実時間音声対話システムTOSBURGの開発 (4) マルチモーダル応答

6N−8

山下 泰樹* 瀬戸 重宣** 橋本 秀樹*** 竹林 洋一**

*(株)東芝 関西研究所 **(株)東芝 総合研究所 ***東芝ソフトウェアエンジニアリング(株)

1.はじめに

これまで音声合成の研究は主として文-音声変換を目的として行われてきており、構文情報を表現するため韻律に関する研究が行われてきた[1,2]。また、最近は人間と計算機の対話システムのための合成が検討されるようになり[3]、意図、状況、感情を伝えるための研究も行なわれている[4]。自然なコミュニケーション手段である音声メディアでは、音声認識の誤りや曖昧性が避けられないという問題があるので、ユーザフレンドリな対話システム実現のためには、ユーザに適切な応答を出力し質問や確認を行ない、スムーズに対話を進行する必要がある。また、従来の対話システムでは、音声認識、合成、画面表示について個々に検討がなされているが、マルチモーダル応答という観点からの検討は不十分である。

今回、我々はハンバーガーショップでの注文をタスクとした、実時間で動作する対話システムTOSBURG(Task Oriented dialogue System Based on speech Understanding and Response Generation)を試作した[5,6]。本稿では、このシステムの特徴であるマルチモーダル応答出力[7]について述べる。

2.マルチモーダル応答

図1は、音声メディア以外に圧力マットと画像表示を用いた音声対話システムの構成図である。音声入力および人検出用圧力サンサーの入力は、音声理解部で処理され、その結果は入力意味表現として対話管理部に出力される。対話管理部ではそれまでの履歴に基づき、出力応答されるべき意味内容を決定し、出力意味表現として応答生成出力部に渡す。応答生成出力部は出力応答を出力意味表現に基づき生成する。

システムの応答は、音声応答に加え店員の表情、品物とその個数、音声入力の可否を示すアイコン、そして応答文の表示を併用しており、これらは対話管理部によって制御されている。合成音声は単に応答文を読上げるだけでなく、ユーザが省略した項目や、システムの理解が不確かな項目を確認するために、その項目を強調した応答を出力する。これによりユーザはシステムの状況を把握しやすくなり、ユーザへの確認がスムーズに行える。また、ユーザは画面上の店員に向かって発声することを自然に行なうことができ、人物の口の動きや表情で対話の進行状況や音声認識の信頼性を把握できる。応答内容は、音声出力とその応答文のテキストの他に、品物とその個数の表示も利用する。このようにして、応答内容をユーザに素早く伝達できる。さらに、システムがユーザの意図と異なる理解をしてしまった場合にでも、上記のマルチモーダル応答からユーザは容易にその誤りに気付くことができるため、安心して対話が行える。

3.意味表現からの応答生成

応答生成出力部は、対話管理部からの出力意味表現を基に応答文、合成音声と店員の表情を生成する。応答文は、出力意味表現により図2に示すような出力発話の文の種類が決まり、さらに注文品の情報をこれに埋め込んで応答文が生成される。合成音声は、生成された文に強調する語の情報を加えて基本周波数パターンを決定し、音声を規則合成する。表情は出力意味表現と対話のやり取りの状態から決定し、店員の口の動きは合成音声の出力時間長に合わせ、また、音声と画像の同期をとって行う。

このようにして、意味表現から音声、画像を生成すること

全注文確認: 全注文内容の確認応答
　例):「ご注文は、ハンバーガーを1つ、コーヒーを3つですね。」
部分確認:　注文の一部の品目についての確認応答
　例):「ハンバーガーは、1つですね。」
追加確認:　追加品目の確認応答
　例):「チーズバーガーを2つ追加ですね。」
置換確認:　置換された品目、サイズ、数の確認応答
　例):「コーヒーは1つではなくて2つですね。」
個別確認:　注文品目を1品目ずつ次々に確認する応答
　例):「1つずつ確認します。ハンバーガーは1つですね。」
再発話要求: ユーザに前発話の内容を繰り返すよう要求
　例):「すみません。もう一度お願いします。」
　　:

図2.出力意味表現と応答文例

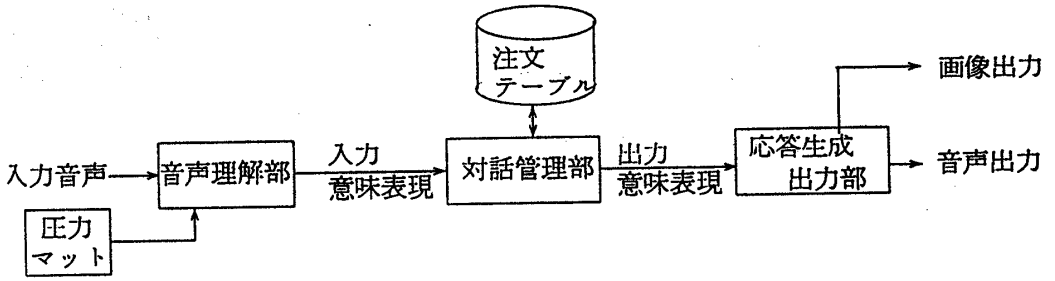

図1. 全体構成図

Development of Real-Time Speech Dialogue System TOSBURG (4) Multimodal Response
Yasuki YAMASHITA*, Shigenobu SETO**, Hideki HASHIMOTO***, Yoichi TAKEBAYASHI**
* Toshiba Kansai Research Lab., **Toshiba R&D Center, ***Toshiba Software Engineering

でマルチモーダル応答をユーザに同時に提示できるようになった。ユーザは必要に応じて視聴覚情報を適宜選択し、各メディアの特長を活かして、フレンドリでスムーズな計算機との対話が行える。

