実時間音声対話システムTOSBURGの開発 (1) システム構成

6N−5

竹林洋一* 坪井宏之** 金沢博史* 貞本洋一* 山下泰樹**
瀬戸重宣* 永田仁史* 新居孝章** 橋本秀樹*** 新地秀昭***
*(株)東芝 総合研究所 **(株)東芝 関西研究所 ***東芝ソフトウェアエンジニアリング(株)

1.はじめに

近年、音声メディアの自然でユーザーフレンドリーな特長を活かした音声対話の研究が活発になってきている[1,2]。音声対話システムは、テキスト入力の様な文字面の入力とは異なり、ユーザと計算機の間の意図、意思の伝達が目的であるため、単語音声や文音声の認識率よりも発話の意味内容の理解が重要となる。また、ヒューマンインタフェースの観点からも、音声対話システムは発話の制約を極力少なくすることが望ましく、さらに、システムを評価するためにも自由発話の実時間の音声理解処理が必要になる。これに対して我々は実際の応用場面を考慮し、ロバスト性を重視した不特定ユーザ向の実時間音声対話システムTOSBURG(Task-Oriented dialogue System Based on speech Understanding and Response Generation)を試作したので報告する。

2.音声対話システムの課題とTOSBURGのコンセプト

2.1.音声対話システムの課題

従来、大語彙/不特定連続音声認識が内外の研究機関で研究されているが、これらのシステムを音声対話システムとして用いると仮定すると、実際の場面で重要なリアルタイム処理や耐雑音性能は必ずしも十分ではなく、さらには発話の文型や形式に関する制約が強いという問題がある。

また、自由な発話(spontaneous speech)を対象とするシステムでは、不要語、言い直し、省略、ポーズ、環境の雑音、対象外の単語などに対処する必要があるが、このような現象を文法として完全に記述することは困難であり、なんらかの方策が必要となる。さらに、認識処理により生じる曖昧性や誤認識などに対しても対話の制御により対処する必要がある。

これらの課題をまとめると以下の4つに分類できる

(1) 発話環境へのロバスト性(耐雑音性)向上[3]
(2) 多様な話し言葉への対処[4]
(3) リアルタイム処理の実現[5]
(4) 誤認識・曖昧性の対話処理での吸収

以下に、システム構成についての説明し、本システムにおける上記課題に対するアプローチについて述べる。

2.2.TOSBURGのコンセプト

現状のマルチメディアの応用の大部分はテキスト、画像、音声などの単なる入出力にすぎず、メディアの理解や生成機能を用いたフレンドリーなヒューマンタフェースの実現には至っていない。これに対してTOSBURGでは簡単なハンバーガショップでの注文システムを想定し、不特定のユーザに対して制約を設けずに、通常の話し言葉で計算機と自然に対話できるシステムの構築を目指している。この様なシステム実現のために、誤りや曖昧性を含む音声理解の他に、対話用の音声合成、対話制御のための知識処理、視覚メディアの利用など、種々の知識やメディアの融合が必須である。また、この新しい人工メディアを用いて評価実験を行い、音声対話システムのユーザの反応を調べ、使い勝手の良いユーザ主体のマルチメディア対話システムの研究が可能となる。

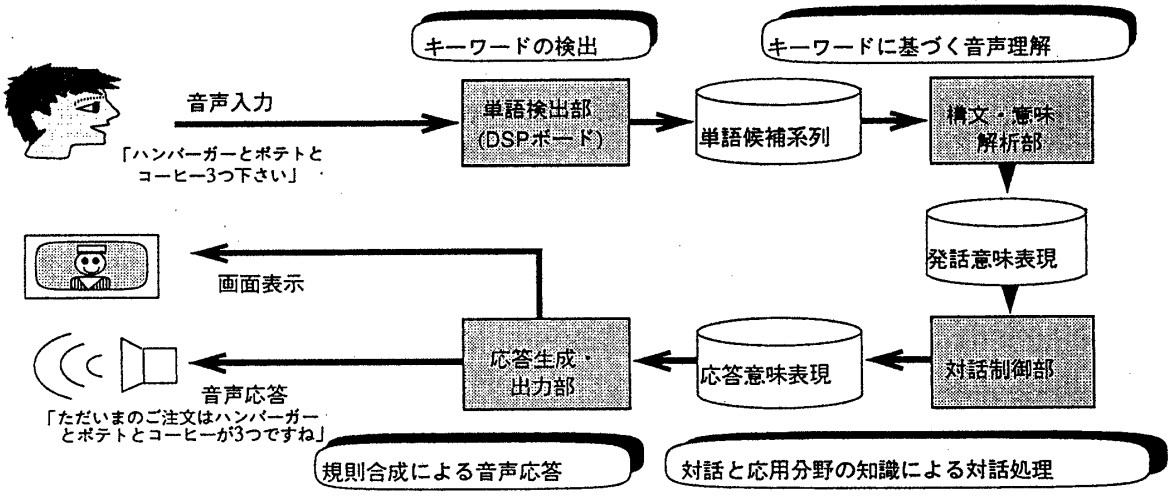

図1. 音声対話システムの構成

Development of Real-Time Speech Dialogue System TOSBURG (1) System Configuration
Yoichi TAKEBAYASHI*,Hiroyuki TSUBOI**,Hiroshi KANAZAWA*,Yoichi SADAMOTO*,Yasuki YAMASHITA**
Shigenobu SETO*,Yoshifumi NAGATA*,Takaaki NII**,Hideki HASHIMOTO***,Hideaki SHINCHI***
* Toshiba R&D Center, ** Toshiba Kansai Research Lab., *** Toshiba Software engineering

