# 音声入力によるＷｅｂ検索のための キーワード認識・抽出法の検討

松下 雅彦[†]　西崎 博光[‡]　宇津呂武仁[††]　中川 聖一[†]

[†]豊橋技術科学大学 情報工学系
[‡]山梨大学 大学院 医学工学総合研究部，[††]京都大学 大学院 情報学研究科

近年，ウェブ検索の分野では，NTCIR ワークショップなどで競争型のコンテストがおこなわれるなど，研究が盛んにおこなわれている．我々は，ウェブ検索の分野の中でも，音声入力によるウェブ検索の有用性に着目し，NTICR-3 のために準備されたデータを用いて実験をおこなった．本研究では，音声認識率を向上させることにより，同時に検索精度をも向上させることを目的としている．そこで今回は，複数音声認識モデル混合の手法を用いて音声認識率の向上を計った．混合手法としては SVM 学習と ROVER 法を用いている．また，学習方法は SVM と同じであるが，適用時に SVM の出力に冗長性を持たせた SVM(冗長) と呼ぶ手法も提案し，評価している．音声認識率を向上させることにより，検索性能も向上させることができた．

# Keyword recognition and extraction for speech-driven Web retrieval task

Masahiko Matsushita[†], Hiromitsu Nishizaki[‡],
Takehito Utsuro[††] and Seiichi Nakagawa[†]

[†]Department of Information and Computer Sciences, Toyohashi University of Technology
[‡]Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi
[††]Graduate School of Informatics, Kyoto University

This paper studied on Web retrieval models which accept speech-driven queries in the NTCIR-3 Web retrieval task. The major focus of this paper is on improving speech recognition accuracy of spoken queries and then improving retrieval accuracy in speech-driven Web retrieval. We experimentally evaluate the techniques of combining outputs of multiple LVCSR models in recognition of spoken queries. As model combination techniques, we compare the SVM learning technique with conventional voting schemes such as ROVER. We also tested the SVM(redandant), which was an expanded use of SVM outputs. We show that the technique of multiple LVCSR model combination can achieve improvement both in speech recognition and retrieval accuracies in speech-driven Web retrieval.

## 1　はじめに

近年の音声認識技術は非常に進展してきており，その重要性も増してきている．実際に，一般のパソコン上でも動作する音声認識ソフトウェアも増えつつあり，音声入力を用いたアプリケーションも徐々に出はじめている．一方，情報検索の研究も非常に盛んに行なわれており，この 2 つを組み合わせた研究も最近になり徐々に増えつつある．

音声データを用いた検索には，検索対象が音声データである場合と，質問を音声で入力する場合 (音声クエリー) が考えられる．音声データの検索としては，TREC-6 の Spoken Document Retrieval(SDR) トラック [7] で音声データを対象にしたテストコレ

クションが整備されたことにより盛んに研究が行なわれ始めた．音声クエリー入力による検索としては，Barnett ら [2] が既存の音声認識システムを用いてテキスト検索を行なっている．Crestani [3] も音声入力による検索を行ない，適合性フィードバックによって検索精度が向上することを示している．上記の 2 つの手法は検索精度の改善についてのみ焦点を当てており，認識精度の改善については触れていない．

音声クエリーの音声認識を改善することにより，検索性能を向上させようと試みる研究も行なわれている．

藤井ら [8, 6] は音声入力による検索において，音声認識と検索精度の両方の改善を行なっている．彼らは，検索対象から言語モデルを作成し，それにより未知語を減少させ，認識精度の改善を行っている．

西崎らは音声キーワードと音声データベースを用いて検索実験を行っている．ここでは，グルーピングと呼ばれる手法と N ベスト出力の結果を用いて，音声認識誤りに対して頑健に対処している．また，音声クエリーとして，キーワード入力法と自然文入力法を比較し，後者の方が高い音声認識精度が得られることを示している [14].

今回の我々の研究も，音声の認識率の向上により，検索性能を改善させることを目的としているが，音声認識の改善方法として，複数音声認識モデル出力の混合 [11] を用いている．すでに，この手法により認識率が改善されることは我々の研究室での研究により判っている．そこで今回は，その複数モデル出力の混合を用いて，音声認識率を改善し，同時に検索精度の改善を行なうことを目的とした．実験では，NTCIR-3 音声入力 Web 検索タスク [8] のデータを用いている．混合手法としては，機械学習法である SVM [9] と多数決法である ROVER 法 [5] を用いている．また，学習方法は SVM と同じであるが，適用時に SVM の出力に冗長性を持たせた SVM(冗長) と呼ぶ手法も提案し実験している．

