及び開発の共通のプラットフォームとして広く使われている。

2.2 放送音声の音声文字変換

1996年早春のDARPAワークショップでは、新聞記事の読み上げ音声から、放送音声の文字変換に焦点が移った。新聞読み上げ音声との

基本的違いは、対象とする音声が「実世界の音声データ」であることである。アナウンサーが原稿をもとに静かな環境で発声した音声だけでなく、音楽・雑音・音声が重畳・混在している音も対象としている。米国を中心に、欧州を含む大学及び研究機関がプロジェクトに参加し、語彙数は 65,000 語が標準的で、言語モデルの作成には、CMU で開発されたツールが広く使われている。

この DARPA の動きに刺激され、日本でも NHK 放送技術研究所を中心として、複数の大学を含むグループで、共同で放送ニュースの音声文字変換研究が始められた。過去５年間以上の放送ニュース原稿データベースと、実際の放送音声が用いられている。統計的言語モデルとしては、ニュース原稿に含まれる比較的頻度の高い 20,000 種類の単語のトライグラムを用いている。日本語では同じ単語に複数の読みがあり、正しい読みは前後関係（文脈）から決定されるため、同じ表記でも読みの違う単語は別の単位としている。更に、話し言葉には必ず挿入される「え」「えー」などの間投詞が、文頭や読点の位置に生じやすい性質に着目して、言語テキストのこれらの位置に間投詞の発音を確率的に与えている[6]。

NHK では、音声認識を用いてニュース音声に字幕を付与するシステムの運用を、2000 年 3 月から開始した[7]。音声認識では、誤認識を完全に回避することはできないため、認識結果を人が即座にチェックして、誤認識をすばやく修正してから放送するシステムになっている。修正がすばやく実行できるレベルとして、95%以上の単語正解精度が必要で、アナウンサーが原稿を読んでいる音声に関してはこれが達成されているが、それ以外のインタビューなどへの音声認識の利用は今後の課題である。ニュースでは日々新しい語彙（未知語）が生ずるため、これに対処するための方法や、文末にデコーダの処理が到達するまで認識結果が確定しない従来の方法では時間遅れが大きくなってしまうので、文中でも常に一定の遅れ以内で認識結果を確定しながら処理を進めるデコーダなどが開発されている。

最近では、DARPA のプロジェクトを中心に、音声認識と情報検索の技術を組み合わせる研究が、活発に行なわれている。放送ニュースの音声を音声文字変換した結果から、その話題（トピックス）を表す重要語やフレーズを自動的に検出し、それを用いてインデクシングを行っている。これにより、大量のニュースを話題に応じて分類するなど、自動的にアーカイブ（データベース）化することができる。放送やビデオの中から、音声の部分だけを取り出す研究や、ある特定の話題に関する部分だけを取り出す研究も行なわれている。

2.3 ディクテーションソフト（音声ワープロ）

音声文字変換ソフト（音声ワープロ）に関しては、1990 年に米国の Dragon Systems が単語区切りで発声した音声を入力とする PC 用プログラムを発売し、その後、IBM、L&H、Philips などの会社が参入した。単語区切りでないと入力できないのはユーザへの負担が大きいので、1996 年頃から単語を続けて発声できる製品が開発されるようになった。これらの製品の多くは、英語版だけでなく、ドイツ語、フランス語、イタリア語、スペイン語、中国語、日本語版などに発展している。主な利用者は、障害者、X 線医師（医療所見の入力）、弁護士などである。日本では、日本ＩＢＭが米国のＩＢＭと協力して、日本語音声のディクテーションソフトの研究開発を進め、今日の製品へ結実している[8]。やや遅れてＮＥＣでも同様の製品が開発されている。

3. 話し言葉の音声文字変換の実現へ

3.1 話し言葉認識の難しさとコーパスの構築

現在の音声認識技術で、自然な話し言葉音声に対して大幅に認識率が低下する原因は、発声が必ずしも明確でないこと、雑音が重畳していること、原稿を読んでいない自発的な話し言葉の音声は、言語モデルの学習に用いたニュース原稿の文とはかなり構造が異なることなどである。話し言葉特有の難しさには、助詞などの言葉の脱落・省略、間投詞などの不要語の付加、倒置、言い間違い、言いよどみ、言い直し、繰返しなどがある。

