

音声からキーワードを検出する技術の  
高度化に関する研究

山梨大学大学院  
医学工学総合教育部  
博士課程学位論文

2014年 3月  
名取 賢

# 音声からキーワードを検出する技術の高度化に関する研究

## 論文要旨

近年、マルチメディアデータの生成・編集環境の普及、ストレージの大容量化、ネットワークインフラの充実により、動画コンテンツに代表される音声やマルチメディアコンテンツが急激に充実してきた。また、会議や講演などにおいて音声の録音や、映像の録画を行う動きも広まってきている。これらのコンテンツはネットワークストレージや動画共有サイトなどにアクセスすることで、容易に利用することができる。そして、いまこの瞬間も、コンテンツの量は急速に増加し続けている。これに伴い、これらの大量のコンテンツから視聴したい場面を検索したいという要求が高まっている。しかし、多くのコンテンツは動画像と音声(一部にジャンルなどのタグ情報など)で構成され、テキスト情報を含んでいない。そのため、音声を含むデータに対しては、音声認識技術を適用してコンテンツを検索する方法が有効であり、音声ドキュメント検索(Spoken Document Retrieval: SDR)として精力的な研究が行われてきた。

アメリカ国立標準技術研究所(National Institute of Standards and Technology :NIST) とアメリカ国防総省内の研究部門の一つである防衛高等研究計画局(Defense Advanced Research Projects Agency : DARPA) によって開催された TREC (Text Retrieval Conference) においては、SDR の Track が 1997 年の TREC-6 から取り上げられ、TREC7~9 を経て 2000 年まで行われた。

一方で、音声中の検索語検出(Spoken Term Detection : STD) の研究が近年注目を集めている。STD は、ある特定の検索語(1 個以上の単語からなる言葉)が、音声ドキュメント群中のどのドキュメントのどの位置に含まれているのかを特定するタスクである。このタスクについても、NIST が中心となって 2006 年にテストコレクションが整理されている。

STD の研究の大部分は未知語と音声認識誤りの問題に焦点を合わせている。

まず、音声認識システムの出力を用いるうえで根本的な問題である未知語と音声認識誤りなどを解決するために、音声認識性能を改善させる手法が提案されている。特に複数の音声認識システムを利用することで、音声認識性能を改善させる手法が多く提案されている。

また、音声認識や検索語の検出をサブワードや音韻単位で行う手法が提案されている。例えば、音素認識結果と単語認識結果を組み合わせた手法や文字系列の異なる音声認識システムの出力を組み合わせる手法、接続確率の高い音素列をサブワードとした言語モデルを利用する手法、複数のサブワード言語モデルを利用する手法が提案されている。さらに、情報検索として適したインデックスの構造を利用する手法が提案されている。例えば、サブワードラティスやコンフュージョンネットワーク(Confusion Network : CN)などを利用する STD の技術が提案されている。

本研究では、サブワードベースの CN を使用した STD 手法を提案する。複数の音声認識システムの出力から構成された音素遷移ネットワーク(Phoneme Transition Network : PTN)から検索語を検出するために、編集距離ベースの Dynamic Time Warping (DTW)フレームワークを利用している。

PTN ベースのインデキシングは、音声認識システムの出力から生成される CN に由来している。

単一の音声認識システムの最尤出力である 1-Best 出力と CN を比較した場合、CN は豊富な情報を持っていることから、STD に対して有効な手法である。また、異なる言語モデルと音響モデルを利用した

複数の音声認識システムとその出力を使用することは、音声認識性能を向上させることにおいて非常に効果的であることが知られている。複数の音声認識システムによる単語(または、サブワード系列)出力の適用は、各音声認識システムの特性が異なっているため、良好な音声認識性能を示すことが可能となる。

本研究は、この複数の音声認識システムとその出力を逸早く STD に応用した。

本研究では、同じデコーダに基づく 12 種類の音声認識システムを使用する。使用するモデルは、2 種類の音響モデル(triphone ベースと syllable ベース)と 6 種類の言語モデル(単語ベースとサブワードベース)を用意した。複数の音声認識システムの出力を、効果的に STD 用のインデックスとするために、CN の構造を利用したネットワーク型インデキシングを行った。

日本語の STD テストコレクションに対し、本手法を用いることで、単一の音声認識システムを利用するより、複数の音声認識システムの出力を利用することが、STD の性能を向上させることに有効であることが示された。さらに、複数の音声認識システムの出力をネットワーク型のインデックスとして利用することが STD に有効であることが示された。また、複数の音声認識システムの出力から得られる情報を利用することによって、誤検出を抑制し STD の性能が向上することが示された。

しかし、PTN の冗長性から、多くの誤検出が発生した。複数の音声認識システムの利用は、より良好な認識性能を達成することができるが、多くの誤検出が同時に発生する。

この誤検出を抑制するために、複数の音声認識システムの出力を利用したネットワーク型インデックスを構築する際に得られる情報を、誤検出を抑制するパラメータとして利用した。

これらの誤検出抑制パラメータを、DTW の距離計算式に導入することによって、誤検出が抑制されることが実験結果より示された。とくに、音素を認識した音声認識システムの数である”Voting”を導入することによって、大幅に検索性能が改善された。

誤検出を抑制する手法として、”Voting”などのパラメータを導入することは検索語を検出するうえで有効であった。しかし、検索語の特性として音素長が短い検索語は検出され易く誤検出が多く、また音素長が長い検索語は誤検出が少ないことが判明した。そこで、検索語の音素数に着目し、音素数が少ない検索語に対して誤検出抑制パラメータの適用法を変更した。

また、ネットワーク型インデックスの「複雑さ」に着目し、誤検出を抑制することが可能ではないかと考え、複数の音声認識システムのエントロピーを利用すること検討した。

検討した手法を広く利用されている日本語 STD テストセットの STD タスクと iSTD タスクに適応した評価を行った。評価結果より、エントロピーベースのフィルタリングは、高 Recall 域での STD 性能の向上に有効であることが示された。また、iSTD タスクに有効であるという結果が示された。

音声ドキュメント検索の一分野である STD の目的は、キーワードが発話されている箇所を音声ドキュメント中から特定することである。現在の STD の研究の多くは、検索性能の改善に焦点を合わせており、実環境下での有効性評価の例は少ない。

STD 技術は、様々な用途において有用であり得る。例えば、会議録音音声からターゲットの内容を検索するために使用することができる。STD 技術を用いたいくつか応用分野があるものの、STD の全体的な有用性は、実際の環境で実用的である情報システムで評価されていない。

そこで、電子ノート作成支援システムでのノート見直し作業を対象に、実環境下での STD 技術の有効性評価を行った。

大学講義や講演などでノートを作成する際、講義・講演や話の進行の速さが原因で、書き漏らしや聞

き逃しが起こるという問題があり、後からノートを参照する際に必要な情報が見つからないことがある。しかし、電子ノート作成支援システムに搭載されている機能で音声を録音しておき、STD 技術を利用することで記録した電子ノートから話し手の話した言葉を精度よく検索できるようになれば、このような問題に対応できると考えられる。

そこで STD 使用者と不使用者の電子ノート見直し作業にかかる時間を比較する被験者実験を行うことで、STD の有効性評価を行った。被験者実験では、被験者全員に講義を受講してもらい、電子ノートを作成して頂いた。講義受講から 1 ヶ月後、各自が作成した電子ノートを用いて、電子ノート見直し作業を行って頂いた。このとき、半分の被験者には STD を使用せず解答するよう指示した。被験者実験の結果から、STD 使用者が不使用者に比べ平均的に、試験問題に速く正答したことを確認できた。このことから、電子ノート見直し作業において、STD は有効である可能性があるということが分かった。

本手法は、STD 性能を向上させるために非常に有効であることが、実験結果から示されている。しかし、検索速度は非常に遅い。今後は、実用化のために、DTW の枠組みの下での高速検索アルゴリズムを開発していきたい。

本論文は以下の内容で構成されている。

第 1 章では、STD にいくつかの先行研究を紹介し、私たちは調査の概要について述べる。

第 2 章では、音声中の検索語検出について述べる。

第 3 章では、音声認識システムの概要と、複数の音声認識システムについて記載する。

第 4 章では、複数の音声認識システムの出力を用いたインデキシングと DTW フレームワークを用いた用語検索エンジンについて述べる。また、未知のクエリ用語のための STD 実験についても述べる。

第 5 章および第 6 章では、誤検出制御手法について記載する。

第 7 章では、提案した STD 手法の応用について考察する。

最後に、第 8 章で本研究をまとめる。

# Study on Improvement of Spoken Term Detection Technique

## Abstract

Recently, the number of information technology environments in which numerous audio and multimedia archives such as video archives and digital libraries can be easily used has increased. In particular, there is a rapidly increasing number of archived spoken documents such as broadcast programs, spoken lectures, and meeting recordings, with some of them being accessible through the Internet. Although there is an increasing need to retrieve such spoken information, there are currently no effective retrieval techniques to meet these needs. Therefore, the development of technology for retrieving such information has become increasingly important.