# 2 NTCIR-3 音声入力 Web 検索タスク

## 2.1 NTCIR ワークショップ

NTCIR ワークショップ [4] は，情報検索とテキスト要約・情報抽出などのテキスト処理技術の研究をより発展させることを目的とした評価会議である．実験用のデータセットと実験結果を評価するための統一された手順が用意されており，参加グループは用意されたデータを用い，様々なアプローチで研究と実験を行なっている．

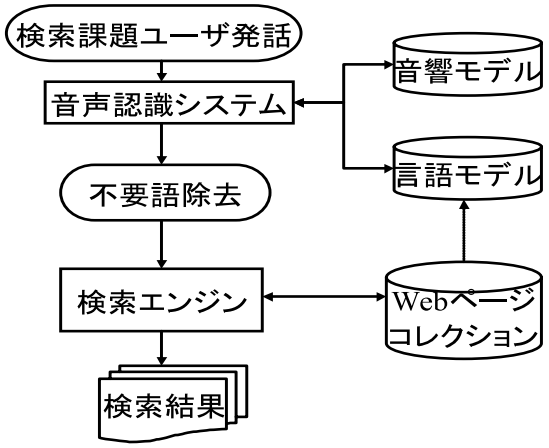


図 1: 音声入力を用いた Web 検索の流れ

また，NTCIR では,「テストコレクション」と呼ばれる実験用データセットを整備・公開している．テストコレクションとは，情報検索分野で情報検索システムの性能評価に用いるデータである．

## 2.2 音声入力 Web 検索タスク

NTCIR-3 の音声入力 Web 検索タスク [8] は筑波大の藤井らがオーガナイザーとなり始められたタスクである．NTCIR 側で用意された文書集合から，音声の検索質問 (query) を用いて検索実験を行なう．実際の実験に用いられているテストコレクションは NTCIR の Web 検索タスクに用いられているものと同じである．検索文書集合は参加者が扱いやすいサイズ (10GB 程度) と，ある程度現実に近いサイズ (100GB 程度) が設定されている．各検索課題ごとに適合度順の検索結果の上位 1000 ページを順位付きで提出し，それを人手で判定するプーリングと呼ばれる手法を用いて正解候補を作成する．ただし，このような便宜的な方法の場合は，正解が一部に偏よる場合があり，実際には正解であっても，不正解であると判定する可能性もある．この時，正解候補は 4 段階 (高度に適合，適合，部分適合，不適合) で判定が行なわれている．

一般的な検索の手順としては図 1 に示す通りである．まず，音声認識システムを用いてユーザーが発話した検索課題音声の認識を行なう．次に，認識結果から不要語を除去しキーワードリストを作成する．最終的に，このキーワードリストを用いて検索を行なう．

# 3 複数音声認識モデル出力の混合

今回，我々は音声入力 Web 検索タスクの音声認識に着目して実験を行なった．音声認識率を向上させ

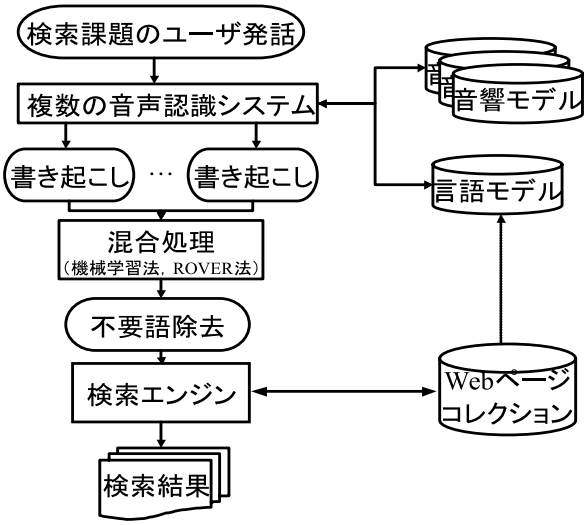


図 2: 複数音声認識モデル混合を用いた Web 検索

ることにより検索精度も同時に向上させることを目的としている．そこで，音声認識率を向上させるために，本稿では複数の大語彙連続音声認識モデルの出力を混合する手法 [11] を適用する．複数モデルの出力の混合法として，機械学習法である SVM(サポートベクターマシン)，および，多数決法である ROVER 法を用いた．

SVM は，一つ一つの属性を次元と見なし，単語の持つ属性の組み合わせをベクトルと考えることにより，2 クラスの分類を行なうための座標変換式を学習する手法である．ROVER 法は，複数のモデル出力間で多数決をとり，混合を行なう手法である．

混合手法を用いた場合の検索手順は図 2 の通りである．まず，複数の音声認識システムを用いてユーザーが発話した検索課題音声の認識を行なう．次に，出力された複数の認識結果を用いて混合処理 (SVM, ROVER) を行なう．その結果より，不要語を除去し，キーワードリストを作成する．最終的に，作成されたキーワードリストを用いて文書集合より検索を行なう．

