音声キーワードによるニュース音声データベース検索手法

西崎博光† 中川聖一†

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，インデックスが付与されていない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すべてのニュースをあらかじめ文字化し，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，自動的に書き起こしたデータベース（誤認識単語を含む，数種類の認識率）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，検索対象となる単語を音声で入力した際の問題点をあげ，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い，その有効性を示す．

A Retrieval Method of Broadcast News Documents in Speech Database via Voice Input Keywords

Hiromitsu Nishizaki† and Seiichi Nakagawa†

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, the retrieval performance was experimentally compared between the system using automatically transcribed database (speech) and the one using manually transcribed database (text). Firstly, the experiment was done using text as the keyword input to the system. As a result, high recall was obtained in spite of low word recognition rate. Next, to solve the inevitable problems which arise when the keyword 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 and represent its effectiveness.

1. はじめに

現在の高度情報化社会により，多くの情報が様々なメディアを介して世界中に配信されている．特に，ニュース記事や各種データベースなど（図書蔵書など）の文書情報は，日々，大量に作り出されている．大量にありすぎて，読みきれない，読みたくても見つからない，整理できない，必要なときに取り出して再利用できないなど，電子化文書というものは思った以上に整理しがたい．このため，こういった情報の検索技術の開発が必要となってくる[1)]．

† 豊橋技術科学大学工学部情報工学系
Department of Information and Computer Sciences, Faculty of Engineering, Toyohashi University of Technology

最近では，文書情報に関する検索技術，および検索システムの開発により，インターネットを使ってどこからでも検索できるようになってきた．しかし，情報には文書情報（テキスト形式）だけでなく，テレビ・ラジオ放送など画像・音声の情報も数多く存在する．こういったマルチメディア情報の検索技術も将来的には必要不可欠になることと思われる．本研究ではその一例としてテレビニュース音声に焦点を当て，これらの検索について考察を行った．

近年では，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．その配信，または放送された過去のニュースの中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[2)]．現在ではインターネットで Web ブラウザを通して過去のニュース記事が

検索できるが[3),4)]，検索された記事はテキストだけであったり静止画像だけのみしか掲載されてなかったりすることが多い．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，様々な検索手法が提案されている．特に，米国規格協会（National Institute of Standards & Technology，NIST）が1992年から開催している，情報検索システム評価会（Text REtrieval Conference，TREC）のプロジェクトの1つにSDR（Spoken Document Retrieval）プロジェクトがあり，音声のドキュメントの検索に関する研究がさかんに行われてている[5)~9)]．しかし，日本での音声ドキュメントの検索に関する研究は，まだ少ない．研究例としては，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[10)]．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とき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[11),12)]．Hauptmannらや Jourlinらや Renalsらは，音声ドキュメントを書き起こしたときの認識率を様々に変化させ，検索パフォーマンスに与える認識率の影響を調べている[13),14)]．日本語では，鷲尾らが検索語と記事間の類似度として相互情報量，TF-IDF法などを使ったニュース記事の検索実験結果を報告している[15)]．ニュース検索システムとしては，Choiらや Kempらの研究例があげられ，ユーザに使いやすいインタフェースについて研究がなされている[16),17)]．

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．遠藤らのように音声特徴パラメータ時系列どうし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[18)]．そこで大量のニュース音声データからの検索の場合は放送されたすべてのニュースをあらかじめ文字化し，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，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，数種類の異なるモデルを用いた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）．結果として数種類の認識率を持つデータベースを構築し，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[19),20)]．また，音声認識の単語正解精度よりも単語正解率が検索性能に影響してくることも実験的に実証できた．本論文で行った検索処理は索引語を使うのではなく，全文検索[21)]（検索キーワードが記事中に存在すれば，その記事を表示する）を行っている．

キーワードを音声で入力することを考えた場合，キーワードが必ずしも正しく認識されるとは限らない．違う単語や同音異義語に認識される可能性が十分にある．また，機械には認識結果が正しいキーワードかどうか分からないため，誤りもありうる認識結果を使って検索を行わざるをえない．また，キーワードのN-best認識仮説を使うことも考慮したり，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ったりする必要がある．こういった場合，実際にユーザが意図しない記事を大量に含む検索結果が得られたり，逆にまった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[22)]．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，キーワード候補はいく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ほかに類を見ない．本論文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い，その有効性を示す．

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（たとえば，ニュース以外の番組など）の検索にも本研究の手法が適用できる．

2. ニュース音声検索システム

2.1 概　　要

今回構築した，ニュース検索システムの概略図を**図1**に示す．

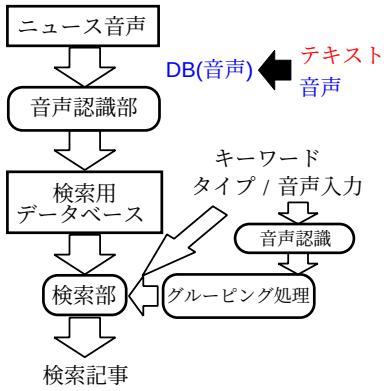


図 1 システムの概略
Fig. 1 Overview of the system.

