音声言語情報処理 29－16
(1999. 12. 20)

# キーワードの音声入力によるニュース音声の検索法

西崎 博光　中川 聖一

豊橋技術科学大学 情報工学系
〒 441-8580 豊橋市天伯町字雲雀ヶ丘 1-1
Tel. (0532)44-6777
E-mail: {nisizaki, nakagawa}@slp.tutics.tut.ac.jp

**あらまし** 近年、多くのニュース番組が放映されており、過去のニュース番組から興味のある記事を見つけたいという欲求が高まっている。

数多くのニュース番組の中から、必要なニュースを見つける場合、各ニュースに対してインデックスが付けられている場合はそれを使って検索することができる。しかし、ニュース音声データからの検索に対する需要もあり、この場合は放送された全てのニュースを予め文字化し、データベースとして蓄積しておく必要がある。この作業を人手で行なうのは不可能に近く、大語彙音声認識システムを用い、自動的に書き起こすこととなる。

本研究では、1) 自動的に書き起こしたデータベース (誤認識単語を含む) での検索性能を、テキスト入力のキーワードを用いて実験的に検討した。実験の結果、単語認識率が低いにもかかわらず、高い再現率を得ることができた。次に，2) 検索対象となる単語を音声で入力した際の問題点を挙げ、それに対する対処法を提案する。実際にキーワードを音声で入力し、提案した方法を使って実験を行ない、その有効性を示す。

**キーワード** 音声認識、情報検索、ニュース音声、キーワード、関連度

# A Retrieval Method of Broadcast News Using Voice input Keywords

Hiromitsu Nishizaki and Seiichi Nakagawa

Department of Information and Computer Sciences, Toyohashi University of Technology,
Tenpaku-cho, Toyohashi, 441-8580, Japan
Tel. (0532)44-6777
E-mail: {nisizaki, nakagawa}@slp.tutics.tut.ac.jp

**Abstract** To retrieve interesting broadcast news documents out of an enormous number of TV news programs, if no indexing is done on the news and word-based retrieval is required, it is inevitably necessary to transcribe all the broadcast news documents automatically and store them as a database. And this task can be done only by using a Large Vocabulary Continuous Speech Recognition(LVCSR) system.

In this paper, 1) the retrieval performance was experimentally compared between the system using automatically transcribed database(speech) and the one using manually transcribed database(text). This experiment was done using text as the input to the system. As a result, high recall was obtained although the word recognition rate was low. Next, 2) to solve the inevitable problems which arise when the input to the system is realized as speech, i.e. misrecognition, a novel method was developed. In experiments, we retrieved news documents through inputted voice keywords to the system by using the method described above and represent its effectiveness.

**key words** speech recognition, information retrieval, news speech, keywords, association degree

# 1 はじめに

近年、多くのニュース記事がインターネットを通じて世界中に配信され、また数多くのニュース番組が放映されている。過去のニュースから興味のある記事を見つけたいというユーザの欲求が高まっている [1][2]。現在では Web ブラウザを通して過去のニュースが検索できるが、テキストレベルの検索・閲覧しか出来ない。しかし、テキストでニュースを閲覧するよりも、テレビで放映された音声・動画のニュースの方が、明らかに見る側にとっては情報を摂取しやすいと考えられる。

ニュース音声の検索に関する研究は数多く行われており、さまざまな検索手法が提案されている。たとえば、Kenny らは単語単位のマッチングではなく、語彙サイズの増大という問題に着目し音素単位でのマッチングによる検索を行っている [3]。また、Robinson らは音声ドキュメントを検索する際、そのドキュメントと類似した別のコーパスを用いることで検索語の拡張を行い、音声ドキュメントを自動で書き起こした時の認識誤りに対してロバストな検索を行う方法を提案している [4]。

数多くのニュース番組の中から、必要なニュースを見つける場合、各ニュースに対してインデックスが付けられている場合はそれを使って検索することができる。RWC のように音声同士のマッチングで行う方法も考えられるが、処理量が大きくなる問題がある [5]。そこでニュース音声データからの検索に対する需要もあり、この場合は放送された全てのニュースを予め文字化し、データベースとして蓄積しておく必要がある。この作業を人手で行なうのは不可能に近く、大語彙音声認識システムを用い、自動的に書き起こすこととなる。

