4E－5 ニュース音声データベースの検索システムの試作

西崎博光　中川聖一
豊橋技術科学大学 情報工学系

1 はじめに

近年、多くのニュース番組が放映されているが、視聴者は、知りたいニュース記事のみを見つけたいという欲求にかられる [1]。必要な記事を効率よく検索するには、放送されたニュース音声を音声パターンのままで検索するのではなく、一旦文字列に書き起こし、テキストデータベース化する必要がある。大量のニュース音声を人手で書き起こすのは不可能に近い。そこで大語彙連続音声認識システムを用い、ニュース音声を自動的に書き起こす必要性がでてくる。

そこで本研究では、自動的に書き起こしたデータベースでの検索性能を調べるため、まず、実際のニュース音声に対して、音声認識システムにより書き起こし、検索用データベースを作成した。このデータベースと正確に書き起こしたデータベースに対して、キーワード群を使って検索された記事の一致率を求め、比較した。実験の結果、単語認識率が低いにもかかわらず高い一致率が得られた。

さらに、キーワードを音声で入力することを考えた場合、必ずしも正しく認識されるとは限らず、この場合余計な記事が検索される恐れがある。そこで、キーワード間の関連度を用いた検索手法についても述べる。

2 ニュース検索システムの概要

今回試作した、ニュース検索システムの概略図を図 1に示す。

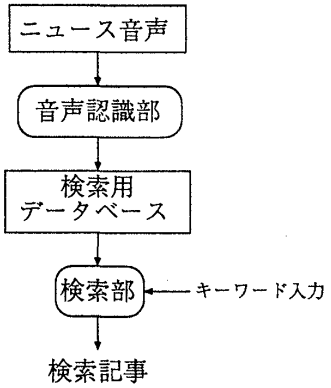

図 1: システムの概略

まず、ニュース音声を音声認識システムに通し、自動的に検索用データベースを作成する。これを元に、入力キーワードに応じて記事を検索部で検索する。

検索部では全文検索を行なっているが、インデックス法 [2] を用いることで、高速な検索を可能にしている。検索キーワードは、テキスト入力でキーワードをいくつか入力する。すべてのキーワードが完全に一致した記事のみを出力する。ただし、入力キーワードが多い場合は、全部が一致しなくてもその大部分が一致している記事を出力する。もし、入力キーワードが未知語だった (音声認識で使用した語彙辞書に入っていない) 場合は、音節列 (かな文字列) 単位の DP マッチングを行なうようにしている。

3 検索実験

前節で説明した検索システムで検索実験を行なった。実験対象の音声データは、NHK ニュース (1996 年 6 月 1 日～7 月 10 日) で、記事の数は 280 記事、文数で 1324 文である。ニュース音声の書き起こしに使用した音声認識システムの条件を表 1に示す。言語モデルは第 1 パスでは語彙サイズ 20000 の単語 bigram、第 2 パスでは trigram を使用している。この音声認識システム [3] を使用した場合、ニュース音声 (背景雑音なども混入されている) の単語カバー率は 91.0%、単語認識率は 49.8%となった。

表 1: 認識実験の実験条件

音響モデル	
5 状態 4 出力分布 (2 混合ガウス分布, 全共分散行列)	
離散継続時間分布付き連続出力分布型 HMM	
音節カテゴリ数	113 音節
サンプリング周波数	12kHz
窓関数	21.33ms ハミング窓
フレーム周期	8ms
分析	14 次元 LPC 分析
特徴パラメータ	
LPC メルケプストラム (10 次元 × 4 フレームの特徴量を KL 展開で 20 次元に圧縮)	
＋ Δ ケプストラム (10 次元)	
＋ ΔΔ ケプストラム (10 次元)	
＋ Δ パワー ＋ ΔΔ パワー	

キーワードを選択するため、1 記事につき 3 人の被験者にキーワードを 3～5 個選んでもらい、3 人とも共通に選んだ単語の集合をキーワード群とした。

自動的に書き起こしたデータベース (A) と、正確に書き起こしたデータベース (B) に対して、キーワード群を使って検索された記事の一致率を求めた。一致率とは、あるキーワード群で (B) のデータベースに対して検索された数個の記事が、同じキーワード群で (A) のデータベースに対して検索した場合に、どれだけ検索されたかを表している。

実験結果を表 2に示す。1 キーワード群当たり検索された記事数は (A) のデータベースで平均 2.4 記事、(B) のデータベースで平均 2.1 記事、一致率は 86.3%であり、対象記事 30 記事中で (A) のデータベースで 28 記事 (93.3%) が正しく検索された。また、余計な記事が検索された割合は、1.7%と非常に低くなっている。

表 2の実験結果を見ると、単語認識率が 49.8%とかなり低い値になっているにもかかわず一致率が高くなっている。これは、キーワードとなりうる単語 (異なり数で 104 単語、総数で 2338 単語、但し複合

An Information Retrieval System of Broadcast News Speech Documents
Hiromitsu Nishizaki, Seiichi Nakagawa
Department of Information and Compure Sciences, Toyohashi University of Technology

