音声言語情報処理 33－2
（2000．10．27）

対話システムにおける音声認識

甲斐 充彦† 伊藤 克亘††

† 静岡大学工学部
〒 432-8561 浜松市城北 3-5-1
†† 電子技術総合研究所
〒 305-8568 茨城県つくば市梅園 1-1-4

あらまし

大語彙ディクテーションタスクを中心として，音声認識システムは大量の音声・言語コーパスを用いた統計的なアプローチにより大きな進展が得られている．また，近年の研究プロジェクトにより，音声認識とその応用システムに関する研究のための共通ソフトウェア基盤も少しづつ整備されつつある．一方，音声対話システムなどの対話音声を扱う領域では，ようやくコーパスが整備されつつあるが，音声対話システム特有の問題など応用システムを視野に入れた様々な研究が進められている．本稿では，音声認識とその対話システム応用に関連する対話音声認識技術の現状や課題について概要を述べ，最後に，本年度から始まるプロジェクトの一部としての対話音声認識の開発計画について触れる．

Speech Recognition in Spoken Dialogue Systems

KAI Atsuhiko† ITOU Katunobu††

† Faculty of Engineering, Shizuoka University
3-5-1, Johoku, Hamamatsu, Shizuoka, 432-8561, Japan
E-mail: kai@sys.eng.shizuoka.ac.jp
†† Electrotechnical Laboratory
1-1-4, Umezono, Tsukuba, Ibaraki, 305-8568, Japan
E-mail: kito@etl.go.jp

Abstract Speech recognition systems have made a rapid advances in the past decade supported by statistical approaches based on large speech and text corpora. Also the ongoing research project for the development of a Japanese Dictation Toolkit is providing software components for sharing of the speech research tools. On the other hand, in the field of conversational speech technologies, although spontaneous speech corpora are planed to be collected on a large scale, researches for the speech recognition technologies which are relevant to spoken language systems has made progress. This paper describes the overview of speech recognition techniques for conversational speech in relation to the application of speech recognition for dialogue systems.

1 はじめに

最近では，特に統計的なアプローチと大量の音声・言語コーパスの利用によって音声認識システムが大きな進歩を遂げ，音声対話システムやマルチモーダルユーザインタフェース (MUI) など，音声言語処理の応用研究に目が向けられてきている．これらは，音声認識システムと効果的に融合することによって，新たな音声言語処理応用システムの可能性を示すことが期待されている．

音声認識応用システムを構築するためには，全体を統合するアーキテクチャの設計から，辞書・文法の作成，個々のコンポーネントとのインタフェースの設計等まで，システムの開発に多大な労力と専門的な知識を必要とするのが現状である．そのような背景から，最近では，情報処理振興事業協会 (IPA) のプロジェクト「日本語ディクテーション基本ソフトウェアの開発」において基本ソフトウェアが公開され，ディクテーションに関連した音声言語処理技術の研究基盤として用いられている．一方，今年度からは，IPA「擬人化音声対話エージェント基本ソフトウェアの開発」(擬人化エージェント) プロジェクトが開始しており，そこでは音声認識を含む各モジュールを連携させる統合・制御モジュールを中心に，擬人化対話エージェントを研究・利用するための共通ソフトウェア基盤を構築する計画になっている．

本稿では，このような現状を踏まえ，まず音声対話システムにおける対話音声認識の現状や課題について述べる．対話音声認識に関する話題は多岐に渡り，これまでにも幾つかの解説論文等 [1, 4] があるが，ここでは音声対話システムの一コンポーネントとしての観点から一部の主な要素技術等に焦点を当てて紹介する．なお，発話の解釈に関することは本稿で扱っておらず，参考文献 [9] 等を参照されたい．最後に，「擬人化エージェント」プロジェクトでの対話音声認識に関する開発計画の概要について触れる．

