自 然 言 語 処 理 149－13
音声言語情報処理 41－4
（２００２．５．２４）

文音声によるクエリーを用いた音声文書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
〒 441-8580 豊橋市天伯町字雲雀ヶ丘 1-1
E-mail:{nisizaki,nakagawa}@slp.ics.tut.ac.jp

本稿では，音声入力クエリーを扱う，日本語音声文書検索システムについて述べる．我々が以前に開発した，キーワードの孤立発声入力型の音声文書検索システムを改良し，自然な連続音声入力型の音声文書検索システムを試作した．そして，キーワード孤立発声型音声入力クエリーの場合と，自然文連続音声入力型クエリーの 2 種類について，検索性能の比較を行った．音声クエリーの場合は，これを音声認識システムで単語に認識させ，それを用いて検索を行うのが一般的な方法であるが，認識誤りという不可避な問題が起こる．これを解決する方法として，クエリーの認識仮説の N-best を利用し，そこから相互情報量に基づいたキーワード選択法により，検索に適したキーワード候補のみを抽出する．検索実験の結果，クエリーとして自然文連続音声入力型のクエリーを使った方（F 値で 73.4）が，キーワード孤立発声型音声入力クエリーを使った場合よりもわずかに性能が良くなった（F 値で 72.9）．

キーワード：音声文書検索，音声入力型 IR，相互情報量，キーワード候補選択

Development of a Spoken Document Retrieval System Accepting Spontaneous Spoken Queries

Hiromitsu NISHIZAKI and Seiichi NAKAGAWA

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

This paper describes a Japanese spoken document retrieval system that uses voice input queries. We deal with two types of spoken queries: isolately spoken keywords and spontaneously spoken queries. To solve a mis-recognition problem of spoken queries, N-best hypotheses of transcripts of queries are used, and keyword candidates are selected from them by mutual information between recognized words. Using spontaneously spoken queries instead of isolately spoken keywords, a retrieval performance in F-measure improves to 73.4 from 72.9, and using N-best hypotheses by recognizing queries, the performance also improves to 73.4 from 68.7.

Keywords: spoken document retireval, voice input IR, mutual information, keyword candidates selection

1 はじめに

近年，文書情報に関する技術の発展により，容易に必要な情報を得ることができるようになっている．しかし，テキスト文書だけでなく，画像・音声情報といった，マルチメディア情報も数多く存在しており，これらの情報を懲り強く検索する技術はまだまだ未熟であり，こららの検索技術も必要不可欠になることと思われる．音声文書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．研究例としては，Renals ら [2] は音声文書を検索する際，その音声文書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，音声文書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．Jourlin ら [3] は音声文書を書き起こしたときの認識率をさまざまに変化させ，検索性能に与える認識率の影響を調べている．これら以外にも英語の音声文書検索の研究は，近年まで開催されていた TREC SDR track[4] で多く発表されてきた．日本語では，鷲尾ら [5] 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．ニュース検索システムとしては，Choi ら [6] や Kemp ら [7] の研究例があげられ，ユーザに使いやすいインターフェースについて研究がなされている．しかし，ほとんどの音声文書検索研究では，テキスト入力の検索語のみを扱い，音声入力の検索語は扱っていないのが現状である．少ない研究例の一つである伊藤ら [8] のシステムでは，音声入力型テキスト情報検索において，大きな問題の一つである未知語問題に焦点をあて，検索語の音声認識部に未知語処理を導入することで，音声認識システムの語彙統制によってテキスト検索に悪影響を及ぼさない方法を提案している．また，今年度から NTCIR3 の Web 検索タスクにおいて，そのサブタスクとして音声入力型サブタスクが新設され，今後音声を扱った情報検索システムの研究がますます盛んになることが予想される．我々も以前ニュース音声データベースの検索を対象とした，キーワード孤立発声入力型音声文書検索システムを構築した [9]．そこでは音声入力の検索語を入力する際，音声認識結果の N-best 仮説から相互情報量に基づいてキーワードを推定し，利用する方法を提案し，有効性を示した．