The National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency hosted the Text REtrieval Conference (TREC) Spoken Document Retrieval (SDR) track in the second half of the 1990s, and many studies on SDR of English and Mandarin broadcast news documents were presented. TREC-SDR is an ad-hoc retrieval task that retrieves spoken documents, which are highly relevant to a user query. In 2006, NIST initiated the Spoken Term Detection (STD) project with a pilot evaluation and workshop. STD intends to detect the positions of target spoken terms from audio archives.

STD requires automatic speech recognition for speech-to-text conversion. Therefore, STD is difficult with respect to searching for terms in a vocabulary-free framework because search terms are unknown before using the speech recognizer. Many studies that address STD tasks have been proposed, and most of them focused on the out-of-vocabulary (OOV) and speech recognition error problems. For example, STD techniques that employ entities such as sub-word lattices and confusion networks (CNs) were proposed.

In this study, I propose an STD technique that uses sub-word-based CN. I use a phoneme transition network (PTN)-formed index derived from multiple speech recognizers' 1-best hypothesis and an edit distance-based dynamic time warping (DTW) framework to detect a query term.

The PTN-based indexing originates from the concept of CN being generated from a speech recognizer. CN-based indexing for STD is a powerful indexing method because CN has abundant information when compared with that of the 1-best output of the same speech recognizer. In addition, it is known that many candidates are obtained by one or more speech recognizers that have different language models (LMs) and acoustic models (AMs).

For example, multiple speech recognizers' outputs improves the speech recognition effectively. The application of the characteristics of the word (or sub-word) sequence output by recognizers may enhance STD because these characteristics are different for each speech recognizer. PTNs that are based on multiple speech recognizers' outputs can cover more sub-word sequences of spoken terms. Therefore, the use of multiple speech recognizers may improve STD relative to that of a single recognizer's output. This is the principal idea in this study.

This study employs 10 types of speech recognition systems with the same decoder used for all types. Two types of AMs (triphone and syllable-based Hidden Markov Models (HMMs)) and five types of

LMs (word- and sub-word-based) were prepared. The multiple speech recognizers can generate the PTN-formed index by combining sub-word (phoneme) sequences from the output of these recognizers into a single CN.

I evaluated the PTN-formed index derived from the 10 recognizers' outputs. The experimental result for the Japanese STD test collection showed that the use of the PTN-formed index effectively improved STD compared with that of the CN-formed index, which was derived from the phoneme-based CN comprising the 10-best phoneme sequence outputs from a single speech recognizer.

The Experimental results showed that the PTN-formed index with the DTW framework improved the OOV STD performance when it is compared with that of the simple and CN-formed indices from the single speech recognizer's output.

However, many false detection errors occurred because the PTN-formed index had redundant phonemes that were incorrectly recognized by a few speech recognizers. The use of more speech recognizers can achieve a better recognition performance, but more errors may occur at the same time.

Therefore, I introduce the concept of majority voting to calculate the edit distance between a query term and the index. In addition, a measure of the ambiguity in PTN is adopted into DTW. New parameters based on majority voting and ambiguity are easily derived from PTN and are considered for distance calculation.

I aim to improve STD by effectively utilizing the advantages realized by using multiple speech recognizers. This is an original concept in the field of STD research.

The PTN was very effective at detecting query terms. However, the PTN generates a lot of false detections especially for short query terms. Therefore, I applied two false detection control parameters to the Dynamic Time Warping-based term detection engine. In addition, I changed the search parameters depending on the length of a query term. And I focus on entropy of the PTN-formed index. Entropy is used to filter out false detection candidates in the second pass of the STD process. Our proposed method was evaluated using the Japanese standard test-set for the STD and the iSTD (inexistent STD) tasks. The experimental results of the STD task showed that entropy-based filtering is effective for improving STD at a high-recall range. In addition, entropy-based filtering was also demonstrated to work well for the iSTD task.

The primary goal of spoken term detection (STD), which is a spoken document retrieval technique, is to precisely indicate the locations (utterances) when a queried term is uttered in a large speech corpus. STD techniques may be useful in a variety of applications. For example, they can be used to search target statements from conference minute speeches. However, although there are some application areas for STD techniques, the overall usefulness of STD has not been evaluated in information systems that are of practical use in real environments.

The usefulness of an STD technique in an electronic note-taking support system is assessed through a subjective evaluation experiment. A user of the note-taking support system can write phrases (or figures) electronically while listening to a target speech. At the same time, the system records and stores the entire speech.

Therefore, the user can review notes while listening to the recorded speech. It may also be useful to play back a speech beginning at a time specified by the time location of a note associated with a word the user wishes to focus on. The STD technique is used to indicate the location of the specified term, and it may also be useful for browsing notes associated with a speech.

In the experiment, subjects responded to questions related to a recorded speech while referring to recorded notes and listening to the speech. The subjects' response times for each correct answer were measured. Half of the subjects browsed their notes using the STD technique; the others did not use the STD technique.

The experimental results show that the subjects who used the STD technique answered all questions faster than those who did not use the STD technique. These results indicate that the STD technique works well for browsing the electronic note-taking support system.

In the future, I intend to develop a fast search algorithm under the DTW framework because the Processing speed of our engine is still very slow for practical applications.

The remainder of this paper is organized as follows.

In Chapter 1, I will introduce a few previous studies on STD, and I describe an outline of the study.

In Chapter 2, I describe the search term detection in speech.

In Chapter 3, I describe a speech recognition system and summary of the multiple speech recognition system.

Chapter 4, I explain the types of indices that deal with the study and the term search engine using the DTW framework. Moreover, the STD experiment for OOV query terms is discussed in this chapter.

Chapter 5 and 6 describe a false detection control technique in the term search engine. I discuss the STD experimental results for OOV set using the improved engine.

In Chapter 7, consider the application of the proposed STD method.

Finally, I summarize this study in Chapter 8.

# 目次

<b>第1章 序論</b>	<b>1</b>
1.1 はじめに	1
1.2 関連研究	1
1.3 本研究の概要	3
1.3.1 未知語検索語に頑健な STD 手法	3
1.3.2 未知検索語に頑健な STD 手法の応用	5
1.4 本論文の構成	5
<b>第2章 音声中の検索語検出 [39]</b>	<b>7</b>
2.1 音声ドキュメント検索の概要	7
2.2 音声中の検索語検出の概要	8
2.3 音声中の検索語検出性能の評価	9
2.4 まとめ	11
<b>第3章 複数の音声認識システム</b>	<b>12</b>
3.1 音声認識システム	12
3.1.1 音声認識の原理	13
3.1.2 音声認識エンジン：Julius	13
3.1.3 連続音節認識	14
3.1.4 音声認識結果の評価	14
3.2 形態素解析システム	14
3.3 音響モデル	15
3.4 言語モデル	17
3.4.1 形態素ベース言語モデル：Word-Base Characters (WBC)	20
3.4.2 平仮名形態素ベース言語モデル：Word-Base Hiragana (WBH)	20
3.4.3 文字ベース言語モデル：Character Base (CB)	20
3.4.4 文字系列ベース言語モデル：Bi-Mora (BM)	20
3.4.5 文字系列ベース言語モデル：Character Sequence Base (CSB)	21
3.4.6 疑似連続音節認識用言語モデル：Non	21
3.5 認識用単語辞書	21
3.6 各モデルの学習条件	21
3.7 複数の音声認識システムを利用した音声認識実験と認識性能	22
3.8 複数の音声認識システムを利用することによる STD 性能の改善余地	24