統計的言語モデルを構築するためには、認識対象に対応した大量のテキストが必須である。書き言葉のテキストは最近では比較的容易に得ることができるが、話し言葉に対しては、実際の音声を人手で書き起こすことが必要である。米国では、DARPA のプロジェクトの中で、”Switchboard”、”Call Home”などと呼ばれる、自然な会話音声を書き起こしたコーパス（データベース）が作られ、これを用いた音声認識研究が行なわれている。我が国でも1999 年から、科学技術振興調整費開放的融合研究推進制度による「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」プロジェクトが開始され、5 年計画で、音声認識の高度化のための、700 万形態素規模の音声言語データベース（コーパス）の構築、解析、認識、要約などの研究が推進されている [9]。組織の主体は、国立国語研究所、総務省通信総合研究所、および東京工業大学であるが、国内外の大学や研究機関からの研究者が非常勤研究員として参加している。

3.2 話題語抽出と音声要約

筆者らは、ニュース音声を自動的に音声文字変換し、その結果としての単語列から話題語（重要語）を自動抽出する試みを行っている[6]。この際、ニュースの話題語はほとんどが名詞であるので、文字変換結果から名詞のみを抽出し、その中から、重要度の尺度によって話題語を抽出する方法を検討した。重要度の尺度は、大量の学習用ニュース記事に出現する各名詞の出現頻度に基づく情報量によって定義した。情報量の大きい単語を上位から順に抽出し、重要語が精度よく抽出できることを確認している。

筆者らはさらに、音声中から話題を担っている話題語をできるだけ文法的・意味的制約に従うように最適に選んで要約文を作成する方法として、各単語の重要度スコア、音声認識結果の信頼度、要約文の統計的言語スコア、単語の意味的関係をもとに、動的計画法を用いる方法を提案した[10]。統計的言語スコアは、音声よりも比較的簡潔な文体となっていると考えられる新聞（毎日新聞）の１年分の記事から計算した単語トライグラムを用いた。認識結果の信頼性は、事後確率によって求め、意味的関係は、係り受け関係によって求めた。放送ニュース音声や学会講演の音声文字変換結果を用いて、文字数にして 20～80%に要約する実験を行っている[10]。本方法は、ニュースの自動字幕作成への適用のほか、講演録・議事録作成など、種々の応用分野への展開が期待できる。これらを、音声認識技術を用いて自動的に要約してデータベース化することができれば、情報検索に限らず種々の用途に極めて有用であろう。講演の要旨を短時間で聞くことができる、「早聞きシステム」の実現も不可能ではない。

3.3 話し言葉の言語的変動への対処

音声認識の応用場面が変わるごとに、言語モデルの学習のために大量の（書き起こし）テキストを集めるのは決して容易ではない。このため、共通の基本的な言語モデルを、限られた量のテキストを用いて、異なる対象に適応化させる研究も行なわれている。言語モデルの適応化のためには、テキストや言語モデルの中で、タスクに依存する部分と、依存しない一般的な部分を分ける必要があり、このような研究も始まっている[11]。

新しいタスクへの適応と関連して重要な課題に、新しい言葉（未知語）をいかにして言語モデルに組み込むかという問題がある。最も一般的な方法は、あらかじめ単語を品詞や意味を用いてクラスに分類しておき、未知語をその中の一つのクラスに対応させてクラス言語モデルを用いる方法である。単語クラスの作り方としては、細分類された品詞、ゆう度最大化基準、潜在的意味解析（latent semantic analysis）、エントロピー最大化基準、相互情報量、シソーラスなどを用いる方法や、これらを組み合わせる方法がある。

4. VoiceXML

電話を用いて、株価、天気予報、航空情報、道路情報、レストラン情報など、Web ベースの情報コンテンツが引き出せる音声ポータルサービスが、米国を中心に広く使われるようにな