そこで本稿では，キーワードの孤立発声入力型の音声文書検索を改良し，自然な連続音声入力型の音声文書検索システムを試作した．そして，キーワード孤立発声型音声入力クエリーの場合と，自然文連続音声入力型クエリーの 2 種類について，検索性能の比較を行った．検索実験の結果，クエリーとして自然文連続音声入力型のクエリーを使った方（F 値で 73.4）が，キーワード孤立発声型音声入力クエリーを使った場合よりもわずかに性能が良くなった（F 値で 72.9）．

2 ニュース音声検索システム

2.1 システムの概要

図 1に今回試作したシステムの概要図を示す．まず，大語彙連続音声認識システムを使って音声文書のディクテーションを行い，認識された各単語をインデックス化しておく．入力されたクエリー（キーワード孤立発声型音声入力クエリー，自然文連続音声入力型クエリーの両方）は，同じく大語彙連続音声認識システムに入力され，認識結果の N-best 仮説を出力する．この N-best 仮説内の単語から，ストップワード（今回の実験では名詞以外）を取り除き，単語同士の相互情報量に基づいて検索エンジンに入力すべきキーワードの候補を選択する．検索エンジンでは，そのキーワードを使って，作成しておいたインデックスに対して文書を検索してくる．

2.2 音声認識システム

音声文書の書き起こしに使った音声認識システムとしては，我々の研究室で開発した **SPOJUS** を用いる．音声認識システムで扱う特徴量，音響モデル，言語モデルをまとめたものを表 1に示す．音響モデルは，日本音響学会の音素バランス文からなる研究用連続音声認識データベース（ASJ-PB），および，新聞読み上げ音声コーパス（ASJ-JNAS）を用いて学習を行っている．言語モデルとしては，約 5 年分の NHK 汎用ニュース原稿から学習した語彙サイズ 2 万（検索対象の音声文書データベースに対するカバー率は 97.6%）の単語 bigram を用いた．

2.3 キーワードのグルーピング

音声によるクエリーの入力では，キーワード孤立発話であれ，自然文連続音声であれ，認識された時，

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たり（異なる音節列），脱落する
4. 発話した内容とまったく違う単語が湧き出る

という場合が考えられるが，機械には認識結果に正しいキーワードがすべて含まれている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．同音異議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，キーワード候補間の関連度を用いたキーワード候補選択手法を適用する（これを我々はグルーピングと呼んでいる）．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，相互情報量を用いた．

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．

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

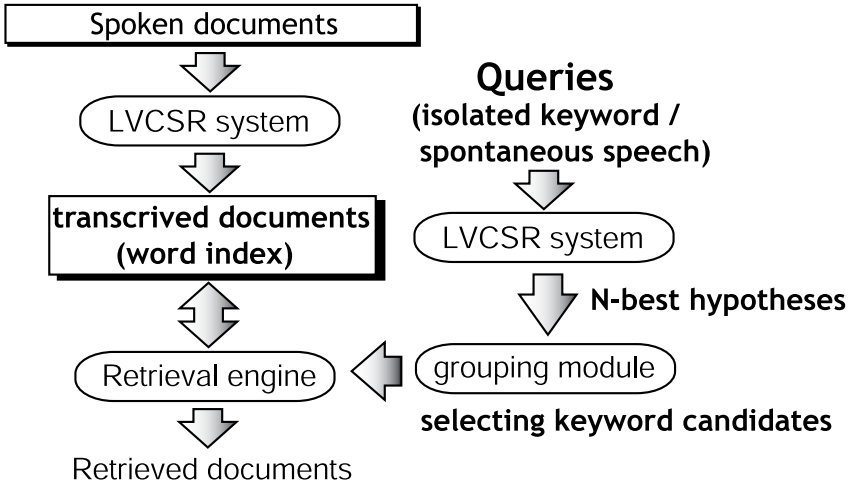

図 1: システムの概要

表 1: 大語彙連続音声認識のモデルの仕様

SPOJUS	
音響モデル	音節モデル（4 混合，全共分散，総状態数 1856） 5 状態 4 ループ，性別依存（男性），継続時間制御， 16kHz サンプリング， 25ms ハミング窓，フレーム周期 10ms
特徴ベクトル	MFCC ＋ Δ MFCC ＋Δ Δ MFCC ＋Δ POW ＋Δ Δ POW（計 38 次元）
言語モデル	NHK 汎用ニュース原稿（5 年分，約 12 万文） 語彙サイズ 20000，bigram，trigram モデル