3.9	まとめ	24
<b>第4章</b>	<b>音声中の検索語検出のための検索用インデックス</b>	<b>25</b>
4.1	単一の音声認識システムの出力を利用したインデックス	25
4.1.1	サブワードベースインデックス	25
4.1.2	ネットワーク型インデックス	26
4.1.3	インデックスの種類	27
4.2	複数の音声認識システムの出力を利用したインデックス	30
4.2.1	サブワードベースインデックス	30
4.2.2	ネットワークワーク型インデックス	31
4.2.3	インデックスの種類	33
4.3	インデックスごとの検索性能	33
4.3.1	動的計画法を用いた検索方法	34
4.3.2	複数の音声認識システムを利用する効果	36
4.3.3	インデックスの形態ごとの評価	39
4.3.4	インデックスを構成する仮説数の評価	42
4.3.5	インデックスを構成する音声認識システム数の評価	44
4.4	まとめ	45
<b>第5章</b>	<b>音声中の検索語検出のための検索方法の改善</b>	<b>49</b>
5.1	誤検出抑制パラメータ	49
5.2	編集距離ベースの誤検出抑制パラメータの組合せによる検索性能 (1)	50
5.2.1	誤検出抑制パラメータの導入方法 (1)	50
5.2.2	抑制パラメータの組合せ	51
5.2.3	評価実験	52
5.3	編集距離ベースの誤検出抑制パラメータの組合せによる検索性能 (2)	56
5.3.1	誤検出抑制パラメータの導入方法 (2)	56
5.3.2	抑制パラメータの組合せ	56
5.3.3	評価実験	56
5.4	まとめ	62
<b>第6章</b>	<b>音声中の検索語検出のための誤検出を改善する手法</b>	<b>63</b>
6.1	検索語長の誤検出傾向に着目した検索語の検出方法	63
6.1.1	検索語の音素長による検索性能	63
6.1.2	検索語の音素長に対する遷移コストの適応	64
6.1.3	評価実験	66
6.2	ネットワーク型インデックスの複雑さに着目した検索語の検出方法	69
6.2.1	ネットワーク型インデックスのエントロピー	70
6.2.2	検索語が含まれる区間のエントロピー	71
6.2.3	評価実験	72

6.2.4	最良の STD 性能時のエントロピー	74
6.3	iSTD タスクにおける PTN の性能	77
6.3.1	iSTD タスク	77
6.3.2	評価実験	77
6.4	まとめ	78
<b>第 7 章</b>	<b>音声中の検索誤検出の応用</b>	<b>80</b>
7.1	音声認識の語彙推定への利用	80
7.1.1	音声認識の語彙推定	80
7.1.2	STD を利用した語彙推定	81
7.1.3	評価実験	81
7.2	音声電子ノート作成支援システムへの応用	83
7.2.1	電子ノート作成支援システム	84
7.2.2	電子ノート作成支援システムへの STD の適用	86
7.2.3	被験者実験	86
7.3	まとめ	89
<b>第 8 章</b>	<b>結論</b>	<b>90</b>
	<b>参考文献</b>	<b>94</b>
付 録 A	日本語 STD 用テストコレクションのコア講演用未知語テストセットの 50 検索語	I
付 録 B	NTCIR-9 SpokenDoc タスク formal-run テストセットの 50 クエリ	III
付 録 C	NTCIR-10 SpokenDoc-2 タスク large-size タスク large-size テストセットの 100 クエリ	V
付 録 D	NTCIR-10 SpokenDoc-2 タスク moderate-size タスク moderate-size テストセットの 100 クエリ	IX
付 録 E	NTCIR-10 SpokenDoc-2 タスク iSTD タスク用テストセットの 100 クエリ	XIII
付 録 F	コンフュージョンマトリクススコア	XVII
付 録 G	コンフュージョンマトリックススコアベースの検索性能	XXI
G.1	コンフュージョンマトリックススコアの導入方法	XXI
G.2	評価実験	XXII
付 録 H	単一の音声認識システムの検索性能	XXIV

付 録 I	既知検索語の検索性能	XXXI
I.1	検索性能の比較実験条件 . . . . .	XXXI
I.2	検索性能の比較結果 . . . . .	XXXII
I.3	考察 . . . . .	XXXII

# 目 次

1.1	提案する STD の流れ	4
3.1	音声認識システムの概要	12
3.2	状態系列と出力信号	16
4.1	SCN のイメージと構築例	27
4.2	PCN のイメージと構築例	28
4.3	DP の傾斜制限と遷移コストの定義	29
4.4	PCN を用いた STD の例	30
4.5	複数の PCN を用いた STD の例	31
4.6	STN のイメージと構築例	32
4.7	PTN のイメージと構築例	33
4.8	PTN を用いた STD の例	34
4.9	サブワードベースインデックスから DP を用いた検索語の検出例	35
4.10	ネットワーク型インデックスから DP を用いた検索語の検出例	36
4.11	単一の音声認識システムの 1-Best 出力を利用したサブワードベースインデックスの検索性能の比較	38
4.12	10 個の音声認識結果を利用したサブワードベースインデックスの検索性能の比較	39
4.13	単一の音声認識システムの出力を利用したインデックスの検索性能の比較	41
4.14	10 種類の音声認識システムの出力を利用したインデックスの検索性能の比較	41
4.15	10 個の仮説数を利用したインデックスの検索性能の比較	43
4.16	100 個の仮説数を利用したインデックスの検索性能の比較	44
4.17	サブワードベースインデックスの検索性能の比較	47
4.18	nPCNs の検索性能の比較	47
4.19	PTN の検索性能の比較	48
5.1	1 種類の誤検出抑制パラメータを導入した検索性能の比較	53
5.2	CM スコアを導入した検索性能の比較	54
5.3	複数の誤検出抑制パラメータを導入した検索性能の比較	55
5.4	1 種類の誤検出抑制パラメータを導入した検索性能の比較	59
5.5	CM スコアを導入した検索性能の比較	59
5.6	Voting に CM スコアを導入した検索性能の比較	60

5.7	ArcWidth に CM スコアを導入した検索性能の比較 . . . . .	60
5.8	Voting と ArcWidth に CM スコアを導入した検索性能の比較 . . . . .	61
5.9	複数の誤検出抑制パラメータを導入した検索性能の比較 . . . . .	61
6.1	検索語の音素長に応じたパラメータ適応による検索性能の比較 (Recall-Precision カーブ) . . . . .	67
6.2	音素長が 10 未満の検索語に対する検索語の音素長に応じたパラメータ適応による検索性能の比較 (Recall-Precision カーブ) . . . . .	69
6.3	音素長が 10 未満の検索語に対する検索語の音素長に応じたパラメータ適応による検索性能の比較 (Recall-Precision カーブ) . . . . .	70
6.4	PTN のエントロピーのイメージ . . . . .	71
6.5	PTN のエントロピーのイメージ (検索語検出区間) . . . . .	72
6.6	エントロピーを導入した際の検索性能の比較 (Recall-Precision カーブ) . . . . .	73
6.7	STD の検出コストとエントロピーの関係図 . . . . .	74
6.8	誤検出を含む STD の検出コストとエントロピーの関係図 . . . . .	75
7.1	PTN による STD を利用した語彙推定の流れ . . . . .	82
7.2	電子ノート作成支援システムの構成と利用概要 . . . . .	84
7.3	電子ノート作成支援システムのユーザ端末画面イメージと使用例 . . . . .	85
7.4	STD による検索結果の表示例 . . . . .	87
G.1	距離計算尺度による検索性能の比較 . . . . .	XXIII
H.1	WBC/Tri の検索性能 . . . . .	XXIV
H.2	WBH/Tri の検索性能 . . . . .	XXVI
H.3	CB/Tri の検索性能 . . . . .	XXVI
H.4	BM/Tri の検索性能 . . . . .	XXVII
H.5	Non/Tri の検索性能 . . . . .	XXVII
H.6	WBC/Syl の検索性能 . . . . .	XXVIII
H.7	WBH/Syl の検索性能 . . . . .	XXVIII
H.8	CB/Syl の検索性能 . . . . .	XXIX
H.9	BM/Syl の検索性能 . . . . .	XXIX
H.10	Non/Syl の検索性能 . . . . .	XXX
I.1	単一の音声認識システムと提案手法の比較 . . . . .	XXXIII
I.2	10 個の音声認識結果を用いた場合の検索性能の比較 . . . . .	XXXIV

# 表 目 次

2.1	日本語 STD 用テストコレクション コア講演用未知語テストセットの内訳	10
3.1	認識用単語辞書の語彙数	22
3.2	CSJ コア講演音声の平均単語認識率 [%]	23
3.3	CSJ コア講演音声の平均音節認識率 [%]	23
3.4	10 種類の音声認識システムの言語モデルの組み合わせ	23
4.1	単一の音声認識システムの出力を利用したインデックスの種類	30
4.2	STN や PTN を構築する際に用いる音声認識システムの種類と N-Best 出力の組合せ例	32
4.3	複数の音声認識システムの出力を利用したインデックスの種類	34
4.4	複数の音声認識システムを利用する効果の比較実験に用いたインデックスの種類	37
4.5	表 4.4 に示すインデックスごとの最大 F-measure と ATWV	37
4.6	インデックスの形態による効果の比較実験に用いたインデックスの種類	40
4.7	表 4.6 に示すインデックスごとの最大 F-measure と ATWV	40
4.8	インデックスを構成する仮説数による効果の比較実験に用いたインデックスの種類	42
4.9	表 4.8 に示すインデックスごとの最大 F-measure と ATWV	42
4.10	インデックスを構成する音声認識システム数による効果の比較実験に用いたインデックスの種類	45
4.11	表 4.10 に示すインデックスごとの最大 F-measure と ATWV	46
5.1	誤検出抑制パラメータを導入する PTN の構成内容	52
5.2	誤検出抑制パラメータの組み合わせ (1)	52
5.3	誤検出抑制パラメータの組み合わせによる検索性能の比較 (1)	52
5.4	誤検出抑制パラメータの組み合わせ (2)	57
5.5	誤検出抑制パラメータの組み合わせによる検索性能の比較 2	58
6.1	“Only EditDist” における音素長別の STD 性能	64
6.2	“Only EditDist” における音素長別の STD 性能	64
6.3	探索パラメータの組み合わせ	67
6.4	検索語の音素長に応じたパラメータ適応による検索性能の比較 (F-measure と MAP)	68