2 対話音声の特徴

人間同士の対話音声のコーパスを基に，様々な話し言葉 (音声言語) 特有の現象が観測されている [2]．話し言葉に特有な言語現象として，つなぎ語 (間投詞)，助詞落ち，倒置の多用，言い間違い，言い直し，言い淀みなどがある [1]．対話音声のコーパス (10 対話 1052 文) の分析によれば，つなぎ語に関しては多くの種類が観測されるが，上位 5 種類で全体の 93%を占め，特定の間投詞の頻度が非常に高い [7]．また，1 文当りの間投詞の出現回数は 1.2 回，言い直しは 0.15 回，助詞落ちは 0.09 回，倒置は 0.02 回観測されている．

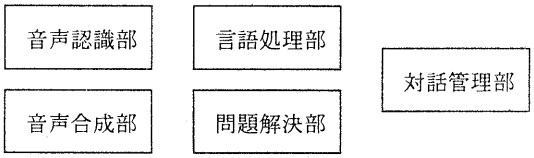

図 1: 一般的な音声対話システムの構成要素

書き言葉と比較した特徴として，縮約 (音便や音韻レベルでの省略・変形) や，文単位の認定の難しさ，発話の時間的な特徴の扱い，対話相手による表現の違い (敬語や言い回し)，発話のターンの管理，等も挙げられる [6]．また，典型的な発話として，「コソアド」の指示語が多く使われ，文型は比較的単純で断片的な句の連続が多い，等の特徴がある [1]．

韻律情報は音声言語特有のもので，言い淀みや非文法的な内容を含む発話を多く含む話し言葉の認識と理解において，様々な手掛かりを与えるものと考えられる．しかし，韻律的なタグを持った話し言葉のコーパスは十分整備されておらず，効果的な応用例はまだ少ない．現在，組織的に話し言葉コーパスの構築が進められており，韻律研究の環境が少しづつ整いつつある [3]．

ここに挙げた言語現象は，おもに人対人の対話コーパスから観測されているが，人対機械の音声言語インタフェースにおいての分析も重要である．そこで，実際の対話システムと被験者との間の対話データや，人間が機械による応答を肩代わり (模擬) する Wizard of OZ (WOZ) 方式による被験者との対話データの収集・分析も行われている [2, 5]．

3 対話音声認識の要素技術

3.1 対話音声認識の課題

対話音声認識は，対話システムの他の部分とも密接に関わっており，各処理レベルを分離して考えることは難しい (図 1)．しかし，対話音声認識という側面から多くの解決すべき課題が挙げられ，それらの多くは今日でも重要な課題といえる [1, 4]．ここでは，対話システム応用に関する対話音声認識の課題に焦点を当てることにする．

対話音声認識においては，音響モデル，言語モデル，探索方式といった一般的な音声認識での要素技術の他，認識結果の信頼度の評価，割込み (Barge-in)・ポーズの扱い，実時間処理，漸次的処理などの扱いが重要となる．

一般的な日常会話や講演等における自然な発話 (spontaneous speech) では，読み上げ音声と比較して発声のゆらぎや不明瞭さが観測される．従って，

従来の読み上げ音声に基づいて学習された音響モデルでは不十分であり，話し言葉のための音響モデル構築は重要な課題となっている [10, 2, 3]．その一方で，情報検索を行うような制限された対話タスクでは実用レベルの対話システムも多く試作され，話者・環境適応化の併用等によって，タスクが制限された小規模な対話システムの構築は現実的なレベルになっている．しかし，対話音声の認識では一般に様々な実環境での利用が想定されるため，環境の違いに対するロバスト性やユーザインタフェースの向上が重要な課題であることが指摘されている [2]．音響モデルはこれらの問題に関わる重要な要素であるが，最近の話し言葉コーパスの整備によって今後多くの成果が期待される [3]．

次節以降では，言語モデルと探索方式，認識結果の信頼度評価，割込み・ポーズの扱いの現状や課題について述べる．

3.2 言語モデルと探索方式

一般に，話し言葉（音声言語）の言語解析では次のような問題がある．(1) 音声や言語（意味，文脈など）のレベルにおいての曖昧さ（不確実さ）が増す．(2) 話し言葉（音声言語）にはいわゆる文法的に不適格 (ill-formed) な発話があり，発話の解析に対する頑健性が要求される [9]．従って，話し言葉の音声認識においては，何らかのトップダウンの言語知識を利用することが不可欠であり，言語情報による曖昧さの解決が必要となる．