## 3.1 大語彙連続音声認識モデル

### 3.1.1 デコーダ

大語彙連続音声認識モデルのデコーダとしては次の 2 種類を用いる．1 つ目としては，我々の研究室 (豊橋技術科学大学 中川研究室) で開発した SPOJUS [1] であり，もう 1 つは CSRC「日本ディクテーション基本ソフトウェアの開発」プロジェクト [10] から提供された Julius(ver.3.2) である．Julius，SPOJUS のどちらのデコーダにおいても 2 パス探索により認識を行ない，1 パス目では単語バイグラム，2 パス目では単語トライグラムを，それぞれ使用した．

(a)Julius

Julius は，1 パス目ではバイグラムを用いた 1best 近似を行ない，1 パス目で得られた単語トレリスを用いて，2 パス目で単語間にわたるクロストライフォンを用いている．

(b)SPOJUS

SPOJUS は，1 パス目ではバイグラムを用いた 1best 近似 (改良版は 1-best と線形辞書の併用[1]) を行ない 200 ベストを求める．本実験においては，2 パス目でコンテキスト依存モデルは用いていない．

### 3.1.2 音響モデル

音響モデルの学習には，日本音響学会読み上げ音声 JNAS を用いた．

(a)Julius

Julius では音素を基本単位とする HMM モデル，および，音節を基本単位とする HMM モデルを用いた．16KHz サンプリング，フレーム周期 10ms，特徴ベクトルは MFCC(12 次元) + ΔMFCC + ΔPOW (計 25 次元) の条件で行ない，HMM としては次の 4 種類を用いて実験を行なう．

- モノフォンモデル
- トライフォンモデ
- 音素内タイドミクスチャ(PTM) モデル
- 音節モデル [13]

(b)SPOJUS

SPOJUS では音節を基本単位とする HMM を用いており，デコーダと同様に，我々の研究室で開発されたものである．16kHz サンプリング，フレーム周期 10ms，25ms ハミング窓，特徴ベクトルは MFCC(セグメント単位): MFCC(10 次元 × 4 フレームを KL 展開で 20 次元に圧縮) + ΔCEP + ΔΔCEP + ΔPOW + ΔΔPOW (計 50 次元)，および MFCC(フレーム単位): MFCC(12 次元) + ΔMFCC + ΔΔMFCC + ΔPOW + ΔΔPOW (計 38 次元) の 2 種類，さらに，継続時間制御/自己遷移ループの 2 種類，合計以下の 4 種類のモデルで認識を行なう．

- MFCC-seg + 継続時間制御
- MFCC-frm + 継続時間制御
- MFCC-seg + 自己遷移ループ
- MFCC-frm + 自己遷移ループ

### 3.1.3 言語モデル

言語モデルは藤井ら [8] のモデルを利用した．この言語モデルは，検索対象の文書集合である 100GB のコレクションを用いて作成された．このコレクション中から頻度上位 2 万語に制限した単語トライグラムおよびバイグラムを作成した (評価データの未知

[1]北岡ら [15] による SPOJUS の改良版

語率は 4.2%，キーワードの未知語率は 18.0%)．言語モデルの作成方法は，大語彙連続音声認識ツールキットでの作成方法 [12] に準じているが，Web 文書の場合，英語の文書なども混ざっているので，ASCII 文字だけからなる段落などを排除する前処理を追加した．

平滑化には，バックオフスムージング (Witten-Bell ディスカウント) を用いた．カットオフはバイグラム，トライグラム共々20 とした．バイグラムは前向きのもの，トライグラムは前向きのものと逆向きのものの両方を ARPA 形式で作成した．

## 3.2 評価データ

検索対象は NTCIR-3[4] で用意された 100GB(1000 万ページ) 程度の Web 文書である．Web 検索タスクでの検索課題 105 文を合計 10 人の話者 (男性 5 人，女性 5 人) がそれぞれ発話したものが音声クエリーとして準備されている．ただし，今回の我々の実験は男性５人の音声クエリーだけを用いて行なっている．

評価データとしては検索課題 105 文中から 53 文を用いる．ただし，実際に評価として用いているのは正解が公表されている 47 文のみである．機械学習で用いる学習データは，評価として用いられない検索課題 (52 文) を，評価者以外の 4 人分 (合計 208 文) を合わせたものを用いる．検索課題の例を以下に示す．

0008 サルサを踊れるようになる方法が知りたい。
0016 ゲノム創薬の最近の動向について述べられている文書を探したい。
0146 ドメスティックヴァイオレンス、DV の現状について調べたい。