6.5	検索語が存在する区間の PTN エントロピー	72
6.6	最大の検出性能 (F-measure) 時の PTN のエントロピー	76
6.7	PTN を用いた iSTD タスク性能	78
7.1	語彙推定による音声認識率の比較結果	83
7.2	実験で使用した STD の性能	87
7.3	STD 使用者と不使用者の正答時間の平均値と標準偏差 [分' 秒"]	88
7.4	STD 使用者と不使用者の設問ごとの正答時間の平均値 [分' 秒"]	88
A.1	コア講演用未知語テストセットの 50 クエリ (1)	I
A.2	コア講演用未知語テストセットの 50 クエリ (2)	II
B.1	formal-run テストセットの 50 クエリ (1)	III
B.2	formal-run テストセットの 50 クエリ (2)	IV
C.1	large-size テストセットの 100 クエリ (1)	V
C.2	large-size テストセットの 100 クエリ (2)	VI
C.3	large-size テストセットの 100 クエリ (3)	VII
C.4	large-size テストセットの 100 クエリ (4)	VIII
D.1	moderate-size テストセットの 100 クエリ (1)	IX
D.2	moderate-size テストセットの 100 クエリ (2)	X
D.3	moderate-size テストセットの 100 クエリ (3)	XI
D.4	moderate-size テストセットの 100 クエリ (4)	XII
E.1	iSTD 用テストセットの 100 クエリ (1)	XIII
E.2	iSTD 用テストセットの 100 クエリ (2)	XIV
E.3	iSTD 用テストセットの 100 クエリ (3)	XV
E.4	iSTD 用テストセットの 100 クエリ (4)	XVI
F.1	ある音素が正解している確率	XVIII
F.2	ある音素が挿入している確率	XIX
F.3	ある音素が脱落している確率	XX
G.1	コンフュージョンマトリックススコアベースの距離計算を行う PTN の 構成内容	XXII
G.2	距離計算尺度による検索性能の比較	XXII
H.1	単一の音声認識システムの検索性能の比較	XXV
I.1	既知検索語の検索性能の比較実験に用いたインデックスの種類	XXXI
I.2	既知検索語の検索性能の比較	XXXII

# 第1章 序論

## 1.1 はじめに

近年、マルチメディアデータの生成・編集環境の普及、ストレージの大容量化、ネットワークインフラの充実により、動画コンテンツに代表される音声やマルチメディアコンテンツが急激に充実してきた。また、会議や講演などにおいて音声の録音や、映像の録画を行う動きも広まってきている。これらのコンテンツはネットワークストレージや動画共有サイトなどにアクセスすることで、容易に利用することができる。そして、いまこの瞬間も、コンテンツの量は急速に増加し続けている。これに伴い、これらの大量のコンテンツから視聴したい場面を検索したいという要求が高まっている。しかし、多くのコンテンツは動画像と音声(一部にジャンルなどのタグ情報など)で構成され、テキスト情報を含んでいない。そのため、音声を含むデータに対しては、音声認識技術を適用してコンテンツを検索する方法が有効であり、音声ドキュメント検索として精力的な研究が行われてきた。

音声ドキュメント検索の一分野である音声中の検索語検出 (Spoken Term Detection : STD) の目的は、検索語 (1 個以上の単語からなる言葉) が話されている箇所を音声ドキュメント中から特定することにある。一般的な STD の手法は、音声認識システムとその出力を利用するものである。この場合、音声認識システムが認識できない語 (これを未知語と呼ぶ) や音声認識性能が低い場合には、単純な文字列検索による検索語の検出は困難となる。本研究では、この検索語が未知語の場合に焦点を当て、未知検索語に頑健な STD 手法を提案することを目的とする。さらに、本研究で提案した未知検索語に頑健な STD 手法の応用について考察する。

## 1.2 関連研究

音声からキーワードを抽出する技術については、これまでに多くの研究成果が報告されている。

音声から直接任意のキーワード (本研究での検索語) を抽出する技術はキーワードスポッティングと呼称されている。これは、音声認識が未熟であった頃に任意のキーワードだけでも認識が可能となるように研究されてきた技術である。このキーワードスポッティングは、大語彙連続音声認識と呼ばれる音声認識手法や、近年計算機の性能が大幅に向上したことにより大量の学習データを用いることが可能となったため技術として衰退した。



しかし、音声データから任意のキーワードが離されている区間を特定するという要求が高まるにつれ、音声認識を用いたキーワードスポッティングが注目されることになった。この音声認識を用いたキーワードスポッティングが、音声中の検索語検出 (Spoken Term Detection : STD) と呼ばれる分野として研究されることになった。

STD に取り組む研究は近年盛んに研究されており、世界中で取り組まれ多くの研究成果が国際学会などにおいて発表されている [1][2][3]。

また、国内においても STD に取り組む研究が多く行われている [4][5][6][7]。

特に 2010 年に開催された INTERSPEECH2010 では、音声ドキュメント検索に関するスペシャルセッションが組まれており、15 件以上の STD に関する発表が行われている [8][9][10][11][12][13][14][15][16][17][18][19]。

STD の研究の大部分は未知語と音声認識誤りの問題に焦点を合わせている。

まず、音声認識システムの出力を用いるうえで根本的な問題である未知語と音声認識誤りなどを解決するために、音声認識性能を改善させる手法が提案されている。特に複数の音声認識システムを利用することで、音声認識性能を改善させる手法が多く提案されている [8][20][21]。

また、音声認識や検索語の検出をサブワードや音韻単位で行う手法が提案されている。例えば、音素認識結果と単語認識結果を組み合わせた手法 [22] や文字系列の異なる音声認識システムの出力を組み合わせて利用する手法 [23]、接続確率の高い音素列をサブワードとした言語モデルを利用する手法 [24]、複数のサブワード言語モデルを利用する手法 [25] が提案されている。

さらに、情報検索として適したインデックスの構造を利用する手法が提案されている。例えば、サブワードラティスやコンフュージョンネットワーク (Confusion Network : CN)[26] などを利用する STD の技術が提案されている [9][27][28][29][30]。

近年の日本語 STD の研究は、検索性能の向上 [31][32] と高速化 [32][33][34][35][36] が主となっている。

伊藤ら [31] は、時間長等が異なる複数のサブワードで音声認識を行い、局所距離にサブワード間の音響距離を利用した各認識結果からの検索結果を統合することで、検索性能の向上を実現している。

神田ら [32] は、まず一定の時間フレームごとに特徴量を切り出し、音響モデルの各状態の音響スコアを算出し、スコアに基づき時間同期の音素認識を行った。この音素認識結果を音素 N-gram インデックスとして登録し、検索語 (クエリ) の発話位置候補を荒く検索した後に、先述した音響スコアによってリスコアリングすることで検索性能の向上と高速化を実現している。

岩見ら [33] は、複数の音声認識システムで音節認識を行い、認識結果を N-gram インデックスとして構築し、辞書順にソートしておくことで高速化を実現している。また、音声認識誤りに対しては、複数候補やダミー音節、音響距離を用いて対処することにより、検索精度を改善させた。

勝浦ら [34] は、Suffix Array を用いた高速キーワード検索手法を提案しており、クエリの分割や反復深化的探索等の技術を複合的に利用することで、高速化を実現して

いる.

斎藤ら [35] は, まずすべての音節 bigram, trigram に対して照合を行っておき, 事前検索結果として照合結果を保存しておく. 次に, クエリに含まれる音節 bigram, trigram から事前検索結果を利用して発話位置候補区間を絞り込み, 厳密に照合する候補を削減することで高速化を実現している.

金子ら [36] は, クエリの音節列と検索対象音声ドキュメントの音節列の距離を音節間距離行列として構築し, 音節間距離を画素濃度とみなすことにより, STD を画像中の直線検出タスクととらえることで高速化を実現している.

