音声認識を用いた放送番組リクエストシステム

安藤彰男　今井亨
ando@strl.nhk.or.jp　imai@strl.nhk.or.jp
ＮＨＫ放送技術研究所音響聴覚研究部
〒157 東京都世田谷区砧 1-10-11

将来のマルチメディア時代において、対話的に視聴する放送サービスを実現するため、音声対話システムの研究を進めている。その第一段階として、音声によって放送番組の検索・リクエストを行う実験システムを構築した。本システムでは、自由発声された音声中から、番組ジャンルや番組名などのキーワードを抽出するワードスポッティング型の音声認識方式を採用している。本システムのために新たに開発したワードスポッティング法を用いて、女性話者４名が発声した音声を認識する不特定話者認識実験を行ったところ、従来法による認識率が４７％であったのに対し、８６％の認識率を得た。

A TV Program Retrieval System Based on Speech Recognition

Akio Ando and Toru Imai
NHK Science and Technical Reserch Laboratories
1-10-11 Kinuta Setagaya Tokyo 157 JAPAN

A TV program retrieval system based on speech recognition has been developed as a part of a spoken dialogue system which will be useful as one of multimedia broadcasting services. The system spots a keyword such a category or a title of a program in spontaneous speech based on a new word spotting method and show the corresponding program on the TV display. Experiments showed that the average recognition accuracy of speaker independent recognition for 4 female speakers is improved from 47% to 86% by using the new word spotting method.

1．はじめに

将来のマルチメディア放送には、ＶＯＤ（Video On Demand）に代表される双方向放送など、新しい放送サービスの実用化が望まれているが、このようなサービスは、視聴者が、高機能、多機能な受信端末を操作することによって初めて享受しうるものである。高機能性・多機能性を損なうことなく、かつ、お年寄りを含め誰にでも使いやすいものとするには、対話することにより視聴者のニーズを的確に把握するインテリジェントな機能を、受信端末に持たせる必要がある。このためには、音声認識の技術が不可欠となるが、端末の使い勝手から考えると、音声認識方式としては、なるべく言い回しを制限しない方式を採用することが望ましい。

本稿では、このような受信端末を実現するための第１段階として構築した音声認識による番組リクエストシステムについて述べる。このシステムでは、言い回しを制限しない音声認識を実現するため、ワードスポッティングによって番組名等を認識することとした。音声対話に関する研究は盛んであるが[1]、ワードスポッティングに基づく方式は数少ない[2]。それは、ワードスポッティングを用いた場合の技術的な難しさに起因すると思われる。本システムでは、新しいワードスポッティング法を導入することにより、スポッティング性能の向上に努めた。以下、２．でシステムの概要を、３．でワードスポッティング方式について述べた後、４．で認識実験の結果を示す。

2．番組リクエストシステムの概要

本システムは、音声認識部、番組検索・表示部から構成された実験システムである（図１参照）。音声認識部で連続発声された音声の中からキーワードを抽出・認識した後、番組検索・表示部で、認識されたキーワードに基づき番組を検索し、検索された番組の動画像をディスプレーに表示する。現在のところ、システムには６１番組の動画および音声各１０秒分が格納されている。番組をリクエストするに当り、番組名あるいは番組を示すアイコン（または静止画等）をディスプレー上に全て表示すると、表示のサイズが小さくなりすぎる。そこで今回は、番組を１２のジャンルに分け、画面上には、番組のジャン

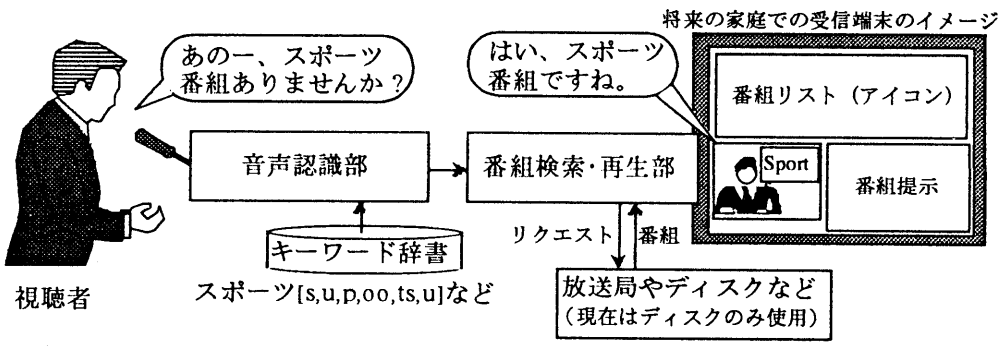

図１ 音声による番組リクエストシステムのイメージ図

ルを示すアイコンと、特定のジャンルが選択された場合には、そのジャンルに含まれる番組を示すアイコン（実際には、番組冒頭の静止画）を表示することとした。従って、本システムを使用する場合には、具体的な番組名を指定する前に、番組ジャンル名を指定する必要がある。表示される画面の例を図２に示す。

