テレビ操作のための音声対話インタフェースの試作

後藤淳　小峯一晃　森田寿哉　金淵培　浦谷則好

日本放送協会 放送技術研究所

{goto,komine,moritat,kimyb,uratani} @strl.nhk.or.jp

1.はじめに

近年、ＢＳ放送とＣＳ放送の普及、ＢＳデジタル放送によるデータ放送開始、地上波デジタル放送の開始が予定されるなど、テレビ受信は多彩になる一方、操作環境は、複雑化しつつある。さらに、ＶＴＲやディスクレコーダ、ＤＶＤ、ゲーム機器など、テレビに接続される周辺機器も多種多様になり、それぞれに異なるインタフェースを持つ機器を操作するのは、高齢者はもとより、一般の人々にとっても、難しい作業になってきている。

我々は、以前、テレビに対する使い易いインタフェースの一つの提案として、音声コマンドを用いたリモコン[1]を試作した。当所で行った一般公開のデモでは、音声を用いた多機能テレビの操作に関心が集まり、大変好評であった。また、最近、音声認識リモコンをオプションとして持つテレビ受信機がメーカーより市販され、音声を用いたテレビ操作インタフェースは現実のものになりつつある。

しかしながら、誰にでも容易に操作できるインタフェースが望まれるなか、自然な言葉でテレビを操作できるまでには至っていない。自然言語を用いて、番組選択や検索、周辺機器操作の指示が可能になれば、多チャンネル化し、機能が複雑化した視聴環境下では、非常に有効なインタフェースになるであろう。

そこで、本稿では、まず、テレビ操作における自然言語対話の傾向を把握するために行ったWOZ（Wizard of OZ）方式による対話データ収集予備実験の内容について述べる。次に、その傾向を踏まえ、試作したテレビ操作インタフェースのシステム構成及び概要を報告する。

2.テレビ操作の対話データ収集予備実験

人は、テレビに知能があるとき、テレビに向かって「どのような言葉を発するのか」、「そのとき、テレビはどのような応答を返すべきか」等の参考データを得るため、WOZ方式を用いた対話データ収集予備実験を実施した。対話者は、男性10名、女性10名（24歳～31歳、平均28.7歳）である。

被験者に対して、「テレビ画面に出現するキャラクタが、発話意図をなんでも理解でき、あなたに代わってテレビを操作してくれます」との事前提示を行い、自由に発話を行わせた。実験装置では、視聴可能なチャンネル数が、BSデジタル放送12chと地上波放送７chの計19chであり、当所で試作したEPG（Electric Program Guide）及び番組検索GUIを使用した。

実験過程では、WOZシステムの裏側にいる実験担当者２名が、被験者が発した言葉を判断し、適切な動作を返すように操作した。音声による応答は、予め準備していた応答文を選択したり、直接キーボードから入力して、音声合成器を用いて生成した。被験者が沈黙した場合には機能紹介や発話を促す応答を返した。

図１に実験で得られた音声を書き起こしたデータの一部を示す。

```
00:27:08, 被験者,「ふふっ、映画を探してます。」
00:30:23, WOZシステム,「ジャンルから選ぶこともできます。
                    ジャンル別番組表をご覧になりますか？」
00:36:25, 被験者,「はい。」
00:38:00, WOZシステム,「はい、了解しました。」
00:47:02, 被験者,「あっ。」
00:47:02, WOZシステム,「お好きなジャンルを選んでください。」
00:50:04, 被験者,「んー、あ、あ、えー、
                    やっぱりバラエティにします。」
00:55:11, WOZシステム,「はい、了解しました。」
01:02:06, 被験者,「うん。」
01:03:29, WOZシステム,「見たい番組を選んでください。」
01:08:27, 被験者,「え～と、
                    もっと下の画面に～いきたいんですが。」
01:12:09, WOZシステム,「はい、了解しました。」
01:15:23, 被験者,「う～ん、もうちょい。」
01:17:27, WOZシステム,「はい、了解しました。」
```

図１ 実験データ例

すべての話者について、番組名、ジャンル、時間、操作コマンドなどのカテゴリに属する言葉を用いた発話が目立った。これは、応答の際、テレビ画面上に表示されるEPGや検索インタフェースなどを基に機能を推測することで、それらを実行するため以外の自発的な発話が制約されたこと