連続音声認識 (CSR) における言語モデル及び探索方式では，読み上げ音声／テキストのコーパスの充実に伴って，統計的言語モデルをベースとした方法が主流となっている [38]．単語 N-gram は，大量のデータを基に言語レベルの曖昧さを効率よく吸収できるため最も良く用いられる．話し言葉であっても内容によって程度の違いが予想され，例えば丁寧な講演調の音声では，十分な量の話し言葉の音声・言語コーパスを用いれば，このアプローチは大きな効果が得られている [10]．同様に，対話音声であっても，情報検索のタスクのように比較的発話の変動が制限され得る場合にも効果が確かめられている [11, 12]．

対話音声の認識においては，スポッティング方式 [16, 14] や構文制御による探索方式も用いられる．前者は，キーワード以外の言語情報を用いない場合には十分な精度が得られず，一般に単語連鎖などの簡易的な言語モデル（フィラーモデル）を用いる．また，言語解析は，結果として得られる単語ラティスから曖昧さを考慮した頑健な言語解析によって行う必要がある [9]．後者の場合は，一般に正規文法または文脈自由文法の言語的な制約により音声認識を行う．探索方法としては，文脈自由文法で予測される探索空間を動的に展開して 1 パスで探索を行う方法と [5, 15, 17]，探索空間の増大に対処するために探索処理を 2 段階に分け，前段で単語連鎖 (bigram)，後段でより高次の言語モデルで再探索を行う方法等がある．前者では効率を考えると近似的に複数候補（N-best 仮説）を求めることになるが，後者では A* 探索アルゴリズムにより効率性を維持してより正確に求めることができる [13]．しかし，発話の終了を検知してから 2 段目の処理が行われるため逐次的な出力ができない問題がある．この問題は，局所的な言語制約のみを用いる N-gram 言語モデルをベースとした探索法では，2 段目を逐次的に行う改良法が報告されている [19]．

文脈自由文法などの形式言語による言語モデル記述は，簡単な対話アプリケーションでは作成及び変更が容易という利点がある．しかし，統計的な言語モデルと比べて比較的カバレージ (coverage) が低く，パープレキシティが高くなる問題があり，対話音声認識の精度に大きく影響する．実際に，タスクに依存した十分な量のコーパスを利用できれば，人手で記述した文法より有効性が示されている [12, 11]．しかし，タスクに依存したコーパスを大量に集めることや，言語モデルのポータビリティの問題等から，形式文法を統計的な言語モデルと併用する方法も提案されており，少量のコーパスしか利用できない場合にも有効性が示されている [12, 20]．この他，統計的なモデルで，構文・意味的な情報を統合的に扱う試みもある [8, 10]．

3.3 認識結果の信頼性評価

一般的な音声認識の枠組みでは，未知語，不要語やシステムの想定外の発話等による誤認識が避けられない．そこで，認識結果の不正解やシステムの想定外の発話であることを判定する「棄却 (rejection)」の機能が要求される．この棄却の判断基準として認識結果の信頼度 (confidence measure) を推定できれば，対話システムでは信頼度に応じて誤り部分の問い返しや応答，対話戦略を変えるなど，柔軟な対話制御が可能になる [22]．

一般的な連続音声認識システムは，入力 $\boldsymbol{x}$ に対して次式で表される事後確率 $P(\boldsymbol{w}|\boldsymbol{x})$ を最大化する $\hat{\boldsymbol{w}} = \max_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{x})$（例えば単語や文）を求める問題として定式化される．

$$P(\boldsymbol{w}|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\boldsymbol{w})P(\boldsymbol{w})}{P(\boldsymbol{x})} \tag{1}$$

分母は仮説の相対的な比較に無関係なため無視され，音声認識では分子のみが評価値として用いられる．この評価値は，信頼度や棄却の基準としては不十分であることから，以下のような信頼度の推定法が提案されている．