なお、本研究では検索タスクとしてニュース音声を対象としているが、音声しか存在しないデータベース (例えば、ニュース以外の番組等) の検索にも本研究の手法が適応できると考えている。

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 (以後、DB(音声) と記す)。このデータベースと正確に書き起こしたデータベース (以後、DB(テキスト) と記す) に対して、キーワード群を使って検索された記事の再現率を求め、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再現率が得られた [6]。

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ない。また、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、誤りもありうる認識結果を使って検索を行なわざるを得ない。また、キーワードに同音意義語が存在する場合は、同じ読みの単語すべてをキーワードとして扱う必要がある。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たり、逆に全く結果が出力されないことになるので、これらの記事をうまく絞り込んでいく必要がある。そこで検索処理に先立って、単語間の関連度を用い、キーワード候補の語数を絞る手法を提案した [6]。単語間の関連度は DB(テキスト)、または DB(音声) より学習し、キーワード候補をグルーピングする。その結果、キーワード候補は幾つかのグループに区分される。そして、単語数の最も多いグループ中の単語を用いて検索処理を行なう。検索用のキーワード候補が実際の入力数よりも増大するというのは、音声でキーワードを入力したときのみに起こる現象であり、検索前に必要なキーワードを選択するというこの手法は他に類を見ない。

本稿では実際にキーワードを音声で入力し、前述の手法で検索実験行ない、その有効性を示す。

# 2 ニュース音声検索システム

## 2.1 概要

今回作成した、ニュース検索システムの概略図を図 1 に示す。

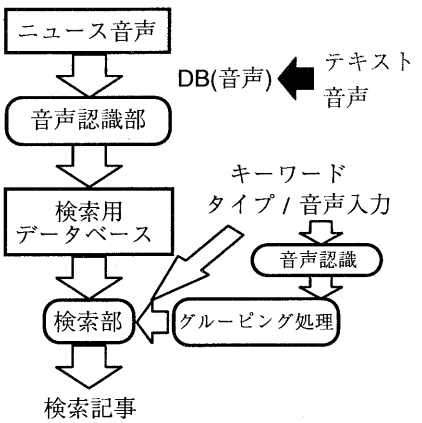


図 1: システムの概略

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。これを基に、入力キーワード (タイプ入力、音声入力) に応じて記事を検索部で検索する。音声入力の場合は、まず、グルーピングモジュールに通し、不必要なキーワード候補を取り除いたキーワードを検索部に入力する。

検索部では全文検索 [7] を行なっているが、インデックス法 [8] 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、これでは制約がきつ過ぎ必要な記事が検索されないので、入力キーワード数が多い場合は、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場合 (音声認識で使用した語彙辞書に入っていない) 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。

## 2.2 キーワードのグルーピング

キーワードの入力がテキストでなく、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、

1. 正解の単語に認識された
2. 正しい音節列ではあるが，異なる語 (同音意義語) として認識された
3. キーワードが違う単語として認識された (異なる音節列)

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、どの場合も得られた認識結果を使って検索処理を開始せざるを得ない。同音意義語が存在する場合は、全ての同音異義語を使って検索する必要があり、同音意義語がない場合でも、認識尤度の高い認識結果候補単語を複数個使って検索する必要性も考えられる。いずれにしても、発声単語数よりも多い単語セット (キーワード候補) を使って検索処理が行なわれるため、必要以上の記事が検索されたり、また逆に全く記事が検索されない恐れがある。こういった不具合を解決する方法として、キーワード間の関連度を用いたキーワードの絞り込み手法を提案する。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、相互情報量を用いる。

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 $W_2$を同じ記事で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。

$$I(W_1;W_2) = \log \frac{P(W_1,W_2)}{P(W_1)P(W_2)} \quad (1)$$

上記の式を変換して、

$$I(W_1;W_2) = \log \frac{\frac{f(W_1,W_2)}{N}}{\frac{f(W_1)}{N}\frac{f(W_2)}{N}} \quad (2)$$

$f(W_i)$: $W_i$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
$N$: 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。