が考えられる。また、「自由に何を喋ってもよい」という提示を行っていても、「テレビに対して、こんなことを言っても仕方がない」「機械に理解できる訳がない」など、現在あるテレビ機能に対する認識による制約が働いたためと思われる。

そこで、現段階での音声対話を用いたテレビ操作インタフェースを考えるとき、『テレビに対する音声発話の大部分は、番組名や時間など番組検索時に使用される単語や、チャンネル変更などの直接操作コマンド語を含んでいる。』という前提のもとで、システムを構築することとした。

3. システム概要

試作したシステムは、誰にでも容易に操作できるように、自然言語による音声操作でテレビ受信機の操作を可能とした。システムの構成は、音声認識、対話処理、出力処理、データ管理、形態素解析システムの各処理プログラムがPC上で動作し、それぞれが相互に通信することで稼働している。また、出力処理プログラムの赤外線制御により、TVチューナーやVTR等の周辺機器を操作することができる。

機能的には、地上波とBSデジタル放送の19チャンネルからの番組選択、番組の録画予約、また、インターネット上の番組データを利用した番組関連情報の検索、番組や出演者関連Webページへのブラウジング機能などがあり、すべて音声による自然言語での対話形式で操作することができる。

図２にシステム構成を示し、以下に音声認識処理、対話処理、出力処理、発話データ管理インタフェースの概要を述べる。

3.1 音声認識処理

ユーザが発した言葉は、当所が開発した連続音声認識エンジンを用い、認識する。本エンジンは、ニュース番組の音声から自動で字幕を作成するために開発されたものであり、逐次的に認識結果を確定していくアルゴリズム[2]を用いることで、ほぼリアルタイムの認識が可能である。テレビ操作における対話に限定し使用するため、ニュース解説用に作成した言語モデルに、対話収集実験で得たデータとテレビ操作用入力テンプレートから生成したデータを追加し、テレビ操作用言語モデルとして作成した。

テレビ操作における番組検索時には、番組タイトルや放送局名、芸能人や歌手等の出演者名などの語彙が高い確率で出現する。また、これらは、新しい名称が頻繁に更新されるため、番組情報データ更新時に、インターネットより得た番組関連情報の語彙が未登録な場合は、形態素辞書に自動登録される仕組みを持たせた。この際、番組名は、例えば「その時歴史が動いた」など、色々な品詞の言葉が組合わさったタイトルがあるため、これらを１つの名詞の形態素として登録する。同時に、対話処理部が持つ入力テンプレートから、新語彙を使った入力文を自動生成し、対応した言語モデルを作成する。

また、毎日更新される番組情報に対して、時期依存言語モデル[3]を用い、システムが動作している日の番組情報に重み付けをつけた言語モデルを自動的に作成することにした。これにより、稼働している日近郊の番組名、出演者等の番組情報を含んだ発話の音声認識の精度を高くすることができる。

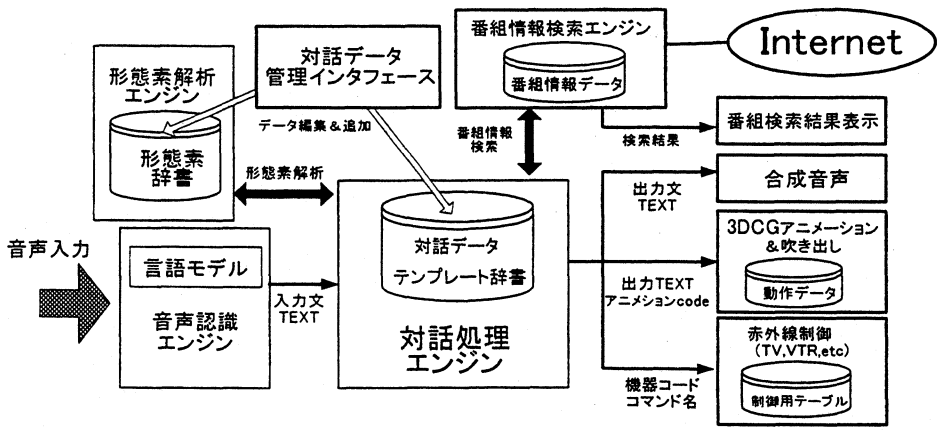

図.2 テレビ操作インタフェース構成

3.2 対話処理

本システムで使用している自然言語対話処理では、形態素単位でのパターンマッチングを用いた方法を使用している。入力文を解析し得られた形態素群が適合するテンプレートを確定し、発話の意図を把握する。このアプローチは、FACTS（FIPA実証プロジェクト）でのエージェントテレビ[4]や文書検索システム[5]などの対話システムにも使用されている。