1. $P(\boldsymbol{x})$ の推定による方法
   音響モデルの HMM の全状態 s の出力分布関数から，$P(\boldsymbol{x}) = \sum_s P(\boldsymbol{x}|s)P(s)$ として求め，信頼度を $P(\boldsymbol{x}|\hat{\boldsymbol{w}})/P(\boldsymbol{x})$ として求める．
2. 音素／音節連鎖モデルによる方法
   任意の発話（単語や文）は音素／音節の連鎖で表されると仮定し，音素／音節系列の認識結果 $\hat{\boldsymbol{p}}$ としての尤度 $P(\boldsymbol{x}|\hat{\boldsymbol{p}})$ を参照スコアとして，尤度比 $P(\boldsymbol{x}|\hat{\boldsymbol{w}})/P(\boldsymbol{x}|\hat{\boldsymbol{p}})$ により信頼度を求める [24, 25, 26]．前記の方法と対比させると，次式による近似と見ることができる．
   $$P(\boldsymbol{x}) = \sum_s P(\boldsymbol{x}|\boldsymbol{p})P(\boldsymbol{p}) \sim max_{\boldsymbol{p}} P(\boldsymbol{x}|\boldsymbol{p}) \quad (2)$$
   この方法は，発話中の未知語部分の検出法としても提案されている [24, 27, 28]．
3. Garbage model／対立モデルによる方法
   語彙（キーワード）以外の任意の音声で学習された音響モデル（Garbage model）を用意し，これによって参照スコアを求める [23]．Garbage model は，一つの HMM でモデル化することで少ない計算量で参照スコアが求まるが，語彙が多い対話音声認識では十分な精度が期待できない．
   改良法として，サブワード単位での対立モデル（anti-model）を用い，認識結果のサブワード毎の尤度比の平均や最小値等を信頼度とする方法がある [29]．

上記の方法は基本的に音響モデルの情報のみに基づくものであるが，更に信頼度の精度を改善するために，その他の様々な情報やモデルを用いる信頼度の推定法が提案されている．それらの一部を紹介する．

1. 言語／辞書的な情報の利用
   スポッティング方式の音声認識において，キーワード外の十分な言語／辞書的な情報によりフィラーモデル（filler model）を与える [31, 21]．
2. 複数認識候補（N-best 仮説）の利用
   大語彙ディクテーションシステムの複数の認識器の出力または単一の認識器による複数候補（N-best 仮説）を利用して，同一単語の出現率や競合仮説数等によって認識単位毎の信頼度を推定する [30]．
3. 複数の手掛かりとなる特徴量の併用
   認識器から得られる認識結果に付随した多くの特徴を用い，線形判別法またはニューラルネット識別器の出力により信頼性を推定する [32]．

3.4 割込み（Barge-in），ポーズの扱い

対話システムでは，機械からの応答の途中でユーザが割り込んで話すような状況（Barge-in）が起こることがある．音声認識では，このようなユーザの発話を許すためには，機械からの音声合成出力をキャンセルするか，その影響を取り除く必要がある．また，対話システムにおいて柔軟な応答，理解，対話制御を行うには，漸次的な音声認識と理解が必要となる．このような研究はまだ始まったばかりであるが，ポーズやその他の音響的な手掛かりを用いて，逐次的なあいづちや細かな発話単位で処理するシステムが試作されている [33]．

ポーズの扱いは，音声認識及び言語処理における発話の単位の扱いに依存する．対話音声では前述のように断片的な発話もあることから，ポーズ単位での文法を用いる方法が提案されている [34]．また，文献 [35] では，逐次的な音声認識・理解のため，言語処理モジュールに対して時間や信頼度を含む認識途中結果を渡し，必要に応じて言語処理モジュールが理解内容を再構成できる仕組みを提案している．ポーズ単位の定義は，このような処理方式だけでなく，実時間性を考慮するかどうかにも関わる．しかし，現在の一般的なネットワーク技術は実時間性を保証するものではないため，複数の計算機でより高度で複雑な対話システム構築を可能にするためにも，実時間での分散処理を可能にする技術基盤の確立が望まれる．