先頭の数字は，検索課題番号を示している．認識結果を評価するための正解データとしては，検索課題を書き起こしたテキストを用いている．キーワードリストの評価用の正解データは，認識結果の出力と混合処理結果の出力からキーワードリストを作成する手順と同じ手順を用いて，検索課題の書き起こし結果から作成する．不要語の処理などでヒューリスティックスを用いているため，正解としているキーワードが必ず検索において適切であるとは限らない．

## 3.3 複数モデルの混合手法

本稿では，先に述べた 8 種類の大語彙音声認識モデルの出力を混合し，単語認識率を向上させる．混合の方法としては，機械学習の手法である SVM，多数決法である ROVER 法を用いる．

### 3.3.1 SVM を用いた混合

SVM は，一つ一つの属性を次元とみなし，単語の持つ属性の組み合わせをベクトルと考えることにより，2 クラスの分類を行うための座標変換式を学習する．複数の大語彙連続音声認識モデルの出力を混合する手法 [11] では，クラス分類時に計算される 2 クラスの境界からの距離を信頼度とすることで，音声認識性能の向上を図っている．ベクトルの素性としては次の 3 つを用い，各単語の正誤を判別対象のクラスとする．

- 単語の品詞
- 音節数
- 単語を出力したモデルの情報

SVM では，個々の素性を次元とする多次元空間中の点として記述された事例の集合に対し，全事例を 2 クラスに分類する境界面を学習し，適用時は評価事例と境界面との間の距離 (信頼度) に閾値を設け，単語ラティス上の競合する単語中で，これらの閾値を越える距離を持つ単語が存在すれば，その中で境界面からの距離が最も大きい単語を 1 つ混合結果として出力する．もし，閾値を越える距離を持つ単語が存在しない場合は，そのラティス上では単語は出力されない．

### 3.3.2 SVM(冗長) を用いた混合

SVM(冗長) は，正解精度を犠牲にして正解率を上昇させることを目的とした手法である．学習は SVM と同じ方法で行う．適用を行なう場合は SVM とは異なり，単語ラティス上の競合する単語中で閾値を越える距離を持つ単語が存在する場合は，閾値を越える距離を持つ単語はすべて出力する．この手法では，単語ラティス上で競合する単語があったときに複数個以上の単語を出力できる可能性があり，SVM の場合より正解率を向上させることができる．

### 3.3.3 ROVER 法を用いた混合

ROVER 法では，学習は行なわず，複数のモデルの出力の多数決により混合を行なう．重みつき多数決を用いる場合は，重みとして，各モデルの単語正解率を用いる．重みを用いない場合で，等しい数の投票があった場合はそのラティス上の単語は出力されない．

# 4 Web 検索

## 4.1 キーワードリストの作成

キーワードリストを作成するために，混合処理などにより得られた結果から内容語 (主に自立語) を

抽出し，さらにそこから不要語を取り除いた．不要語は，ヒューリスティックスにより選択し，検索時によく用いられる単語 (探す，知り，など) と，検索には必要ないと思われる単語 (平仮名 1 文字) を削除した．この処理によって出力された単語を検索のためのキーワードとする．キーワードリストの例を以下に示す (リスト中の * マークが付いた単語は未知語である).

0008 サルサ * 踊れる * 方法
0016 ゲノム 創薬 * 最近 動向 述べ
0146 ドメスティックヴァイオレンス * DV 現状

### 4.2 検索エンジン

キーワードとして抽出された単語を検索エンジンに入力し，検索を行なう．今回実験に用いた検索エンジンは，藤井ら [8] の作成した統計的手法を用いたものである．クエリー $Q$ とテキスト $D_i$ の類似度 $sim(Q, D_i)$ は次式で計算される．

$$sim(Q, D_i) = \sum_t \left( \frac{TF_{t,i}}{\frac{DL_i}{avglen} + TF_{t,i}} \cdot log \frac{N}{DF_t} \right)$$

ただし，$t$ は検索要求に含まれる索引語，$TF_{t,i}$ はテキスト $i$ 中の索引語 $t$ の出現頻度，$DF_t$ は索引語 $t$ を含む検索対象テキストの数であり，$N$ は検索対象のテキスト総数，$DL_i$ はテキスト $i$ の文書長 (バイト数)，$avglen$ は検索対象中の全テキストに関する平均長である．

検索エンジンにキーワードを入力すると，検索対象である文書集合 (100GB) の各ページに対して上式を用いて類似度を計算し，スコアが高いページから順次列挙する．このエンジンは，キーワードの挿入誤りには比較的頑健であることから，単語正解率 (correct) が検索性能に直接影響を及ぼすと考えられる (検索エンジンよっては，挿入誤りは致命的になる場合もある).