関連度の学習は、DB(テキスト)、DB(音声) の両方から学習した (比較実験を行っている)。

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。ここで、N-best(単語列候補が順序づけられている) の下位に出てくる単語はやはり信頼性が低いと考え、下位の方に出てくる単語ほどペナルティを与えていく方が良いと考えられる。このペナルティには音声認識結果のスコア (尤度) を用いる。認識スコアを用いた時の 2 単語間 $W_1$,$W_2$の関連スコアの計算は、単純に認識スコアと相互情報量の値の重み付きの和で表す。つまり、

$$\alpha(L_1 + L_2) + MI \geq TH \quad (3)$$

で 2 単語間の関連スコアを計算する。スコアがある閾値 $TH$を越えた場合、単語 $W_1$,$W_2$をグループ化する。$L_1$ は単語 $W_1$の認識スコア、$L_2$は単語 $W_2$の認識スコア、$MI$は 2 単語間の相互情報量、$\alpha$は重みである。

図 2の例は、7 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同士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。なお、最もキーワード数の多いグループが複数出来た場合、もっとも関連スコアの良いものを選択する。

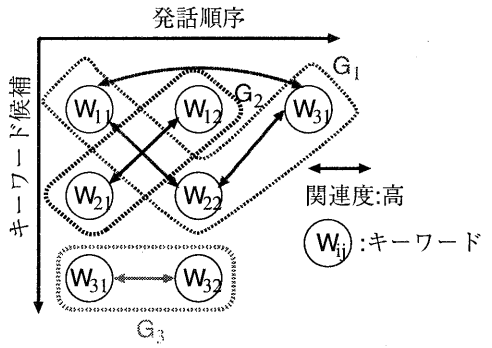


図 2: キーワード候補のグルーピング

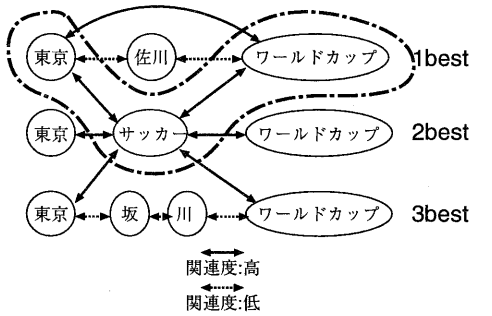


図 3: グルーピングの実例

図 3にグルーピングの実際の例を示す。これは、"東京"、"サッカー"、"ワールドカップ"の 3 つのキーワードを音声入力したときの認識結果の例で、1best～3best までを示してある。"東京"と"ワールドカップ"は 1,2,3best とも正しく認識されているが、"サッカー" については、2best でしか正しく認識されていない。この例では、"東京"_"サッカー"間、"サッカー"_"ワールドカップ"間、"東京"_"ワールドカップ"間で関連度が高いので、これら 3 つのキーワード候補を 1 つのグループとしこれを検索キーワード群として用いる。"東京"_"サッカー"_"ワールドカップ"という 3 つ組は他にも考えられるが、グルーピング時に音声認識スコアも考慮しているため、図 3のグループが採用されることになる。

## 2.3 検索方法

検索用のキーワード入力の違いにより次の 3 通りの方法で実験を行った。

1. タイプ入力

2. 音声入力 (N-best, グルーピングあり)

3. 音声入力 (N-best, グルーピングなし)

N-best とはキーワードの認識結果の N-best 仮説を用いる (実験では、1,3,5,10best) ことを表し、そのときグルーピング手法を使った場合とそうでない場合の実験を行った。

音声入力の場合、すべてのキーワード候補がマッチングする記事のみを検索してくるのは制約がきつすぎるため、大部分がマッチングする記事のみを検索することにする。そこで、入力キーワード数に対し、どれだけマッチングすれば良いとするかの閾値を決めないといけないが、これは記事の**検索率** (特定のキーワード群を音声で入力したとき、正解の記事が検索される確率) の期待値が $Th$ 以上になるように設定した (但し書き起こしが100%正しいと仮定した場合)。つまり、

$$\sum_{i=N}^{M} {}_M\mathrm{C}_i p^i (1-p)^{M-i} \geq Th \qquad (4)$$

$M$ : 入力キーワード数
$N$ : 入力キーワードのうち記事中に実在する数
$Th$ : 期待値の閾値

となるような N を求めた (実際には $Th$ をさまざまに変化させて実験を行った)。$p$ は入力キーワードが正しく音声認識される確率である (図 4参照)。