上記の式を変換して，

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

$f(W_i)$: 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．検索実験では，相互情報量は音声文書データベースの書き起こしからら学習している．図 2にグルーピングの実際の例を示す．これは，”日本”，”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[1]．”日本”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，”日本”–”サッカー”間，”サッカー”–”ワールドカップ”間，”日本”–”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検索エンジンが検索に使用する．

[1]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

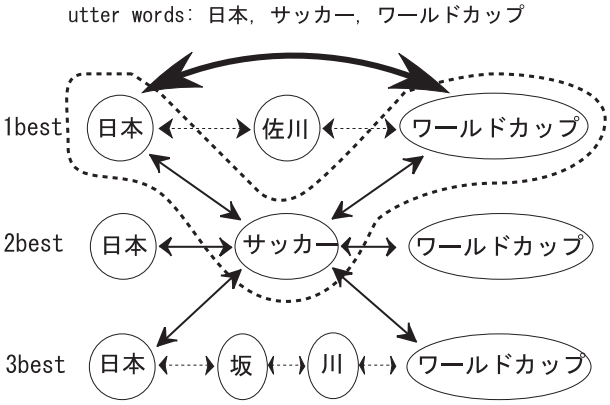

図 2: グルーピングの実例

2.4 検索エンジン

検索エンジンでは，グルーピングモジュールで選択したキーワード候補を使って文書を検索する．すべてのキーワード候補，もしくはキーワード候補の一部が含まれている文書を検索する．このようにこの検索エンジンは単純に単語同士のマッチングしか行わず，文書とクエリー間の類似度の計算などは一切行わない．なぜなら，我々の研究の目的は，さま

ざまな音声処理技術が音声文書検索の性能に与える影響を調べることであって，情報検索性能の絶対値を改善することが目的ではないからである．

3 音声文書検索評価

3.1 データベース

検索実験で使用する音声文書データベースはNHK ニュース音声データベース（1996 年 6 月 1 日〜7 月 14 日）である．このデータベースはあらかじめ記事内容毎に分割されており，その中から 967 記事（7099 文で約 18 時間分の音声）を検索対象とする．この音声文書データベースには，アナウンサーと記者の対話音声や，フィールドレポーターの音声，またバックグラウンドミュージックなどが含まれているため，音声認識が比較的難しい．SPOJUS を使って認識した結果，単語正解率で 73.3% であった（Julius[10] で標準的な音響モデルを使った場合では 62.0%[11]）．