### 4.3 検索精度の評価方法

正解判定は 4 段階 (高度に適合，適合，部分適合，不適合) で行なう．実際に評価を行なう場合は，高度に適合と適合の判定は混合して考えている．その他に，ハイパーリンク情報の利用の有無も考慮するため，最終的には以下の 4 つで評価を行なう．

- RC : (高) 適合，ハイパーリンク情報を用いない．
- RL : (高) 適合，ハイパーリンク情報を用いる．
- PC : 部分適合，ハイパーリンク情報を用いない．

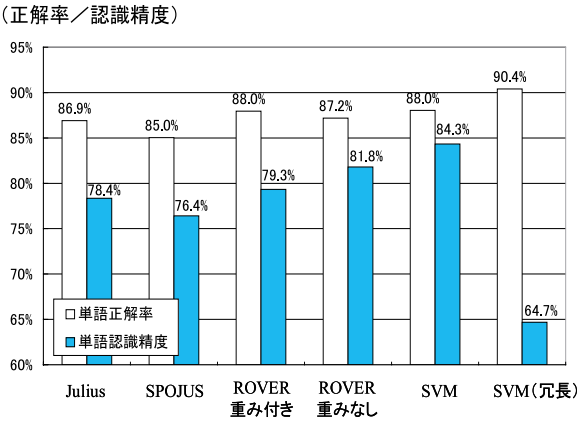


図 3: 評価用検索課題 47 文の単語の正解率・認識精度

- PL : 部分適合，ハイパーリンク情報を用いる．

検索精度は平均適合率を用いて計算を行なっており，各クエリーの検索結果上位 1000 件から計算する．平均的合率は，各再現率レベルでの適合率の平均値 (適合文書が検索された時点での適合率の平均) である．もし，正解ドキュメントが検索されなかった場合は適合率は 0.0%である．最終的に，話者 5 人の平均を求めて評価する．

## 5 実験結果

### 5.1 音声クエリーの認識精度

表 1 は評価用検索課題 47 文を 8 種類のモデルで認識を行なった結果である．図 3，図 4，は音声クエリー 47 文の単語認識率およびキーワード認識率の 5 人の平均を示している．図 4 で示すキーワード認識率は音声認識結果から不要語を除去した出力の認識率である．図中にある Julius と SOPJUS は各デコーダの中で最も単語正解率が高い認識システムの結果 (Julius ではトライフォンモデル，SPOJUS では MFCC-seg + 継続時間制御) を示している．

図より，混合処理を行なった場合には，単独の認識システムの場合と比較して正解率と正解精度の改善が達成されていることが判る．特に，SVM を用いた場合の認識精度の改善は非常に大きいといえる．SVM(冗長) と SVM を用いた場合を比較してみると，正解精度を犠牲にして，正解率の改善を達成できていることがわかる．

### 5.2 検索精度

図 5 に音声入力による Web 検索実験の結果を示す．Julius と SPOJUS はそれぞれのデコーダ中で，

表 1: 単独の認識器での認識精度 [%]

| 認識モデル | | 単語正解率 | 単語認識精度 | キーワード正解率 | キーワード認識精度 |
|---|---|---|---|---|---|
| Julius | モノフォンモデル | 73.2 | 66.9 | 60.3 | 44.7 |
| | トライフォンモデル | 86.9 | 78.4 | 71.8 | 57.8 |
| | PTM モデル | 85.8 | 77.5 | 71.3 | 54.7 |
| | 音節モデル | 84.3 | 77.8 | 68.7 | 56.6 |
| SPOJUS | MFCC-seg + 継続時間制御 | 85.0 | 76.4 | 68.8 | 53.4 |
| | MFCC-frm + 継続時間制御 | 84.3 | 76.5 | 67.6 | 54.1 |
| | MFCC-seg + 自己遷移ループ | 81.8 | 75.0 | 64.9 | 51.4 |
| | MFCC-frm + 自己遷移ループ | 81.8 | 75.2 | 64.8 | 52.8 |
| | MFCC-frm + 自己遷移ループ[1] | 85.1 | 77.6 | 68.8 | 55.8 |