本研究は STD の研究の中でも, 検索性能の向上を目的とした位置づけとなっている. また, 本研究が関連研究と異なる点として以下の 3 点が挙げられる.

1 つ目は, 複数の音声認識システムを利用することである. 形態の異なる複数の音声認識システムを利用することにより, より多くの音素を網羅できると考えた. また, 複数の音声認識システムの出力を CN を利用してネットワーク型のインデックスとして統合した PTN により, 複数の音声認識システムの出力を効率よく表現することが可能となり, インデックスのサイズを抑えることが可能となっている [37]. さらに, PTN が持つ音素の認識数等の情報を利用することで, 外部の情報を必要とすることなく検索精度を向上できると考えた.

2 つ目は, 単純な検索アルゴリズムで高い検索精度を実現可能な点である. 本研究では, 用語検索エンジンに単純な文字列検索アルゴリズムである動的計画 (Dynamic Programming : DP) 法を用いているが, PTN を構築する際に得られる情報を利用することにより, 高い検索精度の実現が可能と考えた.

3 つ目は, STD に入力されるクエリに着目した点である. 未知語のクエリに対応するための検索対象の音声ドキュメントに対する音声認識方法やインデキシング方法に関する研究は数多くあるが, クエリの長さや複雑さ (音声認識の困難さ) を考慮した用語検索エンジンに関する研究は少ない. 本研究では, 前述の用語検索エンジンにこれらの尺度を導入することによってクエリに応じた検索を行い, 検索精度を改善させる.

## 1.3 本研究の概要

本研究では, 検索語が未知語の場合に焦点を当て, 未知検索語に頑健な STD 手法を提案する. さらに, 本研究で提案した未知検索語に頑健な STD 手法の応用について考察する.

### 1.3.1 未知語検索語に頑健な STD 手法

本研究では, 複数の音声認識システムの出力を利用することによって STD 性能を向上させる手法について提案する.

提案する STD の流れを図 1.1 に示す.

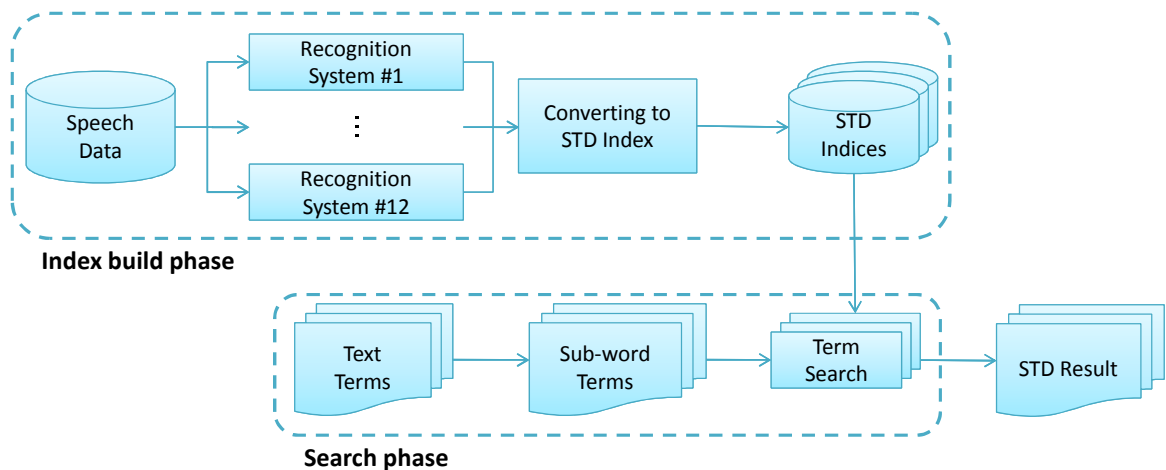


図 1.1: 提案する STD の流れ

本研究が典型的な STD 技術と異なる点は、複数の音声認識システムを使用することにある。複数の音声認識システムの出力を基に、ネットワーク型のインデックスを構築し検索語の検出を行う。

本研究における STD は、検索語を音韻 (音素または音節) 単位で扱う。

本研究では、同一のデコーダを使用した 12 種類の音声認識システムを利用する。使用するモデルは、2 種類の音響モデル (triphone ベースと syllable ベース) と 6 種類の言語モデル (単語ベースとサブワードベース) を用意した。

複数の音声認識システムとその出力を使用することは、音声認識性能を向上させることにおいて非常に効果的であることが知られている。例えば、Fiscus[20] は単語投票方式を採用する ROVER (Recognizer Output Voting Error Reduction) 法を提案している。また、宇津呂ら [21] は音声認識性能を向上させるために、サポートベクタマシン (Support Vector Machine : SVM) を使用することによって、複数の音声認識システムの出力を結合するための技術を見出した。複数の音声認識システムによる単語 (または、サブワード系列) 出力の適用は、各音声認識システムの特性が異なっているため、良い音声認識性能を示すことが可能となる。本研究は、この複数の音声認識システムとその出力を逸早く STD に応用した。

さらに、複数の音声認識システムの出力を、効果的に STD 用のインデックスとするために、CN の構造を利用したネットワーク型インデキシングを行った。

本手法を用いることで、単一の音声認識システムを利用するより、複数の音声認識システムの出力を利用することが、STD の性能を向上させることに有効であることが示された。さらに、複数の音声認識システムの出力をネットワーク型のインデックスとして利用することが STD に有効であることが示された。また、複数の音声認識システムの出力から得られる情報を利用することによって、誤検出を抑制し STD の性能が向上することが示された。

しかし、調査の結果、主に 2 つの要因で誤検出が増加していることが判明した。1 つ

目は、STD における探索パラメータが経験則に基づいて静的に設定されており、クエリによって動的に変更できない点である。2つ目は、PTN の表現力の高さが悪影響を及ぼしていることである。特に、音素数の少ないクエリを入力した場合に誤検出が頻発してしまい、高い検索精度が得られないことが判明した。

そこで、このような語検出の抑制手法を検討し、以下の2つの手法を検討し、検索精度の改善を図った。

1つ目は、音素数の少ないクエリを焦点として、探索パラメータをクエリの音素数に基づいて調整することで、STD 性能を向上させる手法を検討した。

2つ目は、ネットワーク型インデックスのエントロピーを利用した手法である。ネットワーク型インデックスの複雑さに着目し、そのエントロピーを分析した。分析結果を示すとともに、STD の検出候補が持つエントロピーを利用した検出候補のフィルタリング手法を検討した。また、“inexistent Spoken Term Detection (iSTD)” タスク<sup>1</sup>[37]において、ネットワーク型インデックスのエントロピーを利用した iSTD 手法について述べる。

評価実験の結果、クエリの音素数に基づいて探索パラメータを調整することが STD 性能を向上させることに有効であることが示された。また、STD の検出候補が持つエントロピーを利用し、検出候補のフィルタリングを行うことで、閾値を緩くした際の誤検出を大幅に抑えることが可能となった。また、iSTD タスクにおいては、ネットワーク型インデックスのエントロピーを iSTD スコアに加味することで、iSTD の性能を向上させることに有効であることが示された。

### 1.3.2 未知検索語に頑健な STD 手法の応用

本研究で提案した STD 手法を用いることで、STD の性能が向上することが示された。この STD 手法が応用することが可能であることを考察する。

本論文では、電子ノート作成支援システム [38] に提案した STD 手法を利用した。また、大語彙連続認識システムで用いる言語モデルの学習データ選別や、認識単語の選別に用いることで、音声認識性能を向上させることが可能かを考察する。

## 1.4 本論文の構成

本論文は8章から構成されている。

本章に続く第2章では、音声情報検索の基本的な概念や、その中における STD の位置づけ、検索性能の評価方法など、STD の基本的な知識について述べる。

第3章では、音声認識システムの概要と、複数の音声認識システムについて記載する。複数の音声認識システムによる単語(または、サブワード系列)出力の適用は、各音声認識システムの特徴が異なっているため、良好な音声認識性能を示すことが可能

---

<sup>1</sup>ある与えられたクエリが音声アーカイブ内に存在する／しないを検査し、その結果を返すタスク。

となる．本研究は，この複数の音声認識システムとその出力を逸早く STD に応用したものである．

第4章では，複数の音声認識システムの出力を用いたインデキシングと DTW フレームワークを用いた用語検索エンジンについて述べる．また，未知のクエリ用語のための STD 実験についても述べる．本研究では，サブワードベースの CN を使用した STD 手法を提案する．複数の音声認識システムの出力から構成された音素遷移ネットワーク (Phoneme Transition Network : PTN) から検索語を検出するために，編集距離ベースの DTW フレームワークを利用している．PTN ベースのインデキシングは，音声認識システムの出力から生成される CN に由来している．日本語の STD テストコレクションに対し，本手法を用いることで，単一の音声認識システムを利用するより，複数の音声認識システムの出力を利用することが，STD の性能を向上させることに有効であることが示された．さらに，複数の音声認識システムの出力をネットワーク型のインデックスとして利用することが STD に有効であることが示された．