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[21]を行っているが，インデックス法[23]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す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固有名詞など）は，音節列（かな文字列）単位の DP マッチングを行うことにより対処する[31]．また，検索部には角川新類語辞典[24]を使用した同義語処理を組み込んでいる[25]．

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とき，

(1) 正解の単語に認識される，
(2) 正しい音節列ではあるが，異なる語（同音異議語）として認識される，
(3) キーワードが違う単語として認識される（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分からないので，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えない．同音異議語が存在する場合は，すべての同音異義語を使って検索する必要があり，同音異議語がない場合でも，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われるため，必要以上の記事が検索されたり，また逆にまったく記事が検索されなかったりする恐れがある．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[26]およびキーワードの音声認識結果のスコア（尤度）を用いた．相互情報量のほかにも，ある記事においてある 2 単語が同時に同じ記事に出現しやすいかという共起頻度を利用した関連度も使用してみたが，結果的に相互情報量の方が勝っていた[19]．相互情報量を用いた関連研究としては，クロスリンガル情報検索において検索語をターゲット言語に翻訳する際に相互情報量を用いて多語義の曖昧性を解消する方法[27]や，また機械翻訳の分野では単語間の意味的関連性を用いた語義の曖昧性解消方法[28]が提案されている．

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．

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

上記の式を変換して，

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

$f(W_i)$：W_i が出現した記事数（$i = 1, 2$）
$f(W_1,W_2)$：W_1，W_2 がともに出現した記事数
N：総記事数

2 つの単語で，関連性が強いものは I の値が大きくなり，関連性がないものほど 0 に近づく．評価実験では，相互情報量は，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

☆ 本論文で述べるキーワードとは形態素レベルの単語である．すなわち，阪神大震災だと阪神と大震災の 2 つのキーワードとなる．

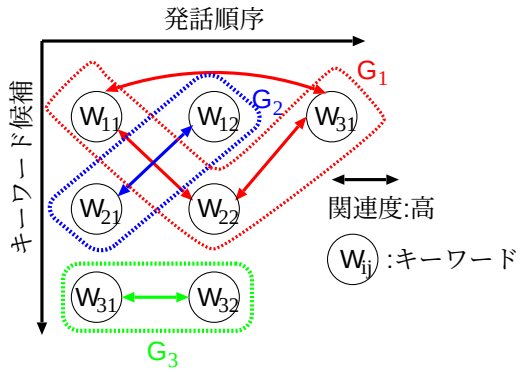


図 2 キーワード候補のグルーピング
Fig. 2 Grouping of keyword candidates.

ティを与えていく方がよいと考えられる．このペナルティには音声認識結果のスコアを用いる．認識スコアを用いたときの 2 単語間 W_1，W_2 の関連度の計算は，単純に認識スコアと相互情報量の値の重み付きの和 $L(W_1, W_2)$ で表す．つまり，

$$L(W_1, W_2) = \alpha(L_1 + L_2) + I \geq TH_1 \qquad (3)$$

で 2 単語間の関連度を計算する．関連度がある閾値 TH_1 を超えた場合，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，I は 2 単語間の相互情報量，α は重みである．

図 2 の例は，3 個のキーワード発話の認識結果のうち，最初の 2 個の単語についてそれぞれ 3 種類の単語候補が得られ，残りの 1 つは 1 種類の単語候補だけにしか認識されなかった場合で，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候補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，最もキーワード候補の数が多い G_1 のグループを使って検索を行う．なお，最もキーワード候補数の多いグループが複数できた場合，最も関連度の良いものを選択する．

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，単語列としての 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 つ組はほかにも考えられるが，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 において一点鎖線で囲まれた組合せのグループが採用されることになる．

☆ 我々の認識システムでは，単語列として最適な N 候補を出力している．

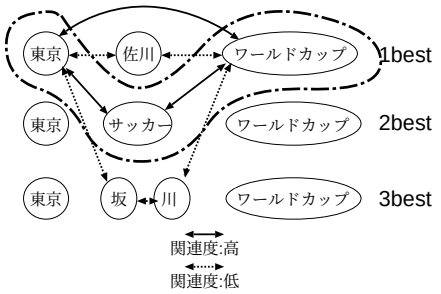


図 3 グルーピングの実例
Fig. 3 Example of the grouping.

2.3 検索方法

全文検索手法を用いた文書検索では，通常は入力キーワードのすべてと一致する単語もしくは文字列が同一の文書中に含まれていないと，その文書は検索されない．したがって，音声データベースの書き起こし文書に対してキーワードを音声で入力する場合は，音声認識システムによりキーワードもしくはキーワードになりうる単語が正しく認識されているとは限らないため，記事が検索されにくくなると予想される．

そこで DB（音声）からの検索やキーワードが音声入力の場合，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，M 個の入力キーワードのうち N 個以上存在する記事を検索する．そこで，M と N の関係を次のように考察した．

(1) DB（音声）中に，キーワード（M 種類中 N 種類）が正しく認識されている確率．タイプ入力の場合が対象．
(2) キーワードの音声入力の音声認識が（M 個中 N 個）正しく行われている確率．音声入力の場合が対象．