4 対話システムにおける音声認識応用

4.1 標準化動向

現在，対話システム等で音声認識応用システムを構築するには，辞書・文法の作成，インタフェースの設計等，システムの設計・開発に多大な労力と専門的な知識を必要とするのが現状である．一方，最近では，商用の音声認識エンジンのベンダーからアプリケーション開発者向けに API(Application Programming Interface) を始めとして実用アプリケーション開発を支援する環境を整備している．個々の音声認識エンジン向けに提供されるもの以外に，プラットフォーム間の互換性を考慮して，サン・マイク

ロシステムズ社を中心に，Javaアプリケーションによる音声認識や音声合成システムとのインタフェースの仕様 (Java Speech API:JSAPI) や，音声認識のための文法記述の仕様 (Java Speech Grammar Format:JSGF) などが提案されている [37].

また，Motorolaを始めとする数社によって創設されたVoiceXMLフォーラムでは，Web技術の応用として音声による対話型アプリケーションを広く提供できるようにするため，拡張マークアップ言語（XML）による対話システム記述言語の標準化仕様 (VoiceXML) を提案している．この仕様では，3つのコンポーネントからなるアーキテクチャのモデルが示され，音声認識エンジンの実装は主にImplementation platformに関する部分として独立性を持たせることが可能になっている．

VoiceXMLの仕様は，複数の対話の状態，ユーザの入力，応答，対話状態の遷移等を記述するため，一般的に扱える対話は，想定される対話スクリプトを記述できるようなものに限られる．しかし，WWWにおけるCGI(Common Gateway Interface)の仕組みと同様に，外部（Document Server）でデータベースの検索結果や特別な判断によって動的に記述を与えることも可能であり，ユーザ主導，システム主導，両者混合の対話戦略による簡単な対話システム実装のプラットフォームとして利用できる．但し，対話システム開発・研究のプラットフォームとして種々の要求を完全に満たすものではない．例えば，発話のターン切替えやタイミングを細かに実時間で制御するような記述は不可能である．

一方，文法記述の仕様であるJSGFは，音声認識器がユーザ発話の言語的な制約として利用する文法を記述する仕様を定義したものである．文脈自由文法の書き換え規則のように，任意の規則名（非終端記号）または単語名（終端記号）の列として規則の集合を定義する．但し，仕様では再帰規則として右再帰のみが許される．文法の一部の記述例を以下に示す．

```
<command> = <action> | (<action> and <command>);
<action> = stop | start | pause | resume | finish;

<command> = please (open {OPEN} | close {CLOSE})
            the file;
```

ここで，“<*name*>”は*name*が規則名（非終端記号）を表しており，“and”や“stop”等は音声認識器の辞書で参照されるエントリ名（終端記号）を表している．また，括弧はグループ化，“|”は代替規則の定義の区切り，“{}”は認識結果で単語（集合）毎に付加されるタグを表している．このJSGFの仕様で注目すべき特徴として，文法をサブセットに分けて分割・分散管理でき，それらを固有のラベルでインポート・参照できるようになっている．

4.2 「擬人化エージェント」プロジェクト

今年度から開始したIPAの「擬人化音声対話エージェント基本ソフトウェアの開発」（擬人化エージェント）プロジェクトでは，その一部として対話音声認識基本ソフトウェア開発が計画されている．ここでは，その計画の概要を述べる．

現在，IPAのプロジェクト「日本語ディクテーション基本ソフトウェアの開発」が今年度までの計画で進められており，既に開発されたソフトウェアの成果が一般に公開されている [38]．一方，「擬人化エージェント」ではこの成果を基に，対話音声認識基本ソフトウェアを開発する．本稿で述べてきた対話音声認識の課題を考慮し，以下のような課題に取り組む．

- 文法に基づく音声認識システムの開発（文法・辞書，デコーダ）
- 音声認識結果の棄却，不要語への対応（信頼度評価）
- 認識処理の動的制御（ポーズへの対応）
- 制御可能なコマンド（通信）体系の設計・実装