ってきた。日本でも、NTTComのVポータル、日本テレコムのVoiziなど、サービスが最近始まっている[12]。これによって、コールセンターなどへの顧客からの要求に、人手をかけることなく音声応答で答えられるようになるので、通信（特にワイヤレス通信）、インタネットプロバイダ含め、各方面から注目されている。ビジネスモデルとしては、小売モデル（B to C型）だけでなく、卸売モデル（B to B型）もある。コンテンツとしては、始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ある。

従来のCTI (Computer Telephony Integration)システムのように、音声認識および合成機能を備えた音声ポータルとコンテンツが同一のシステム上に存在し管理されている場合は、音声ポータルとコンテンツ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た。これらのインタフェースは、API (Application Program Interface)のレベルでは、Microsoftの SAPI (Speech API)など、いくつかの標準化候補が提示されているが[13][14]、未だ統一されていない。このため、音声コンテンツ提供者（プロバイダ）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。

音声対話の記述言語に関する標準化ができれば、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でも、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、音声コンテンツ作成が容易になる。これにより、情報サービスにおいてサービスを始めようとするコンテンツ提供者が、音声認識、音声合成、電話回線制御などの高度な機能を利用して、簡単に電話音声対話システムを構築できるようになる。ユーザの側からは、PC、PDA、公衆電話、携帯電話など、種々の端末で、同様のサービスが受けられるようになる。

この活動として、現在W3Cで標準化が進められているVoice XMLがある。VoiceXMLは、音声対話によるユーザとシステムとのやりとりを記述する言語で、音声でインタネット情報をアクセスする上で、重要な記述言語となると期待されている。VoiceXML Forum（AT&T、IBM、Motorola など60社が参加）が標準化を始め、W3Cによって2001年12月にVoiceXML Version 2.0として、詳細な仕様が制定されている[15][16]。図3に、VoiceXMLシステムの構成を示す。

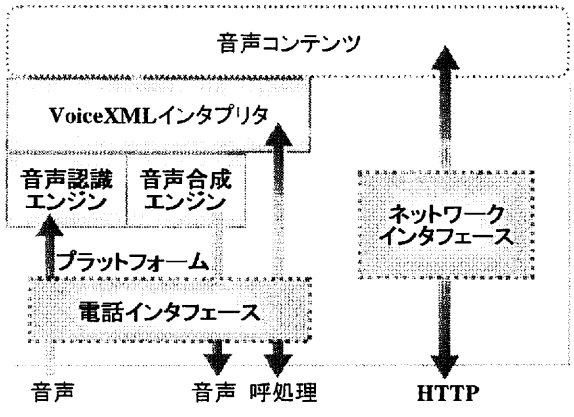

図3　VoiceXMLシステムの構成

VoiceXMLでは、ユーザに示すプロンプト、ユーザの発話文法、対話シナリオの制御など、対話に必要な様々な事項が記述できる。VoiceXMLは、まず<prompt>をユーザに示し、ユーザが<grammar>に従って入力し、それに対してシステムが<filled>内部を実行する、という対話形式を基本としている。すなわち、いわゆる「システム主導」（システムがユーザに入力の催促や質問を行い、それに対してユーザが答える形式）の対話が書き易く設計されている。最近、対話で検索された結果や対話の流れに協調的に対応できるシステムを実現するため、CGIスクリプトによって、VoiceXMLを動的に生成する試みも行われている[17]。