の 2 つの場合を考える．まず (1) の場合，同じキーワードが 1 つの記事中に x 回現れるとすると，この x 個のうち少なくとも 1 つが認識される確率（1 つでもキーワードが認識されていれば記事の検索条件を満たすことになる）は，

$$P_x = 1 - (1 - p)^x \qquad (4)$$

で計算される．p はデータベース中のキーワードの認

識率である．たとえば，データベース中のキーワードの認識率が $p = 0.94$☆で同じ記事に 2 回現れるとするならば，式 (4) より $P_2 = 0.9964$ となり，ほぼ確実に認識されることになる．ここで，入力キーワード数に対し，どれだけマッチングすればよ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入力したとき，正解の記事が検索される確率）の期待値 E_1 が TH_2 以上になるように設定した．つまり，

$$E_1 = \sum_{i=N}^{M} {}_M\mathrm{C}_i P_x^i (1 - P_x)^{M-i} \geq TH_2 \qquad (5)$$

M：入力キーワード数

N：入力キーワードのうち書き起こし記事中に実在する数

TH_2：期待値の閾値

となるような N を求めた（実際には TH_2 を様々に変化させて実験を行った）．

(2) の場合についての検索率の期待値も式 (5) 同様に求めることができる．ただし書き起こしが 100%正しいと仮定した場合の検索率の期待値であるので，実際はこれよりも悪くなる．

$$E_2 = \sum_{i=N}^{M} {}_M\mathrm{C}_i p_k^i (1 - p_k)^{M-i} \geq TH_3 \qquad (6)$$

M：入力キーワード数

N：入力キーワードのうち正しく認識される数

TH_3：期待値の閾値

p_k は入力キーワードが正しく音声認識される確率である（音声入力したときのキーワードの認識率は後述の図 5 を参照．たとえば 1-best のときは $p_k = 0.81$，3-best のときは $p_k = 0.85$）．

これらの式はあくまでも近似式である．たとえば，認識誤りしやすい単語は，何回出てきても誤認識することが多いが，このことは式 (4) には反映されていない．そうでなければ，タイプによるキーワード入力では検索率は 100%近くになると期待されることになる．**表 1** に実際の音声認識結果（$p = 0.94$，$p_k = 0.81$ または 0.85）における M と N の関係を示す．

☆ 表 2 より，音声データベースを書き起こしたときの，検索実験で使用するキーワードの認識率が一番良いときで約 94%．名詞（13,710 種類）の中でキーワードになりうる名詞の方が認識率が高い．

表 1 閾値の設定（$TH_2 = TH_3 = 0.95$）．括弧内数値は検索率の期待値

Table 1 Setting the threshold ($TH_2 = TH_3 = 0.95$). The value in a parenthesis denotes the accuracy of plausible documents to be retrieved.

(a) タイプ入力 (1) の場合

M	N		
	$P_1 = 0.94$	$P_{1.5} = 0.985$	$P_2 = 0.996$
3	2 (0.990)	3 (0.985)	3 (0.982)
4	3 (0.980)	3 (0.999)	4 (0.976)
5	4 (0.968)	4 (0.998)	5 (0.970)
6	4 (0.954)	5 (0.996)	6 (0.965)
7	5 (0.994)	6 (0.996)	7 (0.994)
8	6 (0.990)	7 (0.994)	8 (0.953)

(b) 音声入力 (2) の場合

M	N	
	1-best ($p_k = 0.81$)	3-best ($p_k = 0.85$)
3	2 (0.993)	2 (0.997)
4	2 (0.976)	2 (0.988)
5	2 (0.994)	3 (0.973)
6	3 (0.986)	4 (0.953)
7	4 (0.972)	4 (0.988)
8	5 (0.952)	5 (0.978)

3. 検索実験

3.1 データベース

実験対象の音声データは，NHK のニュース「ニュース 7」と「おはよう日本」（1996 年 6 月 1 日〜7 月 14 日）で，記事の数は 976 記事，文数で 7,099 文である．1 文あたりの単語数は 33 単語となっている．約半分の発話にはノイズ（バックグラウンドミュージック，紙をめくる音など）が混入されている．発声者の内訳としてはアナウンサ 6 人とその他多数のレポータが含まれている．音声データはすべて番組，記事ごとに分類されているので，記事ごとの検索データベースの構築は容易である．ニュース音声の書き起こしには，京都大学で開発された **JULIUS** [29]（16 KHz サンプリング，フレーム周期 10 ms，特徴ベクトルは MFCC，triphone モデル）と我々の研究室で開発した **SPOJUS** [30]（12 KHz サンプリング，フレーム周期 8 ms，特徴ベクトルは LPC メルケプストラム，音節モデル）の 2 種類の大語彙連続音声認識システムを用いた．**表 2** に DB（音声）の認識率を示す☆☆．デコーダ&言語モデルの欄で，**J** は JULIUS，**S** は SPOJUS を表し，**mai** は毎日新聞から学習した言語モデル，**NHK** は NHK 汎用ニュース原稿から学習した言語モデル，**bi**，**tri** はそれぞれ bigram，trigram，最

☆☆ 最新の結果では，JULIUS で単語正解精度 56.3%，単語正解率 66.1%，SPOJUS で単語正解精度で 63.3%，単語正解率 72.1%である[31]．

表 2 音声認識結果 [%]（カバー率：毎日 96.0[%]，NHK：96.7[%]）
Table 2 Recognition rate [%] (coverage: Mainichi 96.0[%], NHK: 96.7[%]).