この中で実現される機能とアプリケーションとの基本的なインタフェース（必要となる情報の入出力の内容）は，前節のAPIやVoiceXMLなどの仕様が参考となる．例えば，ポーズに関する制御のパラメータとして，ユーザの音声入力待ち，音声入力による認識候補確定時，音声入力による認識候補未確定時などの状態毎に最大許容ポーズ長を指定する．また，認識結果の信頼度を利用したり，文法の一部に優先度を与えたり，文法や辞書の一部を動的にアクティブ／非アクティブにする，等の仕様がある．

一方，JSGFの仕様において日本語での単語の読みの問題が考慮されていない点，VoiceXMLにおいて逐次的な処理が考慮されていない点等は，今後検討すべき課題である．

5 おわりに

音声対話システムに関連する研究が盛んになりつつある現状において，本稿では，対話システムへの応用のための対話音声認識技術の現状と課題の概要について述べた．今後，「擬人化エージェント」プロジェクトの一部として，これらの対話音声認識技術に関して更なる検討を行いながら，共通基盤となるシステムの開発を進めていく計画である．なお，

音響モデルや評価の問題などについては触れられなかったが，近年のコーパスやソフトウェアなど研究の共通基盤の整備に伴って，今後の研究成果が期待される．

参考文献

[1] 小特集："音声によるコンピュータとの対話を目指して，" 音響学会誌, Vol.50, No.7, pp.556–580 (1994).

[2] 小特集："音声対話システムの実力と課題，" 音響学会誌, Vol.54, No.11, pp.783–822 (1998).

[3] 板橋, 藤崎, 山本, 板倉, 古井, 広瀬, 田中, 市川, "パネル討論：音声言語関連大型プロジェクトの現状と課題，" 情報処理学会研究会資料, SLP-29-38 (1999.12).

[4] 中川, 岡田, 島津, 堂下, 田窪, 河原, 松本, 新田, 嵯峨山, "音声言語情報処理の現状と研究課題，" 情報処理, Vol.36, No.11, pp.1011–1053 (1995).

[5] K. Itou, S. Hayamizu, K. Tanaka, H. Tanaka, "System design, data collection and evaluation of a speech dialogue system," *IEICE Trans. Inf. and Syst.*, Vol.E-76-D, No.1, pp.121–127 (1993).

[6] 田中穂積監修, 第 10 章："音声対話システム，" 自然言語処理–基礎と応用–，" 電子情報通信学会 (1994).

[7] 中川, 小林, "自然な音声対話における間投詞・ポーズ・言い直しの出現パターンと音響的性質，" 音響学会誌, Vol.51, No.3, pp.202–210 (1995).

[8] R. Pieraccini, E. Tzoukermann, Z. Gorelov, J.-L. Gauvain, E. Levin, C.H. Lee, and J.G. Wilpon, "A speech understanding system based on statistical representation of semantics," *Proc. ICASSP*, pp.I-193–196 (1992).

[9] 河原達也, 松本裕治, "音声言語処理における頑健性，" 情報処理, Vol.36, No.11, pp.1027–1032 (1995).

[10] 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，" 情報処理学会研究会資料, SLP-29-2 (1999.12).

[11] 小暮, 伊藤, 廣瀬, 甲斐, 中川, "CFG/bigram を使用した対話音声認識における意味理解の比較検討，" 情報処理学会, 第 57 回全国大会 2 分冊, pp.239–240 (1998.10).

[12] 中川聖一, 大谷耕嗣, "Ｂｉｇｒａｍの使用による話し言葉用確率文脈自由文法の自動学習，" 情報処理学会論文誌, Vol.39, No.3, pp.575–584 (1998.3).

[13] 河原達也, "探索アルゴリズム— A* 探索を中心に，" 信学技法, SP92-36 (1992.6).

[14] 河原達也 他, "ヒューリスティックな言語モデルを用いた会話音声中の単語スポッティング，" 電子情報通信学会, Vol.J78-D-II, No.7, pp.1013–1020 (1995).