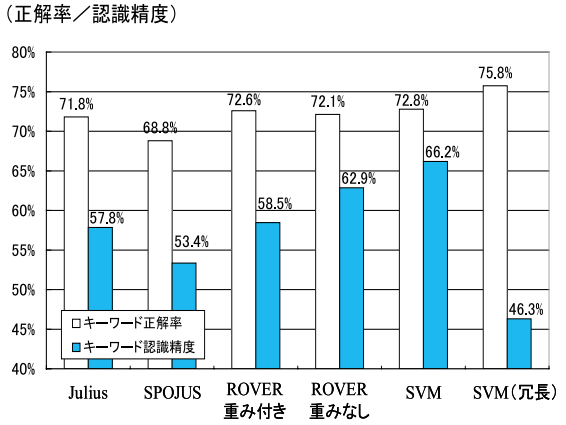


図 4: 評価用検索課題 47 文のキーワードの正解率・認識精度

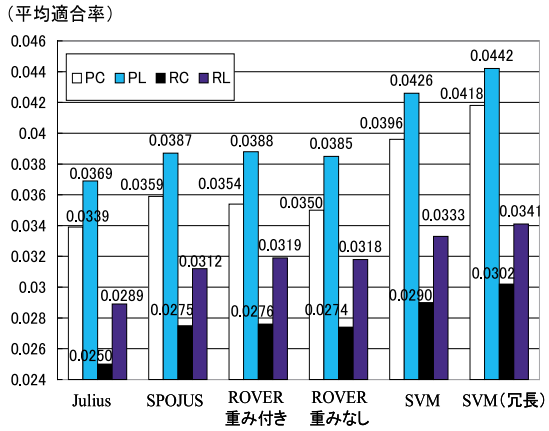


図 5: 検索精度評価

最も単語正解率が良かった場合の結果を示している．キーワードの認識率が高かった Julius より，低い SPOJUS のほうがより高い検索精度を示した．これは，検索エンジンはキーワードの重みを考慮しているため，キーワードの認識率と検索精度が線形な関係でないことによる．したがって，SPOJUS では重みが大きいキーワードがより多く認識されていたと考えられる．ROVER 法を用いた場合の結果は，Julius の単独の結果よりは良くなっているが，SPOJUS 単独の結果と比べると，大きな違いは見られなかった．SVM を用いた場合の結果は，単独の認識モデルや ROVER 法を用いた場合に比べて非常に高い精度を示している．SVM と SVM(冗長) を比較した場合，正解率を重視した SVM(冗長) の場合のほうがより良い結果を得ることができた．これは，音声入力 Web 検索においては認識精度より正解率が重要であることを示している (これは，タスクに依存する．評価用タスクではキーワードが少なすぎる場合，上位 1000 ページに正解候補が少なくなると考えられる)．距離閾値は検索精度が最も良くなる値を用いており，SVM の場合が-1，SVM(冗長) の場合は-1.5 であった．

## 5.3 テキスト検索

表 2 は入力を音声のかわりにテキスト (音声認識率 100%に相当) にした場合の検索結果である．これは，認識結果を評価するために用いている正解データである．テキストの場合に比べて，音声入力による Web 検索の難しさが良く分かる．全文を用いて評価を行なった場合と未知語がクエリー中に存在しない 22 文のみで評価を行なった場合を比較すると，後者の検索精度に大きな向上が見られる．これより，未知語が検索精度の大きな低下の原因になっていることが分かる．したがって，言語モデルの語彙数を増やし，未知語の数を減らすことにより音声入力による検索の精度の向上が期待できると思われる．

## 5.4 SVM(冗長) におけるキーワード数の制限

SVM(冗長) は，各ラティス上の単語の距離が閾値以上の値であれば，複数出力することを許している．しかし，複数個出力することにより，もう一つの実験として，SVM(冗長) におけるキーワードの制限を行なった．キーワードの挿入誤りと脱落誤りは検索精度に悪影響を与える．そこで，SVM(冗長)

表 2: テキスト検索による検索精度評価

| | 全 47 文での評価 | | | | 未知語のないクエリー 22 文での評価 | | | |
|---|---|---|---|---|---|---|---|---|
| | PC | PL | RC | RL | PC | PL | RC | RL |
| テキストクエリー入力 | 0.1181 | 0.1214 | 0.0843 | 0.0960 | 0.0976 | 0.0985 | 0.0728 | 0.0819 |
| 音声クエリー入力：SVM(冗長) | 0.0418 | 0.0442 | 0.0302 | 0.0341 | 0.0700 | 0.0753 | 0.0504 | 0.0582 |