VoiceXML文書の例を図4に示す[18]。VoiceXMLのルートタグは<vxml> (A)であり、内部に<form>(C)、<menu>、<meta>(B)といった要素を持つ。<form>には1単位の対話シナリオが、<menu>にはユーザの選択を伴う対話が記述される。また<meta>には文書に関する情報が記述される。

```
<vxml version="1.0" application="app-root.vxml"> ..... (A)
  <meta name="burger-order"> ..... (B)
  <form> ..... (C)
    <initial> ..... (D)
      <prompt>
        いらっしゃいませ、ご注文は何になさいますか
      </prompt>
    </initial>
    <field name="burger"> ..... (E)
      <prompt> ..... (F)
        バーガーを選んでください
      </prompt>
      <grammar> ..... (G)
        ハンバーガー | チーズバーガー
      </grammar>
    </field>
    <field name="number">
      <prompt>
        お幾つご注文ですか
      </prompt>
      <grammar>
        一つ | 二つ | 三つ
      </grammar>
    </field>
  <filled mode="all"> ..... (H)
      <value expr="burger">を<value expr="number">ですね ... (I)
      <if cond="number == 'ひとつ'"> ..... (J)
      <assign name="Qty" expr="1"/> ..... (K)
      <elseif cond="number == 'ふたつ'"/>
                        :
    </if>
    <if cond="burger == 'ハンバーガー'">
      <assign name="HBG" expr="HBG + Qty"/>
                        :
    </if>
      <goto next="#payment"/> ..... (L)
    </filled>
  </form>
</vxml>
```

図4　Voice XML 文書の記述例 [18]

<form>要素は幾つかの<field>要素(E)、<initial>要素(D)、<filled>要素(H)等から構成される。このうち<field>要素には1ターンのシステムとユーザのやり取りが記述される。<field>要素内の<prompt> (F)にはユーザの入力を促すためのプロンプトが記述され、ユーザは<grammar> (G)で指定される入力文法に従って発話する。<field>要素内に<filled>要素があれば、ユーザの入力を確定後にその内容が実行される。

一般的に、<form>要素の内部は上から順に実行される（対話が進行する）が、この例のように内部に<initial>要素を持つ場合には“混合主導対話”となり、内部の<field>が任意の順番で実行可能となる。このとき複数の<field>の同時入力も可能となる。例えばこのVoiceXMLでは次のような対話進行が考えられる。

対話例（1）
S：いらっしゃいませ、ご注文は何になさいますか？
U：ハンバーガーをください
S：お幾つご注文ですか
U：一つください
S：ハンバーガーを一つですね
　　：

対話例（2）
S：いらしゃいませ、ご注文は何になさいますか？
U：ハンバーガーを一つください
S：ハンバーガーを一つですね
　　：

上の対話例（1）では二つの<field>が逐次満たされ、（2）では同時に満たされている。<field>が満たされた後、このVoiceXMLでは<filled>要素 (H)が実行される。この例のように<filled>要素が<form>(I)の子要素として記述されるとき、<filled>の属性 mode によって<filledd>要素実行の契機となる<field>の数を規定できる。属性modeの値が“all”のとき、全ての<field>が満たされなければ<filled>は実行さ

れない。値が"any"のとき、一つ以上の<field>が満たされなければ<filled>が実行される。

<filled>要素内では、このVoiceXMLの場合、<if>要素(J)による入力内容に応じた条件分岐と、<assign>要素(K)による変数への代入、<goto>要素(L)による次対話への遷移等が実行される。

5. SALT

VoiceXMLは電話を対象にしているため、利用可能なモダリティが、音声およびDTMF（プッシュボタン）に限られている。今後、電話に限らず様々な端末を用いたWebアクセスが行われるようになると、マルチモーダル対話を記述する必要が生ずるが、現在のVoiceXMLでは限界がある。このため、マルチモーダルな要素を取り込めるSALTを提唱するForum（Microsoft、Cisco、Intel、Philips、SpeechWorks、Comverse が参加）が活動を開始している[19]。SALTでは、入力として音声認識、キーボード、キーパッド、ペン、マウスなどを、出力として、スクリーン上のテキスト表示、合成音声、グラフィックス、動画などを対象としており、これらは、互いに独立にあるいは同時に用いられる。2.5/3G 携帯のキラーアプリとして、「いつでも」「どこでも」「どんなデバイスでも」「ユーザが好む方法で」情報が得られるようにすることを目指している。モバイルゲームなどエンタテインメント性も重視されている。SALTは、既存のWeb標準、すなわちHTML、XHTML、XMLなどの拡張として設計されており、ロイヤリティーフリー、プラットフォーム独立を目指している。2002年の第2四半期にはW3Cに仕様を提案する予定になっている。