音声認識部は、ワークステーション（DEC 3900）とＤＳＰによって構成した。また、番組検索・表示部には、Macintoshを用いた。音声認識部で認識されたキーワードは、コード化されてRS232C経由でMacintoshに送られる。Macintoshでは、認識結果が番組のジャンル名であった場合には、対応するアイコンの背景色を変化させることによって、そのジャンルが選択されたことを示し、認識結果が番組名であった場合には、ディスクの中から対応する番組の音声・動画像を呼び出して再生する。図２はスポーツのジャンルが選択された後、野球番組がリクエストされた場合の画面を示している。

発声は、例えば「あのー、スポーツ番組ありますか」、「じゃあ、野球にしてください」など、番組ジャンル名あるいは番組名を含んだものを対象としている。キーワードのみを認識するため、発声様式は自由であり、不要語を含んでも良い。現状では、単一のキーワードのみを認識するシステムであるが、今後は、複数のキーワードが発声された場合や、言い直された場合への対応を検討する予定である。

図２ 表示画面の例

3．新しいワードスポッティング法[3]

本節では、音声認識部で用いたHMM照合に基づくキーワードスポッティング法について述べる。HMMを用いたスポッティング法としては、いくつかの方法が提案されているが（例えば文献(4)など）、始終端フリーのHMMを用いた場合、キーワード以外の部分において誤検出する場合が多いため、ヒューリスティックスなどを用いて認識性能の向上を図ることが一般的である[5]。本節では、これに対して、自由な発話からワードスポッティングすることを目的として考案した、入力音声の音響的特徴のみを用いてHMMスポッティングの性能を向上する方法について述べる。

3.1 処理の流れ

本方式は、母音標準パターンを用いた母音認識結果に基づき、HMMワードスポッティングにおけるスポッティング区間を制御する新しいスポッティング方式である。ブロック図を図3に示す。処理の流れは次のとおりである。

(1)入力音声を、ＬＰＣ分析、零交差波分析により分析する。

(2) 入力の各フレーム毎に、母音標準パターンとの距離を計算し、その距離に基づいて、母音認識を行う。この際、母音認識における置換誤りを考慮し、各母音区間毎に複数の母音候補を出力することとした。また、母音の無声化に対応するため、破裂性子音と摩擦性子音を検出することにより、母音無声化が起こっていると思われる部分を推定する。これらの処理により、最終的には母音ラティスの形で認識結果が得られる。

(3)キーワードの発音記号から抽出された母音列によって（2）で得られた母音ラティスを探索することにより、キーワードが存在する区間を推定する。

(4)（3）で推定された区間について、入力音声を音素HMMを連結したキーワードHMMと照合し、尤度を計算する。HMMとしては、離散型のものを用いた。

(5)（4）で計算された尤度に基づき、スポッティング結果を判定する。

図3の各ブロックと、上記の処理の流れの対応を述べると、

音響分析部：処理（1）

キーワード存在区間推定部：処理(2)、(3)

尤度算出部：処理（4）

スポッティング判定部：処理（5）

である。また、ハードウェアとの対応については、ＤＳＰによって処理(1)及び処理(2)のうち距離計算の部分を、また、ワークステーションによって処理（2）のうち母音認識の部分と(3)(4)(5)の処理を行うこととした。

処理（2）については、別の文献[6]で述べた方式をそのまま利用している。処理（2）の結果、母音区間の情報だけでなく、それぞれの区間における各母音の尤度も算出される。処理（4）は通常のHMMを用いた尤度計算であり、処理(5)は尤度の比較である。処理（3）の部分、即ち母音認識結果に基づくキーワード区間の推定については、次節で詳述する。

3.2 キーワード区間の推定

母音標準パターンを用いて連続音声を認識

する場合には、母音の挿入、脱落を避けることは難しい。本稿では、より致命的な結果をもたらす母音脱落を可能な限り防ぐことを前提とし、その結果生じる多くの母音挿入を許容しながらキーワード区間を推定する方法について述べる。処理（2）によって、母音区間とその認識結果が計算される毎に、それまでに得られた母音ラティスの中から、キーワード中の母音列を探索する（母音列は、キーワードに対応する発音辞書から、母音部分のみを取り出すことにより得られる)。重複した処理を避けるため、探索は後ろ向きに行う。すなわち、最後に得られた母音認識結果と、キーワード中の母音列の最後の母音との照合から探索を開始し、この照合が成功した場合には1つ前の母音部分の照合に移行するという処理を繰り返す。その結果、キーワードの最初の母音まで照合が成功した場合