# 3 検索実験

## 3.1 実験条件

実験対象の音声データは、NHK ニュース (1996 年 6 月 1 日~7 月 14 日) で、記事の数は 976 記事、文数で 7099 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 パスでは単語 trigram を使用している。この音声認識システム [9] を使用した場合、ニュース音声 (バックグラウンドミュージック、紙をめくる音などの背景雑音などが混入されている) の単語カバー率は 96.7%、単語正解率は 54.3%、単語正解精度は 38.0%と非常に低くなった。名詞 (13710 種類) だけの認識率は 79.1%となった。

DB(テキスト) と、DB(音声) に対して、50 組のキーワード群 (1 組は 3~5 個のキーワードからなる) を使って検索し、次の再現率、適合率、F 値を求める。

**再現率 (*recall*)**：あるキーワード群のテキスト入力で DB(テキスト) に対して検索された記事を正解とした場合、同じキーワード群を音声を使って DB(テキスト) または DB(音声) に対して検索した場合に、どれだけ検索されたかを表す割合。

**適合率 (*precision*)**：検索されたすべての記事のうち、正解の記事数の割合で、余計な記事の湧きだしが多いほど小さくなる。

**F 値 (*F-measure*)**：一般的に情報検索の性能評価に用いられる指標。再現率と適合率を同時に評価できる。

$$F = \frac{(\beta^2+1)PR}{\beta^2 P + R} \qquad (5)$$

$P$: 適合率　$R$: 再現率

$\beta$: 適合率に対する重み。本実験では適合率と再現率を等価に扱うので$\beta = 1$ とした。

表 1: 認識実験の実験条件

| | |
|---|---|
| 音響モデル | |
| 5 状態 4 出力分布 (4 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 |
| 学習データ | |
| ASJ ATR503 文 A~J セットの 6 名の男性話者と 216 単語の音声データから初期モデルを作成 | |
| ASJ ATR503 文 A~J セットの 30 名の男性話者と JNAS 新聞記事文 125 名の男性話者を MAP 推定で追加学習 (総発話数 17221 文) | |
| 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 | |

## 3.2 キーワードのタイプ入力による実験結果

キーワードを DB(音声) に対してタイプ入力したときの実験結果を表 2に示す。

再現率で 70.2%、適合率で 52.8%であった。再現率と検索率は本来同等の精度になりうるはずだが、差が生じたのは今回の検索対象記事の書き起こし文の正解率が良かったためと考えられる。また、適合率の値により、検索された記事のうち約半分が余計な記事の湧き出しということになる。

表 2: DB(音声) に対する実験結果 (タイプ入力)

| | | |
|---|---|---|
| 検索対象記事 | : | 50 |
| 再現率 | : | 70.2% |
| 適合率 | : | 52.8% |

表 2の実験結果を見ると、単語正解率 54.3%、単語正解精度 38.0% とかなり低い値になっているにもかかわず再現率が比較的高くなっている。これは、評価実験でタイプ入力したキーワードの書き起こしデータベース中の認識率が全体の認識率よりも高く 93.0%になっている

ためである。音声認識を使って書き起こしたデータベースを用いると、3割程度性能が低下してしまったが、全体の音声認識率は検索性能にそんなに影響しないということが言える。これは、文献 [1] の結果と符合する。

### 3.3 キーワード音声入力による検索実験

#### (a) キーワードの音声認識

3名の話者にキーワード (全部で 175 キーワード) を発話してもらい、認識実験を行なった。

キーワードは連続して発声される場合があり、また複合名詞になっているものも多いので、キーワードの認識にはニュース音声の書き起こし時と同じ大語彙連続音声認識システム (語彙サイズは 20000 単語) を使用した。ただし、以下のように言語モデルをキーワード認識用に変更している。

- 名詞→ストップワード、ストップワード→名詞の接続確率を大幅に低く設定する。(ストップワードとは、キーワードにはなり得ない単語のこと)
- 名詞→名詞の接続確率がある一定値を越えていればその確率値を使用し、一定値に満たない場合は一定値を接続確率に設定する (実験では一定値として $10^{-4}$ を使用)。