[15] 中川 聖一, 甲斐 充彦, "文脈自由文法制御による One Pass 型 HMM 連続音声認識法，" 電子情報通信学会論文誌, Vol.J76-D-II, No.7, pp.1337–1345 (1993).

[16] 坪井, 橋本, 竹林, "キーワードスポッティングに基づく連続音声理解，" 信学技報, SP91-95 (1991).

[17] X. Huang, et al., "Microsoft Windows highly intelligent speech recognizer: Whisper," *Proc. ICASSP*, pp.93–96 (1995).

[18] R. Schwartz and Y.L. Chow, "The N-best algorithm: An efficient and exact procedure for finding the N most likely sentence hypotheses," *Proc. ICASSP*, pp.81–84 (1990).

[19] 今井, 小林, 安藤, "認識結果早期確定のための逐次 2 パスデコーダ，" 日本音響学会講論集, 2-1-4 (1999.9).

[20] Ye-Yi Wang, et al., "A unified context-free grammar and n-gram model for spoken language processing," *Proc. ICASSP*, pp.1639–1642 (2000).

[21] 河原達也, 石塚健太郎, 堂下修司, "発話検証に基づく音声操作プロジェクタとそれによる講演の自動ハイパーテキスト化，" 情報処理学会論文誌, Vol.40, No.4, pp.1491-1498 (1999).

[22] 新美, 小林, "音声認識の信頼性に基づいた対話制御方式，" 信学技報, SP96-30 (1996.6).

[23] Wilpon J. G., Rabiner L. R., Lee Co-H. and Goldman E. R., "Automatic recognition of keywords in unconstrained speech using hidden Markov models," *IEEE Trans. Acoust., Speech & Signal Process.*, Vol.38, No.11, pp.1870–1878 (1990).

[24] Asadi A., Schwartz R. and Makhoul J., "Automatic detection of new words in a large vocabulary continuous speech recognition system," *Proc. ICASSP*, pp.125–128 (1990).

[25] 渡辺隆夫, 塚田 聡, "音節認識を用いたゆう度補正による未知発話のリジェクション，" 信学論, Vol.J75-D-II, No.12, pp.2002–2009 (1992).

[26] 北, 江原, 森元, "連続音声認識における未知語処理，" 日本音響学会講論集, 3-5-3 (1991.3).

[27] 伊藤克亘, 速水 悟, 田中穂積, "連続音声認識における未知語の扱い，" 信学技報, SP91-96 (1991.12).

[28] 甲斐 充彦, 中川 聖一, "冗長語・言い直し等を含む発話のための未知語処理を用いた音声認識システムの比較評価，" 電子情報通信学会論文誌, Vol.J80-D-II, No.10, pp.2615–2625 (1997).

[29] R. A. Sukkar and C. -H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol.4, No.6, pp.420–429 (1996).

[30] D. Willett, A. Worm, C. Neukirchen, and G. Rigoll, "Confidence measures for HMM-based speech recognition," *Proc. ICSLP*, pp.3241–3244 (1998).

[31] M.Weintraub., "Keyword-spotting using SRI's DECIPHER large-vocabulary speech recognition system," *Proc. ICASSP*, pp.463–466 (1993).

[32] T. Schaaf, and T. Kemp, "Confidence measures for spontaneous speech recognition," *Proc. ICASSP*, pp.875–878 (1997).

[33] 平沢, 川端, "音声対話システム Noddy —ユーザ発話途中でのうなずき・相槌生成—"，情報処理学会研究会資料, SLP20-9 (1998).

[34] 竹沢, 田代, 森元, "自然発話の言語現象と音声認識用日本語文法，" 情報処理学会研究会資料, SLP-6-5 (1995.5).

[35] 川端, 宮崎, 平沢, "逐次的音声認識・理解のための ISTAR アーキテクチャ，" 音講論集, 1-1-15 (1998.9).

[36] http://www.voicexml.org/ (Voice XML version 1.0)

[37] http://java.sun.com/products/java-media/speech/

[38] 鹿野, "日本語ディクテーション基本ソフトウェア，" 音講論集, 2-8-1 (2000.3).