同様に、国内でも、マルチモーダル対話記述言語として、XISL (eXtensive Interaction Sheet Language) [18] が提案されている。

6. むすび

音声認識による音声文字変換には、極めて多様な応用が期待できるが、まだ解決しなければならない課題も多い。主たる課題は、話し言葉特有の言語的、音響的変動に対するロバストなモデルと適応アルゴリズムの構築である。その基礎となる話し言葉の大規模コーパスの構築も極めて重要である。音声対話システムでも、音声認識部は、音声文字変換と同様に、大量のコーパスから作成した統計的言語モデルを用いる方法が主流になりつつある。この意味で、音声文字変換技術は音声認識の基本と言える。音声対話システムでは、さらに、コンピュータシステムとの対話のやりとりによって、音声認識誤りを修正あるいは回避しながら、種々の情報が効率よく得られるように、対話の流れ（制御）を設計する必要がある。ユーザにとって真に使いやすいコンピュータとの音声対話システム、あるいはマルチモーダル対話システムの実現法は、まだ確立していない。VoiceXMLやSALTのような標準化対話記述言語の普及によって、システムが容易に作成できるようになれば、急速にノウハウが蓄積され、対話システムの構築技術の進歩と普及が加速するものと期待される。

文　献

[1] 古井貞熙：話し言葉処理のメカニズム － 音声認識から会話意図理解へ --、映像情報メディア学会誌、54、6、pp. 765-768 (2000)

[2] 古井貞熙：音声情報処理、森北出版 (1998)

[3] 古井貞熙：音声トランスクリプションのこれまでと今後の展望、電子情報通信学会論文誌、J83-D-II、11、pp. 2059-2066 (2000)

[4] 松岡達雄、他、"新聞記事データベースを用いた大語い連続音声認識"、電子情報通信学会論文誌、J79-D-II、12、pp. 2125-2131 (1996)

[5] 河原達也、他、"連続音声認識コンソーシアム2000年版ソフトウェアの概要と評価"、音声言語情報処理研資、38-6 (2001)

[6] K. Ohtsuki et al., "Improvements in Japanese broadcast news transcription," Proc. DARPA Broadcast News Workshop, pp. 231-236 (1999)

[7] 今井 亨、他、"逐次2パスデコーダを用いたニュース音声認識システム"、電子情報通信学会技報、SP99-129 (1999)

[8] 西村雅史、"日本語ディクテーションシス

テムの現状と今後の課題”、電子情報通信学会技報、SP99-94 (1999)
[9] 古井貞熙、他、“科学技術振興調整費開放的融合研究推進制度 - 大規模コーパスに基づく『話し言葉工学』の構築 -”、日本音響学会誌、56、11、pp. 752-755 (2000)
[10] 堀　智織、他、“単語抽出による音声要約文生成法とその評価”、電子情報通信学会論文誌、J85-D-II、2、pp. 200-209 (2002)
[11] 河原達也、他、“会話スタイル依存の言語モデルを用いたキーフレーズの検出・検証”、電子情報通信学会技報、SP97-78 (1997)
[12] 許 濱、“音声ポータルビジネスモデルについて”、音声言語情報処理研資、38-9 (2001)
[13] Microsoft SAPI: http://www.microsoft.com/iit/
[14] Java Speech API: http://www.java.sun.com/products/java-media/speech/index/html
[15] 鯨井俊宏、他、“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 -”、音声言語情報処理研資、33-12 (2000)
[16] http://www.voicexml.org/
[17] 安達史博、他、“VoiceXML の動的生成に基づく自然言語音声対話システム”、音声言語情報処理研資、40-23 (2002)
[18] 桂田浩一、他、“音声対話記述言語 VoiceXML と MMI 記述言語 XISL の比較”、音声言語情報処理研資、38-8 (2001)
[19] http://www.saltforum.org/