この変更により複合名詞を発声した場合は bigram の確率が使われるし、孤立単語の発声やキーワード間に文法的な接続関連がない場合でも unigram が適用されるようになる。

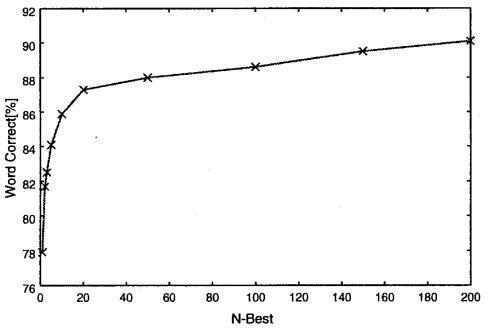


図 4: キーワードの認識率と N-best 仮説の関係

キーワードの音声認識結果を図 4に示す。図 4は単語正解率 (3 人の話者の平均) と N-best 候補 (N 個の単語列候補) の関係を示している。より大きい N-best 候補までを使用することでキーワードの認識率は上昇するのが明らかである。この結果からも、検索時には N-best 候補を用いた方がよさそうであると推測される。

#### (b) 検索実験

キーワードの音声認識結果を入力キーワードとして、検索実験を行なった。DB(テキスト) を使った場合と、DB(音声) を使った場合との検索実験で、検索結果にどれくらいの違いが現れるかを調べた。結果を図 5、図 6、表 3に示す。図 5が DB(テキスト) に対する実験結果、図 6が DB(音声) に対する実験結果である。表 3は、再現率と適合率を同時に評価できる $F$値で、最大の時の値のみを示している。それぞれの図、表はグルーピング処理を行った結果である。これらのグラフ・表より、1best のみの結果を用いるよりも、認識率の高い 3best までを使った方が良いことが分かる。しかし、N の値を大きくし過ぎると逆にパフォーマンスが悪くなっている。これは、キーワード候補数の増加が原因だと考えられる。このことから、キーワード認識の Nbest を使う方が良いが、N が大き過ぎると逆に悪くなるということが言える。また、グルーピングを行う際に、認識スコアによるペナルティを与えず、すべての N-best 候補を等価に扱った場合 ((3) 式の$\alpha$を 0 とした場合) と比べて、recall が 10%程度上昇した。

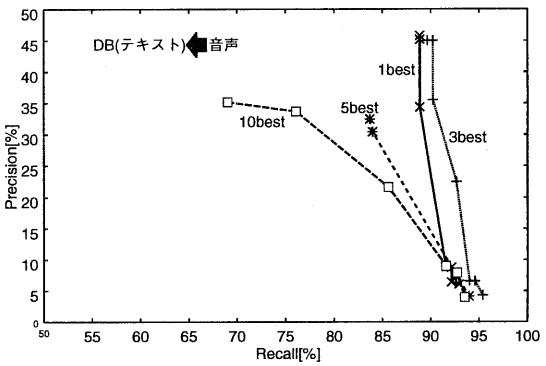


図 5: DB(テキスト) に対する実験結果

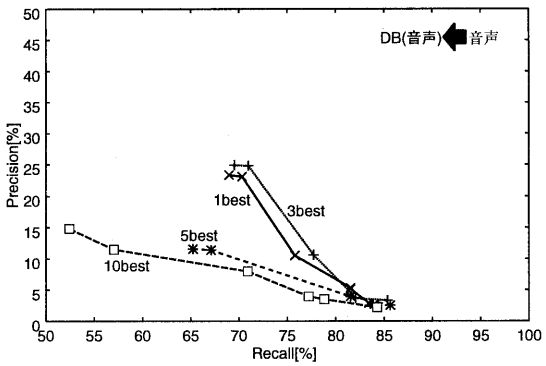


図 6: DB(音声) に対する実験結果

表 3: F-measure による評価

| | DB(テキスト) | DB(音声) |
|---|---|---|
| 1-best | 60.1 | 36.9 |
| 3-best | 61.4 | 38.7 |
| 5-best | 46.8 | 33.7 |
| 10-best | 46.6 | 29.7 |