3.システム構成

3.1.全体構成

TOSBURGは、図1に示すように、キーワード検出、キーワードに基づく音声理解、応用分野の知識を利用した対話処理、音声を含むマルチメディア応答により、音声対話を行う。上述した各要素技術の研究は個別に行われる場合が多いが、我々はこれらを統合し、不特定ユーザ用の簡単なタスクの音声対話システムを、2台のワークステーションとDSPボードを用いて、実時間システムとして試作した。ユーザ主導型のフレンドリーでロバストなシステムであり、音声メディアの他に、人物を検知するマットや、応答メディアとしてテキスト表示、注文品と店員の表情の表示も利用した。

3.2.単語検出部

現状のマルチメディアの応用の大部分はテキスト、画像、音声などの単なる入出力にすぎず、メディアの理解や生成機能を用いたフレンドリーなヒューマンタフェースの実現には至っていない。これに対してTOSBURGでは簡単なハンバーガショップでの注文システムを想定し、不特定のユーザに対して制約を設けずに、通常の話し言葉で計算機と自然に対話できるシステムの構築を目指している。この様なシステム実現のために、誤りや曖昧性を含む音声理解の他に、対話用の音声合成、対話制御のための知識処理、視覚メディアの利用など、種々の知識やメディアの融合が必須である。また、この新しい人工メディアを用いて評価実験を行い、音声対話システムのユーザの反応を調べ、使い勝手の良いユーザ主体のマルチメディア対話システムの研究が可能となる。

3.3.構文・意味解析部

不特定ユーザが対話する際の制約を取除き、自由発話に対応するための実時間処理に適したキーワードに基づく音声の内容を理解する方式を開発した。タスクを小規模なハンバーガ店の注文に限定し、キーワード以外の発話を無視することにより、多様な話し言葉の理解が可能となった。従来の音声ー文変換の枠組みに止どまらず、対話中の省略表現や不要語にも対応できる意図理解を目指し、不明な点や曖昧な点は対話処理で補うこととした。図2は、不要語が発声中に、挿入された場合の、キーワードに基づく解析結果を示す。

3.5.対話制御部

従来の計算機主導型の対話処理ではなく、不特定ユーザが自由に対話できるユーザ主導型の対話システムを目指した。ユーザと計算機との対話モデルは、音声認識の誤りや、曖昧性への対処とマルチメディアの利用を重視し、図3のように、言語行為論の形式で発話意味表現と応答意味表現のアクトをベースに作成した。

3.6.応答生成・出力部

対話用の音声合成は、単にテキストの読み上げを行うのではなく、音声入力の認識誤りや曖昧性に対処し、対話をスムーズに進行するため応答意味表現から音声を合成する。このため、対話中の不明な点の確認等に関して、実時間性と強調点の呈示を効果的に行うためイントネーションや発声速度の制御と実時間処理を重視して音声応答を出力する。また、音声メディアの他に、応答文のテキスト、注文品の品物および個数、店員の表情および動作の表示も併用し、マルチモーダル応答として親しみやすく、計算機の内部状態や対話の状況が把握しやすいように設計した。

4.むすび

本稿では、不特定ユーザを対象とした実時間音声対話システム構築のアプローチとシステム構成について述べた。今後、試作したすてむを用いて実環境の対話音声データを収集するとともに、マルチメデイアインタフェースとして、音声などのメディア理解と生成の効果的利用法と対話モデルについて検討する予定である。

参考文献

[1]小林他,音声研資, S85-15 (1985-6)
[2]速水他, 音声研資, SP91-101 (1991-12)
[3]竹林他, 信学論,J74-D-II Vol.2, (1991-2)
[4]坪井他, 音講論, 1-5-11(1991-10)
[5]坪井他, 音声研資, SP90-37 (1990-8)

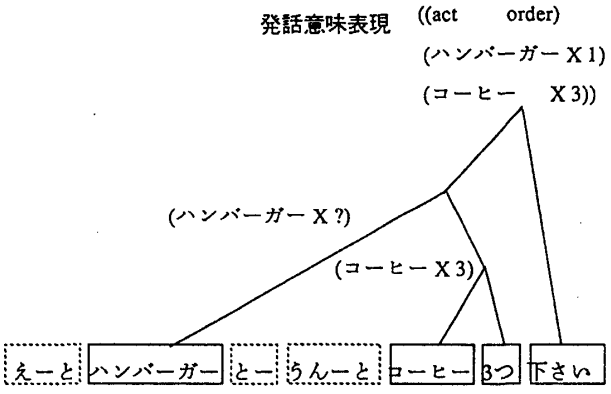

図2. 意味表現と解析木の例

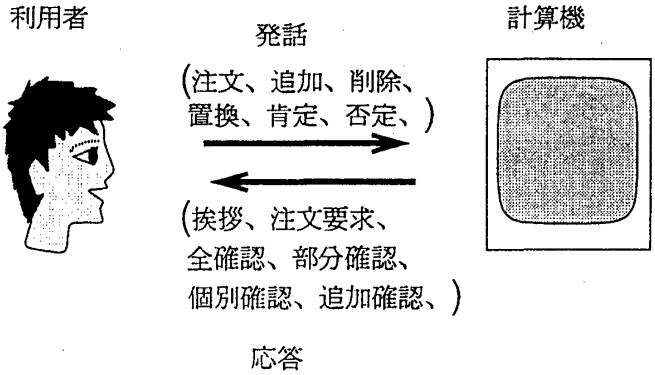

図3. 注文タスクの音声対話モデル