語の表層的な分析で意図を理解するのは、一般会話全般については困難である。しかし、テレビ操作における対話のみにタスクを限定し、WOZ実験の結果より仮定した『テレビ操作の発話は、ある特定のカテゴリに属する語を含んでいる』とすれば、表層的な分析でも、発話の意図を抽出することは、ほぼ可能である。

3.2.1 形態素解析処理

音声認識処理で得られた入力文字列は、形態素解析処理により、品詞や基本形等の情報を持った形態素列に変換される。形態素解析された入力文は、形態素単位で、入力文テンプレートと比較される。なお、形態素解析処理用システムとして、『茶筅 version 2.2.6』[6] を用いた。形態素の定義についても基本的に『茶筅』の辞書に準じている。

3.2.2 テレビ操作における重要度

『茶筅』の形態素辞書に登録済みの形態素に対し、テレビ操作時における重要度に応じたスコアを設定した。重要度の設定は、品詞、単語、単語クラスにより設定することができる。単語クラスは、同種類の意味を持つ形態素の集合である。

テレビ操作において、番組名、人名、ジャンル、時間、放送局、操作動作などを表す単語クラスは、番組選択や検索の際に、重要な意味を持っているため、重要度を特に高く設定している。さらに、上記の単語クラスについては、入力文から抽出するスロットとしても設定している。これらのクラスに属する形態素が、入力文に含まれる場合には、発話履歴として保存し、対話の連続性を持たせるための遷移情報に利用している。

3.2.3 テンプレート

発話を特定するための入力用テンプレートには、表１に示すメタキャラクタ[7]から成るテンプレートパターンを使用している。複数のテンプレートがマッチする場合は、テンプレートに用いた単語及び単語クラスの重要度の和が高いものを選択する。

（例）

テンプレート　*[@時間][@映画]*[見る|探す]*

入力文　今日、ローマの休日が見たいんだけど。
ねえ、マイフェアレディを探して。

表１　メタキャラクタ

メタキャラクタ	機能
*	任意の数の単語
+	1つの単語
!	一致不可の単語
{}	{}内の単語,または、未存在
[]	[]内の単語
()	任意の順序でも可
@	単語クラス
|	単語区切り or の意
,	単語区切り and の意

メタキャラクタを用いたテンプレートパターンの作成は、現在、人手により行っている。そこで、大量の対話データをシステムに反映させるため、メタキャラクタによるテンプレートパターンを作成せずに、発話データそのものを登録可能な手段を用意した。これにより、WOZ方式などで得た対話データをすぐにシステムに反映させることができる。この場合も、テンプレートパターンを使用する時と同様に、登録した発話データと入力文を形態素毎に比較し、マッチした形態素の重要度の和が最も高いものが選ばれる。

3.2.4 入出力対応

入力テンプレートに対する出力は、状況により様々な応答が考えられる。そのため、入力文から抽出したスロット値、過去の応答によるシステムの遷移状態、検索結果、現在時刻などの条件をもとに、予め定義してある複数の出力テンプレートから選択し、出力文が生成される。その際には、機器制御コマンド、及びCGアニメーション用の動作コードも最適なものが選択される。

3.3 出力処理

出力処理には、検索結果の表示、合成音声、CGアニメーション、吹き出しによる音声内容表示、赤外線でのリモコン機器制御がある。

(1)番組検索の結果提示

音声対話を用いて番組情報を得る際に、対話の補助として番組情報の検索結果を画面に表示する。また、出演者名や番組タイトルなど番

組関連の語彙に関連があるホームページ表示が可能である。

(2)合成音声

音声合成の出力処理部は、市販の音声合成SDK（OKI製）を用いて作成した。出力テンプレートから生成されたテキストを受けて合成音声を出力する。音量、早さ、抑揚、声質などの設定が可能である。