	デコーダ & 言語モデル	単語正解率	単語正解精度	脱落率	挿入率	名詞の認識率	キーワードの認識率
（A）	J-mai-bi-20k	51.3	46.3	11.7	5.0	75.4	85.4
（B）	J-mai-tri-20k	58.6	51.6	7.7	7.0	82.2	93.4
（C）	J-NHK-bi-20k	55.8	49.3	11.7	6.5	81.2	93.0
（D）	J-NHK-tri-20k	62.0	51.9	9.2	10.1	83.9	94.2
（E）	J-NHK-tri-20k-ins	61.7	41.9	7.7	19.8	84.0	94.2
（F）	J-NHK-tri-20k-del	49.5	44.9	20.7	4.6	75.5	86.0
（G）	S-NHK-tri-20k	54.5	38.6	13.3	15.9	79.8	93.0

単語正解率 = 100 − 置換誤り率 − 脱落誤り率 [%]
単語正解精度 = 100 − 置換誤り率 − 脱落誤り率 − 挿入誤り率 [%]

後の数値は語彙サイズである．（E）はデコーダのパラメータを変更して強制的に挿入誤りを増やした結果で，同じく（F）は脱落誤りを増やしたものである．名詞（13,710 種類）の認識率は，名詞だけに注目した認識率であり，キーワードの認識率は次で述べる検索実験で使ったキーワード単語（175 種類）だけに関する認識率である．

キーワードの認識率が名詞の平均認識率よりも高くなったのは，複合語（複数の形態素からなる単語列，たとえば，3 キーワードからなる『地下鉄サリン事件』など）がキーワードとして多く使われたからだと考えられる．これらの単語の接続は，音声認識に用いる言語モデルで高い確率値を持っているため認識しやすくなっている．また名詞全般（13,710 語）には 1〜2 音節単語のような認識の困難な単語が多く存在しているが☆，キーワードは比較的長い単語が多いことも理由である．

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（検索漏れ），精度（検索ノイズ），F 値を求める．50 組のキーワード群は，各ニュース記事につけられているタイトル文から，被験者 6 人にキーワードを選択してもらい，6 人中 4 人が一致しているものをキーワードとした．6 人中 4 人というのは，6 人中 6，5 人であると，すべて一致する単語がほとんどなかったからである．また，50 の対象記事は無作為に選んだものである．

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．つまり，

$$\text{再現率} = \frac{\text{検索された記事のうち正解の記事数}}{\text{データベース中のすべての正解記事数}} \quad (7)$$

精度（*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．

$$\text{精度} = \frac{\text{検索された記事のうち正解の記事数}}{\text{検索された記事の数}} \quad (8)$$

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．再現率と精度を同時に評価できる[32]．

$$F = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad (9)$$

P：精度　R：再現率

β：精度に対する重み．本実験では精度と再現率を等価に扱うので $\beta = 1$ とした．

3.2 実験方法

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．

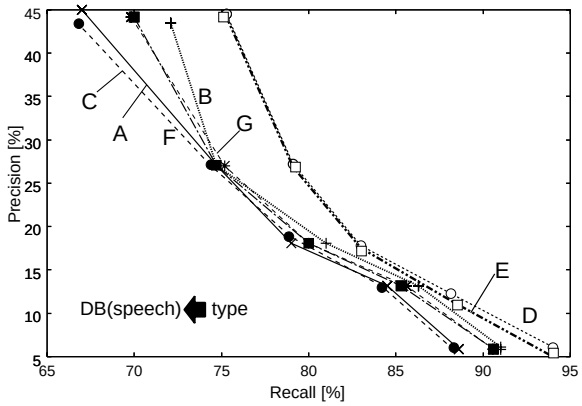
（1） タイプ入力
（2） 音声入力（N-best，グルーピングあり）
（3） 音声入力（N-best，グルーピングなし）

タイプ入力とは，DB（テキスト）に対して正解のキーワードをそのまま入力する実験であり，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．

3.3 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**図 4** に示す．式 (5) において TH_2 を様々に変化させると，M に対する N の値が変化するので（つまり入力キーワード M 個のうち N 個が含

☆ 13,710 語中，1 音節単語の割合が 1.3%，2 音節単語の割合が 10.0%.



F 値の最大値（recall/precision）
A： 53.8（67.3/45.0）
B： 54.3（72.3/43.5）
C： 54.1（70.9/44.0）
D： **55.1（75.2/44.1）**
E： **55.1（75.0/44.0）**
F： 53.6（66.9/43.5）
G： 54.2（70.6/44.0）

図 4 タイプ入力の場合の様々な DB（音声）に対する実験結果
Fig. 4 Result for various DBs (speech) using type-input keywords.