3.2 キーワード 孤立発話と自然文音声クエリー

今回の実験では，クエリーがキーワードだけから構成されているものと，自然文の形式になっているクエリーの 2 種類の形式のクエリーを用意した．まず，キーワードだけから構成されている 50 組のクエリーのセットを用意した．これらは，ランダムに選んだ 50 個のニュース記事のタイトルから被験者に選んでもらった．1 組は 3〜5 個のキーワードからなり，全部で 175 キーワードを用いた．本稿では 1 形態素を 1 キーワードとしている．例えば，“地下鉄　サリン　事件”，“いじめ　国際　シンポジウム” などといったものがクエリーとなる．この 50 組のキーワード構成のクエリーを，被験者が自然な問い合わせ調の文になるように変換してもらった．もちろん，自然文に変換したクエリーは，すべてのキーワードを含んでいる．図 3に本実験で用いたクエリーの例を示す．こうして作成した 2 種類のクエリーを被験者に発声してもらい，実験に用いた．クエリーの認識に使用した音声認識システムは，ニュース音声データベースを書き起こしたときに用いたものと全く同じである．表 2で，クエリー全単語，名詞のみ，グルーピングにより選択したキーワード候補，それぞれの場合で，認識仮説の 1-best のみのを集計したものと，3-best までの仮説の認識率を集計したものを併せて示している．3-best の認識率は，仮説内に出現した全単語の時間情報を基に単語を並べて擬似的に文を生成し，計算した．当然，正解率は向上するが，挿入誤りが増加するため正解精度は減少する．全体的に認識率は自然文のクエリーの認識率の方が，孤立キーワード発話のクエリーよりも認識率が良くなっている．それぞれのクエリーのパープレキシティは，自然文のクエリーのパープレキシティが 398（名詞のみに注目したパープレキシティは 1101）であるのに対して，キーワードのみから構成されているクエリーのパープレキシティが 2248 と大幅に大きくなっている．ニュース原稿から学習した語彙サイズ 2 万の言語モデルを用いているため，孤立キーワード発話の場合は複合語を除いては bigram が存在せず，各単語での unigram が適用されるため，パープレキシティは高くなる．一方で，自然文のクエリーだとキーワード同士を適切な助詞などで結んでいるため言語制約が効きパープレキシティは孤立キーワードのクエリーよりも低くなっている．認識率は複数の仮説を使った方が良くなっていることから，複数の認識仮説を使うことが有効であると考えられるが，その反面挿入誤りが大幅に増加している．これを 2.3 節で述べたグルーピングを用いて，必要なキーワード候補だけを選択すると，単語正解率は落ちるものの，挿入誤りを抑えることが出来ている．

3.3 検索方法

検索性能は，再現率-精度カーブ（ROC）で評価する．理想的な結果と比較するのであれば，すべての入力キーワードが含まれる文書を検索してくれれば良い．しかし，ここでは ROC カーブを描くため，以下に示すような n を定義する．検索語セット Q のある一つの検索語 q_i（$q_i \in Q, Q = \{q_1, \cdots, q_{50}\}$）中のキーワード w がある文書 d に含まれている，含まれていないを，

$$hit(d,w) = \begin{cases} 0 & (w\text{が } d \text{ に含まれない}) \\ 1 & (w\text{が } d \text{ に含まれている}) \end{cases}$$

と定義すると，d に含まれている q_i中のキーワード w の総数は，

$$hit(d,q_i) = \sum_{w \in q_i} hit(d,w)$$

と表すことができる．今回の実験では，$hit(d,q_i) \geq n(|q_i|)$ の条件を満たす文書を検索する．ここで $n(|q_i|)$ はある検索語 q_iを構成しているキーワードの数 $|q_i|$ 毎に設定した閾値である．この条件で検索された文書集合を，

$$output(q_i, n(|q_i|)) = \{d | hit(d,q_i) \geq n(|q_i|)\}.$$

と表すと，検索語 q_iによって検索された文書集合 $output(q_i, n(|q_i|))$ に対する再現率（$Recall_{q_i,n(|q_i|)}$）と精度（$Precision_{q_i,n(|q_i|)}$）は次のように計算できる．

$$Recall_{q_i,n(|q_i|)} = \frac{\text{検索された文書のうち正解文書数}}{\text{正解文書数}}$$

$$Precision_{q_i,n(|q_i|)} = \frac{\text{検索された文書のうち正解文書数}}{\text{検索された文書数}}$$

ここで，閾値 $n(|q_i|)$ の設定は自由度があり，j 種類目の閾値を $n_j(|q_i|)$ と表し，その集合を $N_j = \{n_j(|q_1|), n_j(|q_2|), \cdots, n_j(|q_{50}|)\}$ とする．すなわち，閾値の集合 N_jにおいて，すべての文書に対する再現率 $Recall_{N_j}$と精度 $Precision_{N_j}$は，

$$Recall_{N_j} = \frac{1}{50} \sum_{i=1}^{50} Recall_{q_i,n_j(|q_i|)}$$

$$Precision_{N_j} = \frac{1}{50} \sum_{i=1}^{50} Precision_{q_i,n_j(|q_i|)}$$

▼キーワード構成のクエリー
サッカー，ワールドカップ，日本，韓国
オーストラリア，銃，規制
地下鉄，サリン，事件

↓

▼自然文構成のクエリー
日本と韓国で行われるサッカーのワールドカップに関する情報を見せてください．
オーストラリアでの銃規制に関する記事が欲しい．
地下鉄サリン事件についての記事はありますか？