グルーピングの有効性を調べるため、最もパフォーマンスが良かったN=3の場合について、グルーピングを用いた場合と、用いなかった場合との比較実験を行った。結果のグラフを図7に示す。グルーピングを行った方が適合率が高くなっているが、再現率の点においてはグルーピングを行わない方が良い。グルーピング手法は不必要な記事を検索しないようにするという点では有効である。このグルーピング法はN=1に対しても有効であることを確かめている。

最後にグルーピングを行う際に用いるキーワード間の関連度の違いについて比較した。1つは、DB(テキスト)から学習した関連度で、もう一方がDB(音声)から学習した関連度である。グラフを図8に示す。結果を見ると、ほとんど差がないことが分かる。このことから、DB(音声)から学習した関連度を用いても問題はないと考えられる。

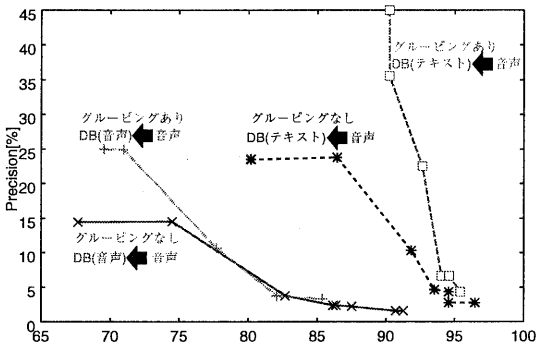


図 7: グルーピングあり・なしの比較 (3-best)

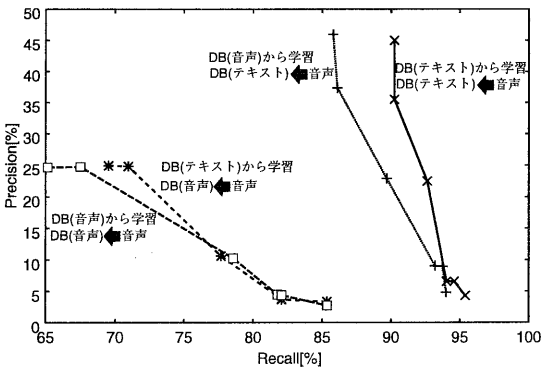


図 8: トレーニングデータの違いによる実験結果 (3-best)

## 4 むすび

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作成し、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。また、キーワードが音声入力の場合に考えられるキーワード候補の増大に対処する方法を提案した。実際に音声入力による実験では、提案したグルーピング手法を用いキーワード候補を絞り込むことで、余計な記事の湧き出しを押さえることができる、キーワード認識のNbestを使う方が良いが、Nが大きすぎると逆に悪くなるということがわかった。

本稿では、検索を行なう前にキーワード候補を絞り込んだが、その検索結果をさらに絞り込む方法として、音響的類似性などを使った方法を試みたい。また、同義語 (例えば、首相⇔総理大臣) や固有名詞の取り扱い方も検討していく必要がある。

## 謝辞

この研究では、NHK放送技術研究所のニュース音声データベース、ニューステキストデータベースを使わせていただいた。これらのデータベースを提供されたNHK放送技術研究所の関係諸氏に深く感謝します。

## 参考文献

[1] A.G.Hauptmann, H.D.Wactlar: Indexing and Search of Multimodal Information, Proc.ICASSP, pp.195–198(1997)

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

[3] Kenney NG: Towards Robust Methods for Spoken Document Retrieval, Proc.ICSLP, pp.939–942(1998.12)

[4] Tony Robinson,Dave Abberley,David Kirby,Steve Renals: Recognition, Indexing and Retrieval of British Broadcast News with the THISL System, Proc.EuroSpeech'99, pp.1267–1270(1999.9)

[5] 遠藤隆, 中沢正幸, 高橋裕信, 岡隆一: 音声と動画像の自己組織化ネットワークによるデータ表現とスポッティング相互検索, 人工知能学会, 人工知能学会全国大会 (第12回) 論文集, S5-04,pp.122–125(1998.6)

[6] 西崎博光, 中川聖一: 音声入力によるニュース音声検索システム, 情報処理学会, 音声言語情報処理研究会,SLP-26-3,PP.17–22(1999.5)

[7] 長尾 真編:自然言語処理, 岩波書店 (1996)

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

[9] 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)