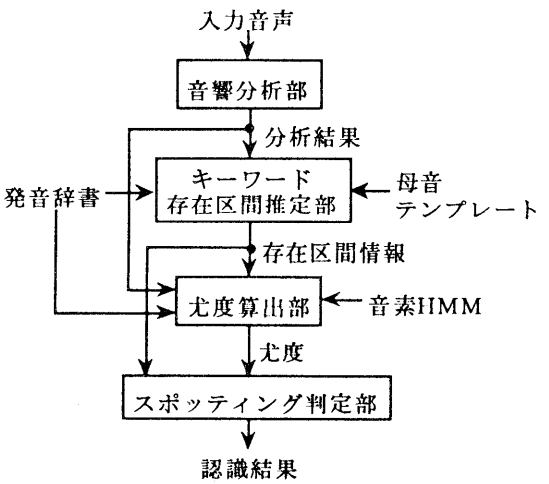

図3 ワードスポッティング法のブロック図

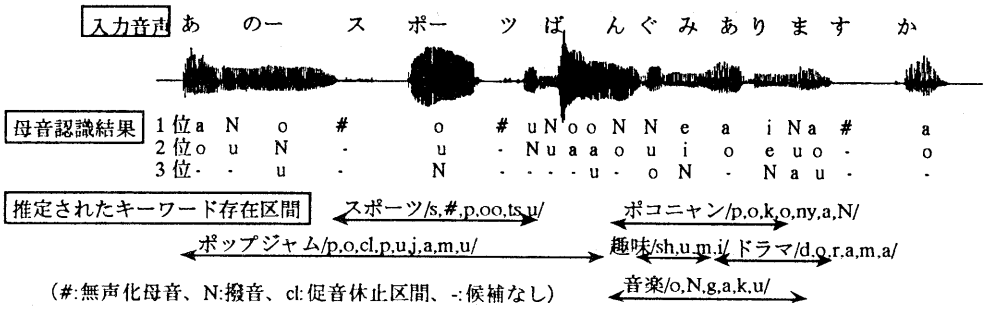

図4 母音認識による区間推定の例

に、対応する入力部分をキーワード区間の候補とする。この際、母音挿入を許すため、ある母音区間において、キーワード中の母音が候補として得られていない場合でも、そのキーワードによる照合を中止せず、1つ前の母音区間へ探索を続行する。また、入力中に母音の無声化が検出された場合には、キーワードの発音辞書に予め登録してある無声化母音のラベルと対応を取る。

通常、1回の発声に対して、各キーワードごとに複数個の区間候補が得られる。この候補数は、母音挿入を許容しているため、場合によっては非常に多くなる。そこで、以後の処理の簡略化のため、各区間候補毎に、対応する母音列の尤度（これは、3.1節で述べた処理（2）によって与えられる）の総和を計算し、1つのキーワードに対する区間候補数が、予め設定した数を越える場合には、この尤度の総和が小さい順に区間候補を削除する。母音認識結果による区間推定の例を図4に示す。

4．認識実験

本システム音声認識部の認識性能を調べるために行った認識実験の結果を示す。ＡＴＲデータベースの女性話者２０名が発声した文音声データ（セットＡ及びＣ）を用いて、母音標準パターン、音素ＨＭＭ（離散型）及びコードブックの学習を行い[7][8]、別の女性話者４名が発声した、番組リクエストのための各４８文を認識する不特定話者認識実験を行った。評価用音声には、各文ごとに1つずつのキーワードが含まれている。発音辞書には、５６のキーワードを登録した。認識実験の結果、母音認識によるキーワード区間推定を行わない場合のキーワード認識率が４７％であったのに対し、本方法の利用により、認識率が８６％まで向上した。実験の詳細については、別の論文（文献（3））にゆずる。

5．まとめ

本稿では、音声による番組リクエストシステムの概要を説明した。今後は、雑音対策など、実環境においても安定して動作するための対策を検討すると共に、意味処理を導入した音声理解を行うことにより、複数キーワード抽出や登録番組数が増えた場合への対応を検討する。最終的には、インテリジェントな受信端末の開発を目指す。

参考文献

(1) 小林："対話音声の認識技術"、音響学会誌 50 no.7 pp.563-567 (1994)

(2) 竹林："音声自由対話システムTOSBURG II－ユーザ中心のマルチモーダルインターフェースの実現に向けて－"、信学論 (D-II) J77-D-II no.8 pp.1417-1428 (1994)

(3) 今井、安藤："母音認識による区間推定を用いたＨＭＭワードスポッティング"、音講論集 平成7年9月 1-Q-19 (1995)

(4) 今村："ＨＭＭによる電話音声のスポッティング"、信学技報 SP90-18 (1990)

(5) 河原、堂下："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"、信学論 (D-II) J78-D-II no.7 pp.1013-1020 (1995)

(6) A.Ando and E.Miyasaka："A New Method for Estimating Japanese Speech Rate", Proc of ICSLP94 pp.731-734 (1994)

(7) 安藤、尾関："誤認識関数を最小化する標準パターン学習アルゴリズム"、信学論 (A) J76-A no.4 PP.580-588 (1993)

(8) T.Imai and A.Ando："An HMM Learning Algorithm for Minimizing An Error Function on All Training Data", J. Acoust. Soc. Jpn. (E) 13 no.6 pp.369-378 (1992)