(3)CGアニメーションによるキャラクタ動作

OpenGLを用い、予め用意してある動作データをリアルタイム3DCGで出力する。同時に音声合成に合わせた応答内容をキャラクタの吹き出しとして生成する。

(4)赤外線による機器制御

入力テンプレートと、対話システムの状態パラメータから、制御する機器及び制御内容を抽出し、赤外線によりテレビ、デジタルHVチューナー、VTR等を制御する。

学習機能を持ち、リモコン信号を制御用データベースに登録することで、リモコン操作できる殆どの家電が制御可能である。

3．4 番組情報検索処理

番組の放送時刻や放送事業者名及び詳細情報などの番組関連データは、インターネットより取得し、番組情報データベースへ格納する。対話処理エンジンからの検索依頼により、対応するデータを返し、検索結果を情報提示部へ送信する。

現在、番組関連データは、インターネットＴＶガイド(www.tvguide.or.jp)及びＮＨＫオンライン(www.nhk.or.jp)などの各放送事業者のホームページから取得している。

3．5 対話データ管理インタフェース

対話データの入力や出力テンプレート及び入出力条件を作成するには、非常に手間と時間が必要とされる。そこで、データ作成支援ツールとして、対話処理エンジンを動作させながら、入出力テンプレート、応答の条件、形態素辞書、クラスなどを編集可能な対話データ作成インタフェースを実装した。

この管理用インタフェースを用いることで、データの追加や変更及びシステム動作のチューニングを容易に行うことができる。また、スロットの追加が、単語クラスを追加する場合にオプション設定が可能であり、様々な用途に応じた対話に対応できるようにしている。

3．6 システム動作例

試作したテレビ操作インタフェースシステムを用いて、動作実験を行った。その結果例を図３に示す。実験のための入出力テンプレートや、CGアニメーション、各機器の動作などのデータは、対話データ管理インタフェースを用いて、WOZ方式を用いた予備実験の発話データを参考に手動で設定したものである。

ユーザ　：今日のドラマは、何がある？
システム：今日のドラマは、「ほんまもん」のほか、
　　32番組があります。どれにしますか？
ユーザ　：じゃあ、16時からのだけ、出して。
システム：「水戸黄門」のほか、３番組があります。
ユーザ　：それじゃあーね、水戸黄門が見たいな。
システム：水戸黄門は、１時間20分後です。
　　録画予約しますか？
ユーザ　：お願い。
システム：録画予約しました。

図３ システムの動作例

4．まとめ

誰にでも使いやすいテレビ操作インタフェースを目指し、自然言語を用いた音声対話でテレビ操作を行うインタフェースシステムを試作し、その概要を報告した。現在、本システムを用い、テレビ操作における音声対話インタフェースの有効性と問題点を検証しているところである。

今後は、WOZ方式によるコーパス収集実験を引き続き行い、テレビとの対話データを集めると共に、試作インタフェースシステムの開発を継続し、柔軟な出力文の生成や対話履歴からの学習など、今回、実装できていない機能を追加する予定である。

参考文献

[1]小峯他:"音声リモコンシミュレータの開発"
第53回NHK技術報告会予稿集,pp19-20
[2]T.Imai:"Progressive 2-pass Decoder for real-Time Broadcast News Captioning",
ICASSP-2000,Vol3of 6,pp1559-1562(2000)
[3]小林他:"ニュース音声認識のための時期依存言語モデル",情報処理学会論文誌、Vol40 No.4(1999)
[4] 村崎他:"エージェントを応用した次世代テレビエンターテーメントシステム",信学技報、AI2000-25(2000)
[5] 高野他: "自然言語を用いた対話形式による文章検索における辞典情報の利用",
信学技法,NCL2000-7,pp49-54(2000)
[6] 松本他:形態素解析システム「茶筅」ver.2.2.6 使用説明書,2000.4
[7] 金淵培:"エージェント技術の放送への応用",映像情報メディア学会誌vol.52,No.4,pp447-451(1998)