の出力結果の単語の数を制限することで，その受ける影響を調査した．具体的には次の 2 種類の単語数の制限を行なった．

### 5.4.1 閾値によるキーワード数の制限

ここでは，SVM(冗長) 適用時の距離閾値を変化させることにより単語数の制限を行なう．図 6 に距離閾値を変化させた場合のキーワード認識率と検索精度の結果を示す．距離閾値の値は，学習した 2 クラスの境界面を距離閾値だけ移動させる．今回は 0 からマイナスの値で調査を行なっているので，境界面はマイナス (誤りのクラス) 方向に移動させている．距離閾値の値が大きければキーワードの脱落誤りが多くなり，距離閾値の値が小さくなればキーワードの挿入誤りが多くなる．図 6 を見ると，0 から-3 に距離閾値を変化させた場合，-1.5 で検索精度のピークを得た．距離閾値を下げることにより，挿入誤りも増えているが，正解も取り込めているため検索精度は向上している．ただし，過度に距離閾値を下げてしまうと，挿入誤りが増加しすぎることにより検索精度を低下させてしまう．図 6 のキーワードの認識率を見てみると，正誤の境界面が現在の境界面より距離が-0.5 移動した付近で取られることが理想であると判る．すなわち，SVM の学習により完璧な正誤の境界が得られてはいないため，距離閾値などで境界面を調整するこが望ましい．距離閾値が最適な値になっていない場合は，挿入誤りと脱落誤りのバランスがうまく取れず，検索精度を悪化させてしまう．距離閾値を適切に設定することにより，正解単語をより多く抽出し，より高い検索精度を得ることができる．ただし，あまり挿入誤りが多い場合は，逆に検索精度を低下させてしまうため注意が必要である．

### 5.4.2 競合単語数の制限

次に，SVM(冗長) 適用時の競合単語の出力数を変化させることにより単語数の制限を行なう．競合単語とは，SVM(冗長) 適用時に各ラティス上で境界面からの距離が最大でなく，かつ閾値以上である単語を示す (SVM 適用時には出力されず，SVM(冗長) 適用時のみ出力される単語)．図 7 は競合単語数を制限した場合のキーワード認識率と検索精度の

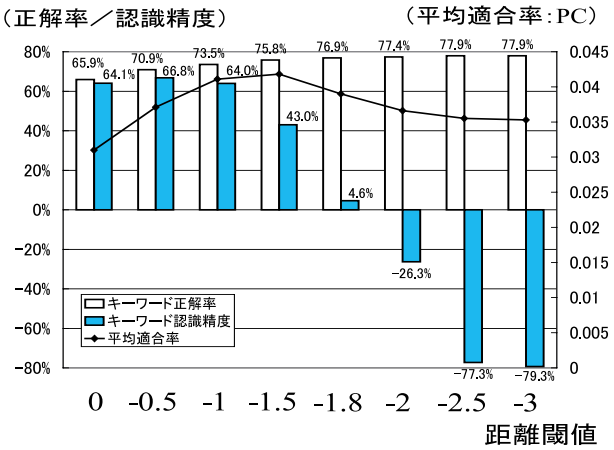


図 6: 閾値によるキーワード数の制限

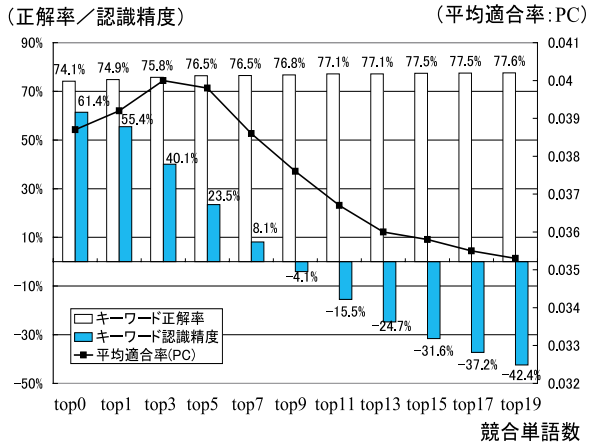


図 7: 競合単語の制限によるキーワード数の制限

結果である．topN とは，SVM の出力に SVM(冗長) のみで出力されている単語の中で境界面からの距離が大きいものから N 個加えることを意味している．top0 の場合は SVM の出力と同じである．N 個は，各クエリー中の全ての競合単語の中から選ばれる．今回は SVM(冗長) 適用時の距離閾値は設定しない (全ての競合単語を出力) で実験を行なっている．検索精度は N が 2 の場合に最も良くなっている．この場合も N の値が小さい場合はキーワードの脱落誤りが多くなり，N の値が大きすぎる場合はキーワードの挿入誤りが多くなり，検索精度を低下させてしまう．