表 2: 実験結果

検索対象記事：	30
一致率	：86.3%
検索率	：93.3%

語が多い) の認識率 (95.0%) が高くなっているためである。全体の音声認識率は検索性能にあまり影響しないということが言える。

4 キーワードの音声入力

前述の実験では、キーワードの入力がテキストであったが、音声での入力も考えられる。音声によるキーワード入力では、キーワードが認識された時、

- キーワードが正しく認識された
- 違う単語に誤って認識された
- 同音意義語が存在する

という場合が考えられるが、機械には認識結果が正しいキーワードかどうかわからないので、どの場合も単語のマッチングを行なうしかない。同音意義語の場合は、同じ読みの単語すべてを行なう必要があるし、また同音意義語でなくても、認識結果の尤度の高い候補単語を複数個使って検索することも考えられる。

こういった場合、実際にユーザーが意図しない記事を大量に含む検索結果が得られることになるが、これらの記事を絞り込んでいく必要がある。

その方法として、キーワード間の関連度を用いたキーワードの絞り混み手法を提案する。関連度とは、ある2つのキーワードがどれくらい関係しているかを表す尺度で、以下の値を用いる。

- 共起頻度の利用

  2つの単語間の関連度を求める際に、ある記事において、ある単語とどの単語が同時に同じ記事に出現しやすいかという情報を用いる。

  2つの単語をそれぞれ、W_1,W_2としたとき、これらの W_2 の W_1 に対する関連度 $R(W_1,W_2)$ を以下のように計算する。

$$R(W_1,W_2)=\frac{1}{2}\{\frac{f(W_1,W_2)}{f(W_1)}+\frac{f(W_1,W_2)}{f(W_2)}\}$$

  $f(W_i)$:W_iが出現した記事数 $(i=1,2)$

  $f(W_1,W_2)$:W_1,W_2が共に出現した記事数

- 相互情報量の利用

  相互情報量は、単語の共起や関連を客観的に表す尺度として用いられる。2つの単語 W_1,W_2 の相互情報量 $I(W_1;W_2)$ は、W_1 と W_2 を同時に観測する確率 $P(W_1,W_2)$ を、W_1 と W_2 を独立に観測する確率 $P(W_1)$,$P(W_2)$ と比較する。

$$I(W_1;W_2)=\log\frac{P(W_1,W_2)}{P(W_1)P(W_2)}$$

  上記の式を変換して、

$$I(W_1;W_2)=\log\frac{\frac{f(W_1,W_2)}{N}}{\frac{f(W_1)}{N}\frac{f(W_2)}{N}},\quad N:\text{総記事数}$$

  2つの単語で、関連度が強いものは I の値が大きくなり、関連度がないものほど 0 に近づく。

ニュース記事から学習した前述の指標を使って、図 2に示すように関連度の高いキーワード候補どうしをグルーピングする。この例は、6 個のキーワードの候補がありうる場合を示している。矢印で結んであるキーワードどうしが関連度の高いキーワードで、1 グループを形成している。ここでは 3 つのグループが作られているが、最もキーワードの数が多い G_1のグループを使って検索を行なう。

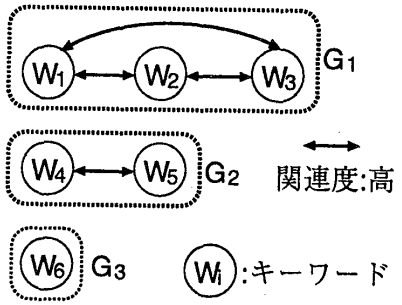

図 2: キーワード候補のグルーピング

図 3に実際の例を示す。これは、"公開"、"官庁"、"公務員"の 3 つのキーワードを入力したときの例である。

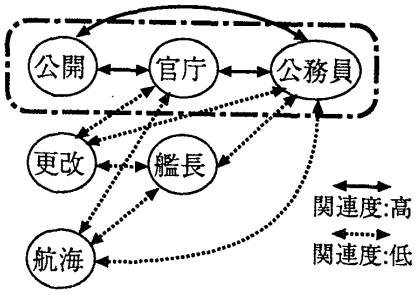

図 3: グルーピングの実例

5 おわりに

今回、ニュース音声データベースから、ニュース記事の検索システムを試作し、音声認識による書き起こしのデータベースを用いても検索能力が高いことを示した。

今後は、実際に音声でキーワードを入力し、前述した検索手法がどの程度有効であるか、また他の関連度の尺度についても検討したい。また、検索対象の記事を増やして実験を行なう予定である。

参考文献

[1] Dave Abberley,Steve Renals,Gary Cook: Retrieval of Broadcast News Documents with the THISL System, Proc.ICCASP,pp.3781–3784(1998.5)

[2] 福島, 赤峯:全文検索システム RetrievalExpress の開発と評価, 言語処理学会, 第 3 回年次大会,pp.361–364(1997.3)

[3] 赤松, 花井, 甲斐, 峯松, 中川:新聞・ニュース文をタスクとした大語彙連続音声認識システムの評価, 情報処理学会, 第 57 回全国大会, pp.35–36(1998.10)