まれている記事を検索してくる），このときの再現率と精度，F 値の値をプロットしている．図中の（A）〜（G）は表 2 の（A）〜（G）と対応している．単語正解精度（Accuracy）というよりは単語正解率（Correct）が高くなると，それに対応して再現率も上がる傾向にある．わざと挿入誤りを増やした（E）のデータベースと同程度の正解率の（D）とを比較してみると，再現率はほとんど変わらず，精度も若干値が下がっているが大幅に下降したとはいえない．また，わざと脱落誤りを増やし単語正解率を落とした（F）の場合は，正解精度が（E）よりも高いのにもかかわらず再現率が大幅に低下した．以上の実験結果をまとめると，さほど高くない認識率でも再現率が比較的高くなっており，これはキーワードの認識率が高いためである．また，同程度の単語正解率であれば，単語正解精度が 10%も違っていても検索パフォーマンスにはほとんど影響しないが，正解率が変わると，それにともなって性能が変化する．検索パフォーマンスに影響するのは，データベースの単語正解精度ではなく単語正解率であり，音声認識の挿入誤りは検索に影響がほとんどないことが実験的に分かった．

3.4 キーワード音声入力による検索実験

3.4.1 キーワードの音声認識

3 人の男性話者にキーワード群 50 組（全部で 175 キーワード）を発話してもらい，認識実験を行った．

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 20,000 単語）を使用した．これにより，キーワードを連続して発声することができる．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．

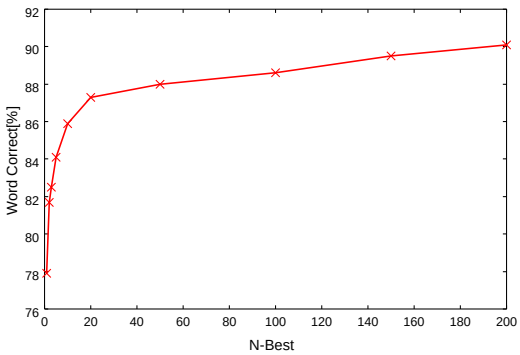


図 5 キーワード発話の認識結果
Fig. 5 Keyword recognition result.

- 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えない単語のこと）．
- 名詞→名詞の接続確率がある一定値を超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．

検索用のキーワード発声の認識の困難度は，20,000 語の孤立単語の認識と同定度と考えられる．実際のキーワードの音声認識結果を**図 5** に示す．図 5 は単語正解率（3 人の男性話者の平均）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．

3.4.2 グルーピング実験

図 6 は重み α（キーワードの音声認識尤度の重み）を変化させたとき，音声認識で得られたキーワード候補のグルーピングにより入力したキーワードを推定した結果を再現率と適合率で評価したものである．この図は，認識仮説の 3-best の場合である．$\alpha = 0$ の点がグルーピングに使用する関連度（式 (3)）において，相互情報量の値だけを使用した場合の結果である．α を変化させる，つまりキーワードの音声認識尤度の重みを高くすると，再現率と適合率が変化している．この図 6 では，F 値（再現率，適合率を同時に評価できる指標，式 (3) 参照）が最大になっていることから（F 値 = 73.7）$\alpha = 0.015$ で最も良い結果になっていることが分かる．つまり，相互情報量だけを関連度として使用するよりも，音声認識尤度も使用した方がグルーピングの能力が高くなることがいえる．実験では，最適な α を用いて実験を行った（次節参照）．

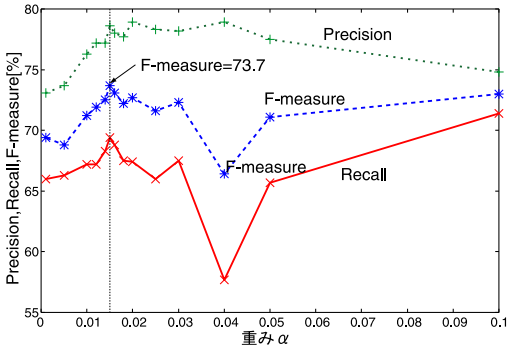


図 6 重み α とグルーピング能力の関係（3-best）
Fig. 6 Relationship between weight α and grouping ability.

3.4.3 検索実験

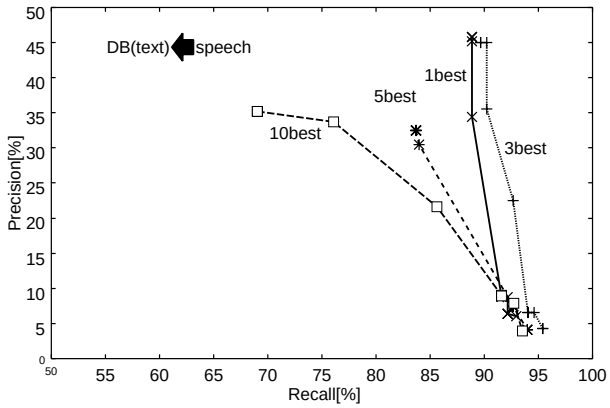
(a) N-best の効果

キーワードの音声認識結果を入力として，まずは DB（テキスト）に対する検索実験を行った．実験結果を**図 7** に示す．音声入力の場合は，式 (6) の TH_3 を変化させて値をプロットしている．この図は，キーワードの認識結果の N-best の値をいろいろと変化させた場合の図であり，前述のグルーピング処理を行っている．このグラフより，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 の値を大きくしす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 N-best を使う方がよいが，N が大きすぎると逆に悪くなるということがいえる．また，図には示していないが，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（式 (3) の α を 0 とした場合）と比べて，再現率が 10%程度上昇し，尤度の利用の効果を確かめている[19),22)]．