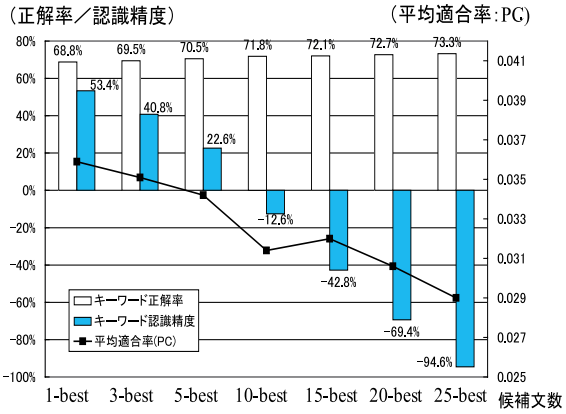


図 8: 単独の認識器の N-best での組み合わせ

### 5.5 単独の認識器 (SPOJUS) の N-best 出力を混合した場合の検索精度

混合手法としては，複数の音声認識モデル出力の混合の他にも単独の音声認識モデルの N ベストの出力を混合する手法が考えられる．さらに，今回のように機械学習や ROVER 法などを用いず，単純に加える (OR 混合) だけの場合も考えられる．コストの面で考えるならば，複数より単独，機械学習法や ROVER 法より OR 混合の方が良い．そこで，今回は，一番コストのかからない，単独の音声認識モデルの N ベスト出力の OR 混合を実験した．SPOJUS を用いて音声認識を行なった場合，認識結果の上位 N ベストの認識結果を得ることができる．これを用いて上位 N ベストの結果を用いてキーワードリストを作成し，検索実験を行なった．図 8 に結果を示す．検索精度に関してはまったく向上が見られなかった．1-best から 3-best への認識率の変化と図 7 の top0 から top1 への認識率の変化を比較すると，明らかに，top0 から top1 への変化の場合の方がより良い改善がなされている．これは，混合手法によってより，正確なキーワードが取り出されていることと，異なるシステムを用いることにより，比較的異なる単語が抽出されているためである．比較して，

## 6 おわりに

今回の実験で，音声認識率 (特に単語正解率) を改善し Web 検索性能を向上させることができた．今後，さらに検索性能を上昇させるためには，未知語率を減少させると共に，音声認識結果において，キーワード認識精度よりもキーワード正解率を優先させて検索のキーワード候補を多めに抽出することが望ましい．これは，SVM(冗長) を用いた実験からも明らかである．ただし，挿入単語が多すぎる場合は，逆に検索精度を低下させる原因になるため，距離閾値や競合単語の制限などで，キーワードとなる単語の数のバランスを取ることが重要である．

## 謝辞

この研究では，NTCIR-3 のデータを数多く利用させていただいた．これらのデータベース提供してくださった NTCIR ワークショップの関係諸氏に深く感謝致します．また，本研究を進めるにあたって適切な助言を頂いた，筑波大学 図書館情報学科の藤井敦助教授に深く感謝致します．

## 参考文献

[1] 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. 情処第 57 回全大講論集, pp. 35–36, 1998.

[2] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. W. Kuo. Experiments in spoken queries for document retrieval. In *Eurospeech97*, pp. 1323–1326, 1997.

[3] F. Crestani. Word recognition errors and relevance feedback in spoken query processing. In *Fourht International Conference on Flexible Query Answering Systems*, pp. 267–281, 2000.

[4] K. Eguchi, K. Oyama, E. Ishida, and K. Kuriyama. Overview of Web retrieval task at the third NTCIR workshop. In *Working Notes of the 3rd NTCIR Workshop Meeting*, pp. 1–24, 2002.

[5] J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proc. ASRU*, pp. 347–354, 1997.

[6] Atsushi Fujii and Katunobu itou. Building a test collection for speech-driven web retrieval. In *Eurospeech2003*, pp. 1153–1156, 2003.

[7] J. S. Garofolo, E. M. voorhees, V. M. Stanford, and K. S. Jones. Trec-6 1997 spoken document retrieval track overview and results. In *In Proceedings of the 6th Text Retrieval Conference*, pp. 83–91, 1997.

[8] 伊藤, 藤井. NTCIR-3 ワークショップにおける音声入力型ウェブ検索タスク. 情報処理学会研究報告, Vol. 2002, No. (2002–SLP–43), pp. 25–32, 2002.

[9] T. Joachims. Making large-scale svm learining practical. In *Advances in Kernel Methods Support Vector Learning*. MIT Press, 1999.

[10] 河原, ほか. 日本語ディクテーション基本ソフトウェア (99 年度版). 日本音響学会誌 (技術報告), Vol. 57, No. 3, pp. 210–214, 2001.

[11] 小玉康広, 渡邊友裕, 宇津呂武仁, 西崎博光, 中川聖一. 機械学習を用いた複数の大語彙連続音声認識モデルの出力の混合. 情報処理学会研究報告, Vol. 2003, No. (2003–SLP–45), pp. 95–100, 2003.

[12] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄（編）. 音声認識システム. IT Text. オーム社, May 2001.

[13] 山本, 池田, 松本, 西谷, 宮澤. コンパクトで高精度な音節モデルの検討. 日本音響学会秋期講演集, Vol. 2002, No. (1-9-22), 2002.

[14] 西崎, 中川. 音声キーワードによるニュース音声データベースの検索手法. 情報処理学会論文誌, Vol. 42, No. 12, pp. 3173–3184, 2001.

[15] 北岡, 高橋, 中川. N-best 線形辞書探索と 1-best 近似木構造辞書探索の併用による大語彙音声認識. 電子情報通信学会技術報告, SP2003-26, 2003.