図 3: 実験で使用した 2 種類のクエリーの例

表 2: 2 種類のクエリーの単語認識率 [%]

クエリーのタイプ	全単語					名詞のみ		グルーピング後	
	Cor.	Acc.	Sub.	Ins.	Del.	Cor.	Acc.	Cor.	Acc.
孤立キーワード (1-best)	77.5	61.1	15.5	16.4	7.0	77.5	61.1	64.7	57.8
孤立キーワード (3-best)	78.3	45.0	15.0	33.3	6.7	78.3	45.0	65.8	57.2
自然文 (1-best)	88.3	80.1	11.2	8.1	0.6	85.8	65.3	75.8	74.1
自然文 (3-best)	89.0	55.1	10.4	33.9	0.6	88.3	41.4	76.1	70.6

Acc. = 100 - Sub. - Ins. - Del.[%]
Cor. = 100 - Sub. - Del. [%]

で求まる．本実験では $j = 1 \sim 10$ と 10 種類のキーワードの数毎の閾値の集合を用意し，再現率-精度カーブを求める[2]．各カーブでの最大の F 値

$$F\text{値} = \frac{2RP}{(R+P)} \quad \text{R：再現率} \quad \text{P：適合率}$$

での評価も行う．

3.4 検索実験結果

図 4に，孤立キーワード発声のクエリーを入力した場合の再現率–精度カーブを示す．①のカーブ（1-best の認識仮説の認識結果をそのまま使用）と③のカーブ（3-best までの認識仮説をそのまま使用）を比較してみると，最大となる F 値で①のカーブの方が（$F = 71.3$）③のカーブよりも（$F = 68.5$）検索性能が良くなっている．これは，クエリーを発話した内容とまったく関係のない単語が多く挿入されたからだと考えられる．しかし，N-best の認識仮説からグルーピングにより適切なキーワード候補を選択し，それを検索に使うことで，④のカーブのように（$F = 72.9$）もっとも良い検索結果が得られた．①，②のカーブによりグルーピングは 1-best 候補に対しても若干有効であることがわかる．

次に，図 5に自然文のクエリーを音声入力した場合の再現率– 精度カーブを示す．図 4の場合と同様に，④の 3-best までの認識仮説に対して，グルーピングにより適したキーワードを選択した場合のカーブがもっとも高い F 値（$F = 73.4$）を得た．図 4と図 5で，グルーピングを適用していない場合は，孤立キーワードの音声クエリーを用いた方が良い．今回試作した検索システムでは，音声クエリーの認識結果から単にストップワードを取り除くという単純な処理しか行っていないため，より長い自然文の音声クエリーの認識結果からストップワードを取り除いてもまだなお，関連のないキーワード候補が残ってしまう．そのため，孤立キーワードの音声クエリーの方が良くっている．しかし，グルーピングにより，それら関連のないキーワードをさらに取り除くことによって，かなりの性能改善を得ることができた（図 5での③→④）．

最終的に，グルーピング結果を併用することで，もっともクエリー音声の単語正解率の高かった（88.3%）図 5の④のカーブが検索性能も良くなり，より発声しやすく認識しやすい自然文の音声クエリーを使う方が良いということが分かった．