第5章および第6章では，誤検出制御手法について記載する．複数の音声認識システムの利用は，より良好な認識性能を達成することができるが，多くの誤検出が同時に発生する．この誤検出を抑制するために，複数の音声認識システムの出力を利用したネットワーク型インデックスを構築する際に得られる情報を，誤検出を抑制するパラメータとして利用した．これらの誤検出抑制パラメータを，DTW の距離計算式に導入することによって，誤検出が抑制されることが実験結果より示された．

しかし，検索語の特性として音素長が短い検索語は検出され易く誤検出が多く，また音素長が長い検索語は誤検出が少ないことが判明した．そこで，検索語の音素長に着目し，音素長が短い検索語に対して誤検出抑制パラメータの適用法を変更した．

また，ネットワーク型インデックスの「複雑さ」に着目し，誤検出を抑制することが可能ではないかと考え，複数の音声認識システムのエントロピーを利用すること検討した．検討した手法を日本語 STD テストセットの STD タスクと iSTD タスクに適応した評価を行ったところ，エントロピーベースのフィルタリングは，高 Recall 域での STD 性能の向上に有効であることが示された．

第7章では，提案した STD 手法の応用について考察する．STD 技術を用いたいくつかの応用分野があるものの，STD の全体的な有用性は，実際の環境で実用的である情報システムで評価されていない．そこで，電子ノート作成支援システムでのノート見直し作業を対象に，実環境下での STD 技術の有効性評価を行った．STD 使用者と不使用者の電子ノート見直し作業にかかる時間を比較する被験者実験を行うことで，STD の有効性評価を行った．被験者実験の結果から，STD 使用者が不使用者に比べ平均的に，試験問題に速く正答したことを確認できた．このことから，電子ノート見直し作業において，STD は有効である可能性があるということが示された．

最後に，第8章において本研究を総括し，今後の課題について述べている．

## 第2章 音声中の検索語検出 [39]

STD とは音声ドキュメント検索の一分野であり，音声ドキュメント検索とは情報検索 [40] の一分野である．情報検索とは，コンピュータを用いて大量のデータ群の中からユーザの要求に合致した情報を見つけ出すことである．

本章では，STD の音声ドキュメント検索分野に対する位置づけや検索性能の評価方法について述べる．

### 2.1 音声ドキュメント検索の概要

本論文で扱う情報検索は，検索対象のデータ群として音声ドキュメント集合，ユーザの要求として検索語(クエリ)を用いる音声ドキュメント検索である．音声ドキュメント検索においては，ニュース記事や講義音声，ビデオデータなど音声を含むデータを音声ドキュメントと呼び，複数あるいは大量の音声ドキュメントがある中で，検索要求(クエリ)に関連する内容を持つ音声ドキュメントを特定することを，アドホック (ad-hoc) 音声ドキュメント検索，あるいは単に音声ドキュメント検索 (Spoken Document Retrieval : SDR) や音声内容検索 (Spoken Content Retrieval : SCR) と呼ぶ．

SDR の基本的な枠組みでは，まず音声ドキュメント群を単語ベースにて音声認識を行い，その認識結果である単語系列に対してテキスト検索 [40] の技術を用いてどの音声ドキュメントかを特定する．

現在，音声ドキュメント検索は大きく分けて SDR と STD の 2 分野があり，さらにタスクごとに細分化されている．

アメリカ国立標準技術研究所 (National Institute of Standards and Technology : NIST) とアメリカ国防総省内の研究部門の一つである防衛高等研究計画局 (Defense Advanced Research Projects Agency : DARPA) によって開催された TREC(Text Retrieval Conference) においては，SDR の Track が 1997 年の TREC-6 から取り上げられ，TREC7～9 を経て 2000 年まで行われた [41]．これを機に，海外では音声ドキュメント検索に関しての研究，特に英語と標準中国語のニュースドキュメント検索に対する多くの研究成果が発表されるなど，音声ドキュメントに関しての研究が推進・活性化された．

日本においても，情報処理学会音声言語情報処理研究会 (SIG-SLP) において，国内の音声ドキュメント処理研究の推進・活性化を目的として，2006 年に音声ドキュメント処理ワーキンググループ (Spoken Document Processing Working Group : SDPWG)

を立ち上げ、これまでに SDR 評価用テストコレクションを構築・公開している [42].

## 2.2 音声中の検索語検出の概要

STD は、ある特定の検索語 (1 個以上の単語からなる言葉) が、音声ドキュメント群中のどのドキュメントのどの位置に含まれているのかを特定するタスクである. この STD は、以前からワードスポッティングという形で多くの研究が行われてきた. ワードスポッティングとは、あらかじめ定められた単語 (単語辞書) のみを音声から抽出する技術である.

従来のワードスポッティングでは、音響的な特徴に加えて文法的な制約やあらかじめ定められた単語 (単語辞書) のみを音声から抽出するといった方法が主流であった. このワードスポッティングも多くの手法が提案されている [43].

アドホック音声ドキュメント検索により、クエリと関連あるドキュメント群が特定されたとしても、その結果は一覧性や確実性に欠け、最上位のドキュメントでさえ、あるキーワードが含まれているかは実際に視聴しないことには確認できない. 検索語が話されている箇所を音声ドキュメント群中から特定したいというニーズは音声ドキュメント検索において不可避である.

また、検索語が音声認識システムにおける未知語になる場合は多く [44], 未知語の検索機能は不可欠である. このような背景もあり、NIST では 2006 年に STD を新たなテーマとして設定 [45] し、STD の試験評価とワークショップを行っている.

このような状況を踏まえ、SDPWG は日本語 STD 用テストコレクションの構築を 2008 年度から開始し、2010 年 5 月に公開した [19]. この日本語 STD 用テストコレクションは『日本語話し言葉コーパス (Corpus of Spontaneous Japanese : CSJ)』<sup>2</sup>[46] を対象としたテストセットとなっている. CSJ は実際の学会などの講演音声と模擬講演、朗読音声などから構成されており、全部で 3,302 の音声データが収録されている.

このテストコレクションの構築・公開に伴って、日本語 STD に関しての研究が推進・活性化されており、国内や国外の学会において多くの研究発表が行われている.

日本語音声ドキュメント処理研究推進の場として、NTCIR<sup>3</sup>においても音声ドキュメント処理のタスクが設定された. 2011 年に開催された NTCIR-9 においては、SpokenDoc のサブタスクとして STD のタスクが設定され、多くの研究が発表された [47]. また、2013 年に開催された NTCIR-10 においては、STD のタスクに加えて iSTD タスクが設

---

<sup>2</sup>『日本語話し言葉コーパス』は、東京工業大学の古井貞熙 (サダオキ) 教授を総括責任者として、独立行政法人国立国語研究所と独立行政法人通信総合研究所が推進してきている文科省科学技術振興調整費開放的融合研究制度研究課題「話し言葉の言語的・パラ言語的構造の解析に基づく『話し言葉工学』の構築」プロジェクト (1999-2003) の一環として構築されたものである. このコーパスは日本語の自発音声を大量にあつめて多くの研究用情報を付加した話し言葉研究用のデータベースである.『日本語話し言葉コーパス』には全体で約 660 時間の自発音声 (語数にして約 700 万語) が格納されている. 音声信号はヘッドセット式コンデンサマイクロホンと DAT によって収録したものを 16 ビット、16KHz にダウンサンプリングして格納してある. 音声は、本コーパスのために考案された特別な正書法に従って書き起こされており、漢字仮名混じりと仮名のみの 2 種類の書き起こしテキストとして提供されている. また、書き起こしテキストには品詞分析が施されている. この分析もまた、長短 2 種類の単位による結果がそれぞれ提供される.

定された [37]. この iSTD タスクは音声ドキュメント内に存在していない単語を、どれだけ検出しなかったのかを評価するタスクである. この NTCIR の STD タスク, iSTD タスクにおいて多くの STD 手法が競われるなど, 現在においても音声中の検索語検出は盛んに研究されている [48][49][50][51][52][53][54][55][56][57][58][59].

## 2.3 音声中の検索語検出性能の評価

検索性能を評価する際, 音声認識では音声ドキュメントの「質」(発話の丁寧さや, 録音の精度など)に主に影響されるが, 音声ドキュメント検索では音声ドキュメントの「質」だけでなく「長さ」や「正解箇所の数」にも影響される. 例を挙げると, 1 時間の音声ドキュメント群から検索する場合と, 10 時間の音声ドキュメント群から検索する場合や正解が全く含まれていない音声ドキュメント群から検索する場合では, これらの検索性能の比較は困難である. このため, 音声ドキュメント検索では共通の音声ドキュメント群やクエリ (STD においては検索語), 正解位置に基づいて評価が行われることが望ましい.