(b) 認識率と検索性能の比較

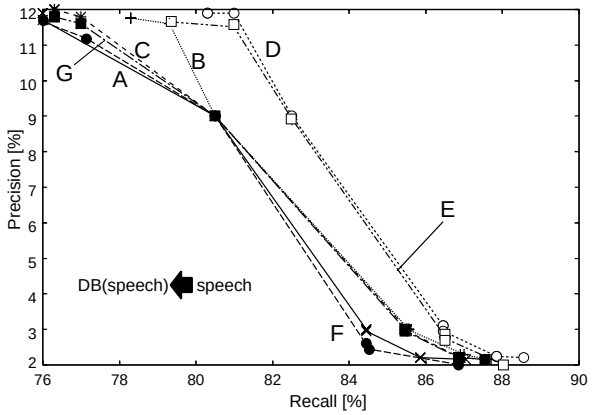
続いて，同じくキーワードの音声認識結果を入力キーワードとして，3.3 節と同様に認識率の違う複数の DB（音声）に対する検索実験を行った．実験結果を**図 8** に示す．この図は，キーワードの認識結果 3-best までを用いてグルーピングした結果である．音声入力のキーワードを用いているため，パフォーマンスは図 4 よりは落ちているが，全体の傾向は変わっていない．

また，DB（音声）（D のデータベースを使用）に対して，図 7 と同様の実験を行ったのが**図 9** である．こ



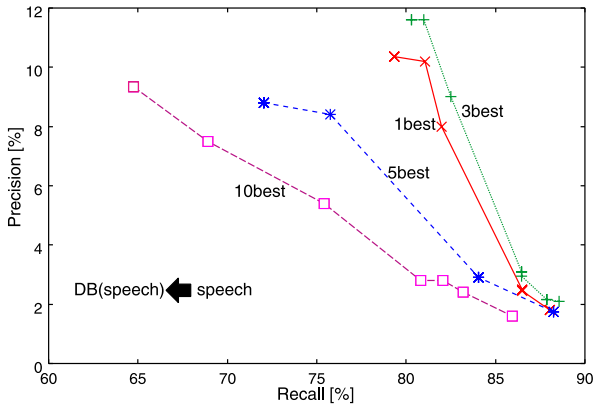
F 値の最大値（recall/precision）
1best： 60.1（88.9/45.5）
3best： **61.4（90.5/44.9）**
5best： 46.8（83.7/32.5）
10best： 46.6（75.5/33.8）

図 7 音声入力の場合の DB（テキスト）に対する実験結果
Fig. 7 Result for DB (text) using voice-input keywords.



F 値の最大値（recall/precision）
A： 20.1（76.0/11.6）
B： 20.3（78.2/11.7）
C： 20.2（76.2/12.0）
D： **20.7（81.9/11.9）**
E： 20.5（81.0/11.7）
F： 20.1（76.0/11.6）
G： 20.2（76.2/11.8）

図 8 音声入力の場合の様々な DB（音声）に対する実験結果
Fig. 8 Result for various DBs (speech) using voice-input keywords.



F 値の最大値（recall/precision）
1best： 19.1（81.1/10.3）
3best： **20.5（81.9/11.9）**
5best： 16.1（72.1/8.9）
10best： 15.8（64.8/9.2）

図 9 N-best 使用時の検索結果（DB（音声）には表 2 の（D）を使用）
Fig. 9 Result for the various N-best (for the transcript (D) shown in Table 2).

の図からもキーワードの認識結果の N-best を用いた方がよいことがいえる．

(c) グルーピングの有効性

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 3-best 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（全候補をキーワードとする）との比較実験を行った．結果のグラフを**図 10** に示す．グルーピングを行った方が精度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よい．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有効であり，グルーピングなしでは精度を上げることはできない．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[20]．

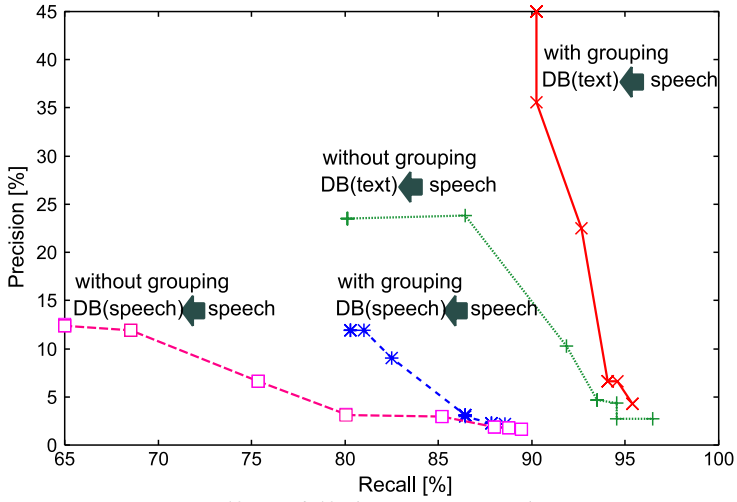
(d) トレーニングデータの違い

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 11** 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．

4. おわりに

本研究において，ニュース音声データベースからニュース記事を検索するシステムを構築した．音声データを大語彙連続音声認識システムにより書き起こしたデータベースに対して記事を検索することになるが，この認識率が検索性能に与える影響を，様々な音声認識率を持つデータベースを使って実験的に考察した．また，ニュース記事を音声入力のキーワードで検索する場合の音声認識誤りに対して頑健に対処する方法として，キーワード間の相互情報量と音響尤度を用いたグルーピング手法を提案した．