[2]もちろん同じ jのときはキーワードの数が同じ検索語は同じ閾値に設定してある

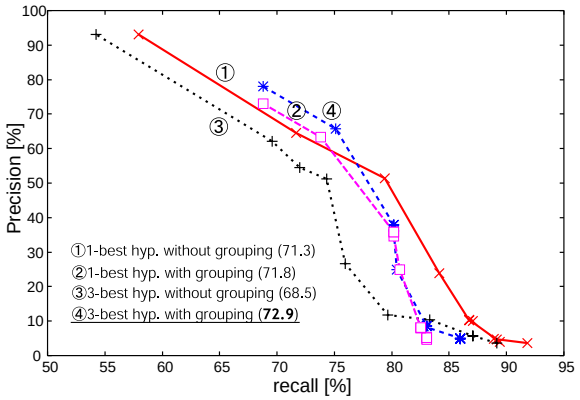

図 4: 孤立キーワード音声クエリーを用いた場合の検索性能

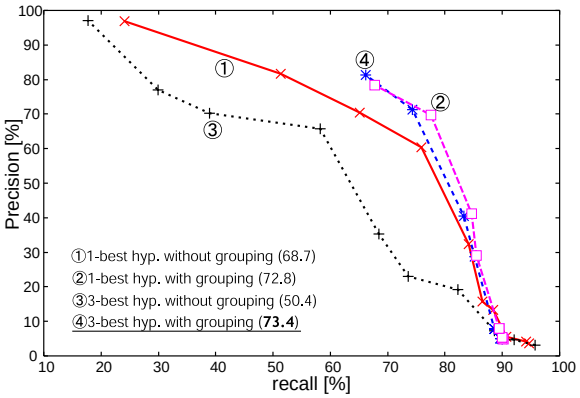

図 5: 自然文音声クエリーを用いた場合の検索性能

4 おわりに

本稿では，今後必要となってくるであろう音声クエリーを用いた音声文書検索システムについて述べた．キーワード孤立発声の音声クエリーと，自然文の音声クエリーの 2 種類の形式のクエリーを用意し，検索実験を行った．パープレキシティの値からもわかるように，自然文の音声クエリーを使う方がキーワード孤立音声よりも認識率が良くなった．また，認識結果の N-best 仮説を用いることで，さらに認識率の向上が図れるが，湧き出し単語が増えてしまう．しかし，相互情報量に基づくキーワード候補選択手法をクエリーの認識結果に適用することで，一番認識率の高かった自然文音声クエリーの N-best 出力を使った場合の検索性能が最も良くなった．検索対象の音声データの書き起こし認識率は，挿入誤りが若干多くても単語正解率を高くする方が良いが [9][11]，音声クエリーの場合は，単語正解率と正解精度の両方が高い（挿入誤りが小さい）ことが望ましいことが分かった．

今後の課題は，この音声入力型の検索システムを改良し，音声による質問を受け付けることのできる質問応答システムを開発する予定である．

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感謝致します．また，本研究を進めるにあたって適切な助言をいただいた，豊橋技術科学大学の宇津呂武仁講師に深く感謝致します．

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] 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).

[3] 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).

[4] John.S.Garofolo, et.al, : The TREC Spoken Document Retrieval Track: A Success Storyl, *Proc. of the 8th Text Retrieval Conference*, pp. 107–130, 1999.

[5] 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, 電子情報通信学会, pp. 97–102 (1999).

[6] 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. of ICSLP'98*, pp. 2867–2870 (1998).

[7] Kemp, T., Geutner, P., Schmidt, M., Tomaz, B., Weber, M., Westphal, M. and Waibel, A.: The Interactive Systems Labs VIEW4YOU Video Indexing System, *Proc. of ICSLP'98*, pp. 1639–1642 (1998).

[8] 伊藤克亘, 藤井敦, 石川徹也: 未知語検出を用いた語彙統制のない音声検索, 日本音響学会講演論文集 I, pp. 103–104 (2002).

[9] 西崎博光, 中川聖一: 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, 情報処理学会論文誌, Vol. 42, No. 12, pp. 3173–3184 (2001).

[10] 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).

[11] 西崎博光, 中川聖一: 未知語を考慮した音声文書検索の検討, 音響学会秋期講演論文集 I, 2-1-11, pp. 71–72 (2001.10)