現在, 音声ドキュメント検索の評価では, 参考文献 [19] に示されるような評価用テストコレクションや評価尺度が用いられている.

日本語 STD 用テストコレクションは, CSJ の音声データの内, 学会講演 987 講演, 模擬講演 1,715 講演の計 2,702 講演, 約 604 時間の音声ドキュメントを検索対象データとする全講演テストセットと, 2,702 講演の内, 「コア」と称する 177 講演 (学会講演 70, 模擬講演 107) 約 39 時間の音声ドキュメントを検索対象データとするコア講演セットが存在する.

日本語 STD 用テストコレクションの内, 本研究ではコア講演用未知語テストセットを用いて, STD 性能の評価を行っている. コア講演用未知語テストセットの内訳を表 2.1 に示す.

本研究では, 評価尺度に Recall-Precision カーブ, F-measure, MAP (Mean Average Precision), MRP (Mean R-Precision) を用いている. また, 海外での研究との比較のために ATWV (Actual Term Weighted Value)[45] を一部で用いている. 以下に, 評価式を示す.

---

<sup>3</sup>エンティサイル (NII Testbeds and Community for Information access Research : NTCIR) は, 情報検索, 質問応答, 要約, テキストマイニング, 機械翻訳など膨大な情報の中から所望の情報にアクセスし, 情報の理解や活用を支援する技術の大規模な評価基盤を国内外の多数の研究者が共有し, その共通基盤の上でそれぞれの研究を進め, 検証, 比較評価し, 相互に学びあうフォーラムを形成するプロジェクトである. 1997 年末にプロジェクトが開始され, より豊かな情報アクセス技術の実現と未来価値創成を標榜し活動が行われている. NTCIR ワークショップは, 1998 年から概ね 1 年半を 1 サイクルとし, 毎回いくつかのタスク (研究部門) を選定し, 国内外の 100~130 の研究団体が協力し研究基盤として新しい手法の有効性の検証とベンチマークのためのデータセットを構築し, 同じ基盤の上で相互比較をし, 協調と切磋琢磨をしながら研究を集中的に推進する活動である. 各サイクルの最後には, NTCIR カンファレンスを国際会議として開催している. NTCIR カンファレンスでは, タスク参加チームの研究成果や比較評価によって得られた知見が発表されている. また, 情報アクセス技術の評価手法に関する研究論文を広く一般から公募し, 発表する場として EVIA(International Workshop on Evaluating Information Access : EVIA) を連続開催している. プロジェクトを通じて構築した, 正解データ付きの実験用データセット (テストコレクションと呼称される), リソースやツールの多くは研究目的で公開されている.



表 2.1: 日本語 STD 用テストコレクション コア講演用未知語テストセットの内訳

検索対象音声ドキュメント	検索語種	正解位置
CSJ コア講演音声 (177 講演, 約 39 時間)	50	234

$$Recall(t) = \frac{N_{corr}(t)}{N_{true}} \quad (2.1)$$

$$Precision(t) = \frac{N_{corr}(t)}{N_{corr}(t) + N_{spurious}(t)} \quad (2.2)$$

$$F-measure(t) = \frac{2 \times Recall(t) \times Precision(t)}{Recall(t) + Precision(t)} \quad (2.3)$$

$$MAP = \frac{1}{Q} \sum_{q=1}^Q AveragePrecision(q) \quad (2.4)$$

$$AveragePrecision(q) = \frac{1}{N_{true}(q)} \sum_{k=1}^R \delta_k \times Precision_{rank}(k) \quad (2.5)$$

$$Precision_{rank}(k) = \frac{\text{第 } k \text{ 位までに得られた正解数}}{k} \quad (2.6)$$

$$MRP = \frac{1}{Q} \sum_{q=1}^Q R-Precision(q) \quad (2.7)$$

$$R-Precision(q) = \frac{N_{true}(q) \text{ 位までに得られた正解数}}{N_{true}(q)} \quad (2.8)$$

$$ATWV(q) = 1 - (P_{miss}(q) + \beta P_{fa}(q)) \quad (2.9)$$

$$P_{miss}(q) = 1 - Recall(q), P_{fa}(q) = \frac{N_{spurious}(q)}{Total - N_{true}(q)} \quad (2.10)$$

$t$  は閾値を表しており, Recall-Precision カーブは閾値ごとの評価値によって描かれる.  
 $q$  は検索語を表しており, 検索語ごとに算出されることを示している. また,  $Q$  はテストセットの検索語数を表す.

$N_{corr}$  は検出された適合検索語の出現数を表し,  $N_{spurious}$  は誤検出された検索語の出現数を表す.  $N_{true}$  は音声データ中に本来存在する検索語の出現総数を表す.

Recall-Precision カーブと F-measure は全検索語の合計検索結果から算出したものを用いている.

式 (2.5) の  $R$  は最後に正解が表れた順位を表し,  $\delta_k$  は  $k$  位の区間が正解であれば 1, 不正解であれば 0 となる. 式 (2.6) は第  $k$  位の候補における Precision を示す. MAP は

Average Precision(AP)を全検索語で平均したものであり、APは正解出現時のPrecisionを平均したものである。

MRPはR-Precision(RP)を全検索語で平均したものであり、RPは検索結果をスコア順にソートし、上位から検索語に対する正解数までの検索結果のPrecisionである。

式(2.9)の $Total$ は音声データの持続時間(秒)を表し、158,400秒を設定した。 $\beta$ は本稿では144を設定している。最終的なATWVは、各検索語に対する評価値の平均となる。

## 2.4 まとめ

本章では、STDの音声ドキュメント検索分野に対する位置づけや検索性能の評価方法について述べた。音声ドキュメント検索においては、経緯や関連研究を踏まえて、その概要について述べた。

本研究では、ここで紹介した日本語STD用テストコレクションのうち、コア講演用未知語テストセットを用いる。また、評価尺度としてRecall-Precisionカーブ、F-measure, MAP, MRPと、一部でATWVを用いる。

本章で述べたSTDの知識を前提に、第4章から本研究で行った実験について述べる。

次章では、本研究で提案するSTD性能改善に用いた複数の音声認識システムについて、音声認識の原理とともに述べる。

## 第3章 複数の音声認識システム

本章では，提案する STD 性能改善に用いた複数の音声認識システムについて述べる．また，複数の音声認識システムを構築する上で重要な技術である音声認識技術と形態素解析について簡単に述べ [60]，複数の音声認識システムによる音声認識実験の結果について述べる．

音声認識システムは同一の音声認識エンジンを用い，そこで用いるモデルを変更することによって複数の音声認識システムを構築した．

音声認識で用いるモデルは，音響モデルを 2 種類，言語モデルはその形態の違いにより 6 種類を用いた．すなわち音響モデルと言語モデルの組み合わせにより 12 種類の音声認識システムを用意した．

用意した 12 種類の音声認識システムのうち，10 種類は言語的な問題が軽減される平仮名单語認識システムである．

### 3.1 音声認識システム

音声認識システムの概要を図 3.1 に示す．音声認識システムは音声波形から声の特徴を抽出する音響分析部，音響モデルや言語モデル，単語辞書を参照しながらその特徴量を単語列に変換する音声認識プログラムから成る．

以下では本研究に用いた音声認識システムである大語彙連続音声認識 (Large-Vocabulary Continuous Speech Recognition : LVCSR) エンジンについて簡単な説明を行う．

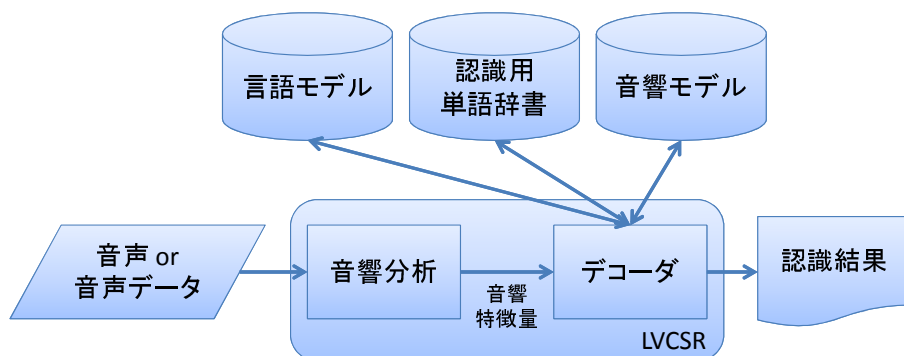


図 3.1: 音声認識システムの概要

### 3.1.1 音声認識の原理

音声認識の原理は、発話者がある単語列  $W = \{w_1, \dots, w_n\}$  を発話して、その音声  $A$  が観測されたという条件で、事後確率が最も高い単語列  $\tilde{W} = \{\tilde{w}_1, \dots, \tilde{w}_n\}$  を求めることである (式 (3.1) )。

$$\tilde{W} = \underset{W}{\operatorname{argmax}} P(W|A) \quad (3.1)$$

しかし、この確率を求めることは非常に困難なため、ベイズの定理を用いて以下のように変形する (式 (3.2) )。

$$\tilde{W} = \underset{W}{\operatorname{argmax}} \frac{P(A|W)P(W)}{P(A)} \quad (3.2)$$