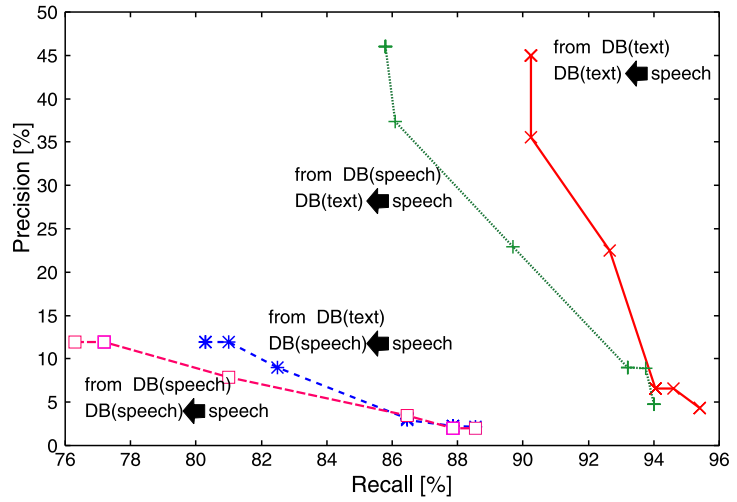
音声認識率の差の影響を見る実験では，検索性能は単語正解精度よりも正解率に影響されることを実証し



F 値の最大値（recall/precision）

DB（text）		DB（speech）	
with grouping	without grouping	with grouping	without grouping
61.4（90.5/44.9）	36.9（87.2/24.5）	20.8（81.9/11.9）	20.4（65.0/12.1）

図 10 グルーピングの効果（3-best，DB（音声）には表 2 の（D）を使用）
Fig. 10 Effect of the grouping method (with 3-best, for the transcript (D) shown in Table 2).



F 値の最大値（recall/precision）

データベース	DB（text）		DB（speech）	
学習データ	DB（text）	DB（speech）	DB（text）	DB（speech）
F-measure	61.4（90.5/44.9）	59.9（85.8/46.0）	20.8（81.9/11.9）	20.9（76.3/12.1）

図 11 トレーニングデータの違いによる実験結果（3-best，DB（音声）には表 2 の（D）を使用）
Fig. 11 Difference of training data for mutual information (with 3-best, for the transcript (D) shown in Table 2).

た．続いて，キーワードを音声入力で入れた場合では，提案したグルーピング手法を用いて，N-best 候補の中から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おさえることができる，キーワード認識の N-best 候補を使う方がよいが，N が大きすぎると逆に悪くなる，という知見が得られた．なお，相互情

報量を学習する際のトレーニングデータによる検索性能の違いはあまり見られなかった．

最後に今後の課題について述べる．本論文では，検索を行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検討する必要がある．また，認識辞書に登録されていない固有名詞の取り扱い方も検討していく必要がある[25]．たとえば，データベース中に含まれる未知語（固有名詞）は音節列で書き起こしておき，単語のマッチングの代わりに音節列どうしのマッチングを行う方法が考えられる[31),33),34)]．さらに対話を通して段階的に検索できるようになることが望ましい[35)]．

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) 藤田澄男：自然言語処理を利用した情報の検索・分類のアプローチ，情報処理，Vol.40, No.4, pp.352–357 (1999).
2) Abberley, D., Renals, S. and Cook, G.: Retrieval of Broadcast News Documents with the THISL System, *Proc. ICASSP'98*, pp.3781–3784 (1998).
3) http://www.asahi.com
4) http://www.mainichi.co.jp
5) http://www.itl.nist.gov/iaui/894.01/sdr99/sdr99.htm
6) Garofolo, J.S., Voorhees, E.M., Cedric G.P., Auzanne, V.M.S. and Lund, B.A.: 1998 TREC-7 Spoken Document Retrieval Track Overview and Results, *Proc. 7th TREC*, pp.115–119 (1997).
7) Johnson, S.E., Jourlin, P., Moore, G.L., Jones, K.S. and Woodland, P.C.: Spoken Document Retrieval for TREC-7 at Cambridge University, *Proc. 7th TREC*, pp.115–119 (1997).
8) Dharanipragada, S., Franz, M. and Roukos, S.: Audio-Indexing For Broadcast News, *Proc. 7th TREC*, pp.115–119 (1997).
9) Garofolo, J.S., Voorhees, E.M., Stanford, V.M. and Jones, K.S.: TREC-6 1997 Spoken Document Retrieval Track Overview and Result, *Proc. DARPA98*, pp.166–174 (1998).
10) Kenny, N.G.: Towards Robust Methods for Spoken Document Retrieval, *Proc. ICSLP'98*, pp.939–942 (1998).
11) Robinson, T., Abberley, D., Kirby, D. and Renals, S.: Recognition, Indexing and Retrieval of British Broadcast News with the THISL System, *Proc. EuroSpeech'99*, pp.1267–1270 (1999).
12) Renals, S., Abberley, D., Kirby, D. and Robinson, T.: Indexing and retrieval of broadcast news, *Speech Communication*, Vol.32, No.1-2, pp.5–10 (2000).
13) Hauptmann, A.G. and Wactlar, H.D.: Indexing and Search of Multimodal Information, *Proc. ICASSP'97*, pp.195–198 (1997).
14) Jourlin, P., Johnson, S.E., Jones, K.S. and Woodland, P.C.: Spoken document representations for probabilistic retrieval, *Speech Communication*, Vol.32, No.1-2, pp.21–36 (2000).
15) 鷲尾誠一，緒方 淳，有木康雄：ニュース音声に対する検索方法の比較，電子情報通信学会技術研究報告，SP99-109, pp.97–102 (1999).
16) Choi, J., Hindle, D., Hirschberg, J., Magrin-Chagnolleau, I., Nakatani, C., Pereira, F., Singhal, A. and Whittaker, S.: SCAN—Speech Content Based Audio Navigator: A Systems Overview, *Proc. ICSLP'98*, pp.2867–2870 (1998).
17) Kemp, T., Geutner, P., Schmidt, M., Tomaz, B., Weber, M., Westphal, M. and Waibel, A.: The Interactive Systems Labs VIEW4YOU Video Indexing System, *Proc. ICSLP'98*, pp.1639–1642 (1998).
18) 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，人工知能学会全国大会（第 12 回）論文集，pp.122–125, 人工知能学会 (1998).
19) 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，情報処理学会研究報告，99-SLP-26-3, pp.17–22 (1999).
20) Nishizaki, H. and Nakagawa, S.: A Retrieval System of Broadcast News Speech Documents through Keyboard and Voice, Lecture Notes in Computer Science, Vol.1692, pp.286–289, Springer (1999).
21) 長尾 真，佐藤理歴，黒橋禎夫，角田達彦：自然言語処理，岩波講座ソフトフェア科学，Vol.15, No.15, 岩波書店 (1996).
22) 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，電子情報通信学会技術研究報告，SP99-108, pp.91–96 (1999).
23) 福島俊一，赤峯 享：全文検索システム RetrievalExpress の開発と評価，第 3 回言語処理学会年次大会発表論文集，pp.361–364 (1997).
24) 大野 晋，浜西正人：角川類語新辞典（CD-ROM 版），角川書店 (1989).