4. 対話における応答生成

実時間音声対話システムTOSBURGにおける応答例を図3に示す。まず、(a)人物検知によりスタートし、システムは合成音声で呼びかけ、同時にその応答文のテキストと微笑んだ店員の姿を表示する。このとき、応答出力はユーザの立ち止まりを検出した後、タイミングを考慮して、スムーズにユーザを対話へ導く。

(b)システムは、ユーザの発話を正しく認識できない場合、謝っていることを表す合成音声を出力するとともに、その応答文のテキストと申し訳なさそうな表情の店員を表示する(図4)。(c)ユーザの注文の発話を認識すると、確認のための合成音声を出力するとともに、そのテキストと普通の表情の店員の姿、注文テーブルの内容を表示する。表示する店員の表情は、出力意味表現と対話のやり取りの状況から決定する。注文テーブルの表示により、ユーザは、自分の注文をシステムが理解しているかどうかを短時間に確認できる(図5)。

また、(d)ユーザが個数を省略して追加注文した場合、システムはユーザが所望する個数を推論して、確認のため"1つ"を強調した合成音声を出力するとともに、注文テーブルの内容の表示にコーラの絵と数字1を加える。このように、音声合成出力では追加の確認を行う一方、表示ではこれまでの注文した内容を示すことにより、音声メディアの有する一過性の欠点を補うことができる。

(e)全ての注文の確認を終えると、お礼の挨拶を表す合成音声を出力するとともに、その応答文のテキスト出力とお辞儀をする店員の姿を表示して一連の対話を終了する。

5.むすび

本文では実時間音声対話システムの応答生成部について説明した。本システムでは、意味表現から音声や画像や文字からなるマルチモーダル応答を生成することにより、ユーザフレンドリかつ自然な応答が出力できる。今後、実システムを用いて各種評価実験を行い、音声およびマルチモーダル応答についての評価を行う。

参考文献

[1]箱田、佐藤: 文音声合成における音調規則、信学論D Vol.J63-D No.9,pp.715-722(1980).

[2]広瀬、藤崎、山口、横尾: 統語構造を利用した日本語文音声の基本周波数パタンの合成、音響学会音声研資S83-70 (1984).

[3]山下、水谷、溝口:合成音出力における概念表現の利用、信学技報SP89-115 (1990).

[4]K.Sheahan,Y.Yamashita,Y.Takebayashi : Synthesis of Nonverval Expressions for Human-Computer Interaction,音講論 2-4-6,pp237-238,(1990-3)

[5]竹林、他: 実時間音声対話システムTOSBURGの開発(1)システム構成、 本大会予稿集 (1992-3).

[6]新地、他: 実時間音声対話システムTOSBURGの開発(3)対話処理、 本大会予稿集 (1992-3).

[7]M.M.Taylor : The Structure of Multimodal Dialogue ,North-Holland(1989).

(a)ユーザ： 圧力センサ-を踏む
(b)システム:「いらっしゃいませ、ご注文をどうぞ。」
(c)ユーザ: 「え〜、ハンバーガーを2つとあの〜コーヒーを2つあ〜お願いします。」
(d)システム:「はっきり聞きとれませんでした。もう一度お願いします。」
(e)ユーザ: 「ハンバーガー2つとコーヒー2つ下さい。」
(f)システム:「御注文はハンバーガーを2つ、コーヒーを2つですね。」
(g)ユーザ: 「コーラの大も下さい。」
(h)システム:「コーラの大を1つですね。」
(i)ユーザ: 「はい。」
(j)システム:「御注文はハンバーガーを2つ、コーヒーを2つ、コーラの大を1つですね。」
(k)ユーザ: 「それでいいです。」
(l)システム:「有難うございました。」

図3.対話の例

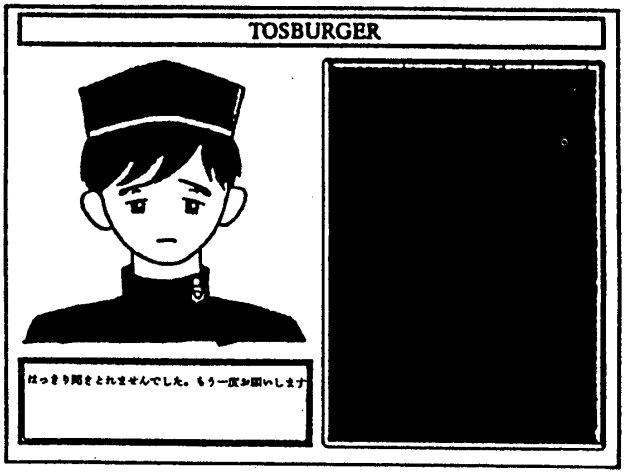

図4. 表示画面の例(申し訳なさそうな表情)

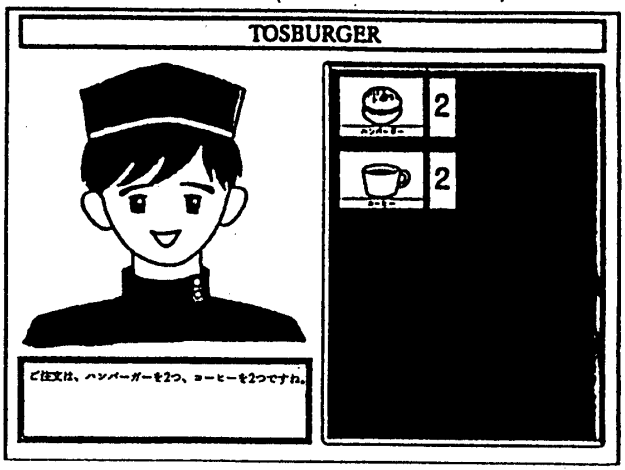

図5. 表示画面の例(普通の表情)