25) Flank, S.: A Layered Approach to NLP-based Information Retrieval, *Proc. COLING-ACL'98*, pp.397–493 (1998).
26) 北　研二, 中村　哲, 永田昌明：音声言語処理—コーパスに基づくアプローチ，森北出版 (1996).
27) Jang, M.-G., Myaeng, S.H. and Park, S.Y.: Using Mutual Information to Resolve Query Translation Ambiguities and Query Term Weighting, *Proc. ACL'99*, pp.223–229 (1999).
28) 菊井玄一郎：ターム間の意味的関連性に基づくタームリストの翻訳多義解消，自然言語処理，Vol.7, No.3, pp.79–96 (2000).
29) 河原達也, 李　晃伸, 小林哲則, 武田一哉, 峯松信明, 嵯峨山茂樹, 伊藤克亘, 伊藤彰則, 山本幹雄, 山田　篤, 宇津呂武仁, 鹿野清宏：日本語ディクテーション基本ソフトウェア（99 年度版）の性能評価，情報処理学会研究報告，2000-SLP-31-1, pp.9–16 (2000).
30) 赤松裕隆，花井建豪，甲斐充彦，峯松信明，中川聖一：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，第 57 回情報処理学会全国大会講演論文集，pp.35–36 (1998).
31) 西崎博光，中川聖一：未知語を考慮したニュース音声記事の検索，電子情報通信学会技術研究報告，SP01 (2001.12).
32) Argamon, S., Dagan, I. and Krymolowski, Y.: A Memory-Based Approach to Learning Shallow Natural Language Pattern, *Proc. COLING-ACL'98*, pp.67–73 (1998).
33) Ng, C., Wilkinson, R. and Zobel, J.: Experiments in spoken document retrieval using phoneme n-grams, *Speech Communication*, Vol.32, No.1-2, pp.61–77 (2000).
34) Wang, H-M.: Experiments in syllable-based retrieval of broadcast news speech in Mandarin Chinese, *Speech Communication*, Vol.32, No.1-2, pp.49–60 (2000).
35) 小暮　悟，中川聖一：移植性の高いデータベース検索用音声対話システムの試作，情報処理学会研究報告，99-SLP-27, pp.99–104 (1999).

(平成 12 年 10 月 16 日受付)
(平成 13 年 10 月 16 日採録)

西崎　博光（学生会員）

昭和 50 年生．平成 10 年豊橋技術科学大学情報工学課程卒業．平成 12 年同大学大学院修士課程情報工学専攻修了．現在同大学院博士後期課程電子・情報工学専攻在学中．音声言語処理に関する研究に従事．電子情報通信学会，日本音響学会各会員．

中川　聖一（正会員）

昭和 23 年生．昭和 51 年京都大学大学院博士課程修了．同年京都大学情報工学科・助手．昭和 55 年豊橋技術科学大学情報工学系講師．平成 2 年教授．昭和 60〜61 年カーネギーメロン大学客員研究員．音声情報処理，自然言語処理，人工知能の研究に従事．工学博士．昭和 52 年電子通信学会論文賞，1998 年度 IETE 最優秀論文賞，平成 13 年電子情報通信学会論文賞受賞．著書「確率モデルによる音声認識」（電子情報通信学会編),「音声・聴覚と神経回路網モデル」（共著，オーム社),「情報理論の基礎と応用」（近代科学社),「パターン情報処理」（丸善）等．