この式 (3.2) での変数は  $W$  であり、 $P(A)$  は変化しないので、以下のように変形することができる (式 (3.3) )。

$$\tilde{W} = \underset{W}{\operatorname{argmax}} P(A|W)P(W) \quad (3.3)$$

この式 (3.3) が音声認識の基本式となる。  $P(A|W)$  は単語列  $W$  を仮定したときの特徴ベクトル  $A$  の確率 (帰属確率) であり、この確率を求めるために作成されるモデルを音響モデルと呼ぶ。  $P(W)$  は単語列  $W$  が観測される確率 (事前確率) であり、この確率を求めるために作成されるものを言語モデルと呼ぶ。

音響モデル、言語モデルでは、確率を対数で表しており、これを対数尤度と呼ぶ。確率を対数尤度で表す理由は、確率を使用した場合、事前確率、事後確率を計算する際、有効桁数の桁落ちが発生する可能性があるためであり、有効桁数の桁落ちがない対数尤度を使用する。また、音響モデルの最小単位は音素または音節、言語モデルの最小単位は単語であるため、最終的な全体の尤度を音響尤度と言語尤度の重み付き和で求めることが多い。通常は、以下の式 (3.4) を用いる。ここで  $\lambda$  は言語の重みであり、全体の尤度にしめる言語尤度の割合を決定するパラメータである。

$$\tilde{W} = \underset{W}{\operatorname{argmax}} \{ \log P(A|W)P + \lambda \log(W) \} \quad (3.4)$$

### 3.1.2 音声認識エンジン : Julius

本研究では、音声認識エンジンとして Julius ver. 4.1.3<sup>4</sup>を用いる。Julius とは、IPA「日本語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 [61] から提供された大語彙連続音声認識エンジンである。

Julius は、2 パス方式の探索を行っている。1 パス目では単純な言語モデルを用いた近似計算を行い、1 パス目で得られた単語トレリスを用いて、2 パス目で複雑な言語モデルを用いて最適な認識単語列を出力する。

大語彙連続音声認識エンジンは、探索結果の尤度順に複数の音声認識結果を出力することができる。この出力は N-Best 出力と呼ばれる。

<sup>4</sup><http://julius.sourceforge.jp/> (現在の最新バージョンは ver. 4.3.1)

### 3.1.3 連続音節認識

本研究では連続音節認識の結果も用いる。

連続音節認識とは、言語モデル(言語の制約)を利用しない音声認識のことを言う。本研究で用いる連続音節認識では、Julius で用いる言語モデルにおいて、全ての音節の bigram, trigram 確率が等しいモーラ単位の言語モデルを利用することで、擬似的に実現している。また、式 (3.4) の  $\lambda$  を 0 とするように Julius の認識パラメータを設定する。これによって、Julius における音響特徴量(音響モデル)のみに依存した認識結果が得られる。言語モデルがモーラで構成されているため、認識結果はすべて平仮名列となる。

### 3.1.4 音声認識結果の評価

一般に音声認識結果の評価には、音声認識率を用いる。本研究では、音声認識率を音節単位で算出する音節認識率によって行う。音節認識率には音節正解率と音節認識精度があり、その定義を式 (3.5) および式 (3.6) に示す。音節認識精度は挿入誤り数を考慮しており、挿入誤りが非常に多くなると負の値もとる。

$$\text{音節正解率 (Corr.)}[\%] = \frac{N - D - S}{N} \times 100 \quad (3.5)$$

$$\text{音節正解精度 (Acc.)}[\%] = \frac{N - D - S - I}{N} \times 100 \quad (3.6)$$

ここで、 $N$  は総音節数、 $D$  は脱落誤り音節数、 $S$  は置換誤り音節数、 $I$  は挿入誤り音節数をそれぞれ指す。

## 3.2 形態素解析システム

大語彙日本語連続音声認識を行う場合に、日本語の言語モデルを作成する必要がある。この言語モデルを作成する際に、形態素解析という処理が必要になる。

形態素とは、日本語の文を構成する最小の単位で、名詞、動詞、形容詞などのことを言う。文章から形態素を切り出すことを、形態素解析という。日本語の文章は英語の文章と異なり単語ごとにスペースや区切りがなく、形態素として抽出しにくい。また、計算機にとって文章は、ただの文字列であり文法や意味を持っているものではないため、形態素に分解することは容易ではない。形態素を解析するツールとして、奈良先端科学技術大学自然言語処理講座で作られた「茶筌 (ChaSen)」[62] や京都大学情報学研究科-日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクトによって作られた「和布蕪 (MeCab)」[63] が存在する。この形態素解析は、

1. ある特定の位置から始まる全ての形態素を形態素の辞書を引くことによって得る

2. 辞書を引くことによって得られた個々の形態素に対して、その直前の位置に存在する全ての形態素との接続可能性のチェックする
3. 形態素コスト、接続コストの計算を行なう

という順序を経て形態素を抽出している。

茶筌と和布蕪の大きな違いとして、コスト計算をおこなうアルゴリズムが挙げられる。茶筌では隠れマルコフモデル (Hidden Markov Model : HMM) を用いている。一方、和布蕪では条件付き確率場 (Conditional Random Field : CRF) を用いている。

### 3.3 音響モデル

音響モデル (Acoustic Model : AM) とは、ある観測信号が得られた場合、統計的にどの音声の信号に最も近いかを求めるために使用されるモデルである。

音響モデルは、通常は隠れマルコフモデル (Hidden Markov Model : HMM) によってモデル化されることが多い。HMM とは、時系列信号の確率モデルであり、複数の定常信号源の間を遷移することで、非定常な時系列信号をモデル化したものである。

HMM からの信号出力確率の計算方法を説明する。HMM  $H$  から信号系列  $O = o(1), \dots, o(N)$  が出力される確率  $P(O|H)$  を求める場合について考える。このとき、 $(N + 2) \times (M + 1)$  の格子点上には以下の状態停留確率が存在する。ここで  $M$  は状態数、 $N$  は時刻を表している。

$$\alpha(n, m) = P(s(n)|O) \quad (s(n) = S_m) \quad (3.7)$$

状態停留確率とは、信号系列  $O$  が与えられたとき、時刻  $n$  において状態  $S_m$  に停留している (すなわち、 $s(n) = S_m$ ) 確率であり、格子点  $(n, m)$  ごとに与えられる。また状態系列  $S = s(0), \dots, S(N + 1)$  は、 $(0, 0)$  から  $(N + 1, M)$  まで格子点上をたどる 1 つの経路となる。

図 3.2 のように、観測された信号系列  $O$  を出力することができる状態系列は複数あるが、HMM  $H$  のある状態遷移系列  $S$  に沿って信号系列  $O$  が出力される確率  $P(O, S|H)$  は、以下の式で表される。

$$\begin{aligned} P(O, S|H) &= P(O|S, H)P(S|H) \\ &= \left\{ \prod_{n=1}^N b_{s(n)}(o(n)) \right\} \cdot \left\{ a_{0s(1)}a_{s(N)M} \prod_{n=1}^{N-1} a_{s(n)s(n+1)} \right\} \\ &= a_{0s(1)} \left\{ \prod_{n=1}^{N-1} b_{s(n)}(o(n))a_{s(n)s(n+1)} \right\} b_{s(N)}(o(N))a_{s(N)M} \end{aligned} \quad (3.8)$$

さらに、異なる状態系列同士は排反であるため、以下の式により、すべての可能な状態系列を介した出力確率の和で観測信号の出力確率を求める。

$$\begin{aligned} P(O|H) &= \sum_S P(O, S|H) \\ &= \sum_S \left[ a_{0s(1)} \left\{ \prod_{n=1}^{N-1} b_{s(n)}(o(n))a_{s(n)s(n+1)} \right\} b_{s(N)}(o(N))a_{s(N)M} \right] \end{aligned} \quad (3.9)$$

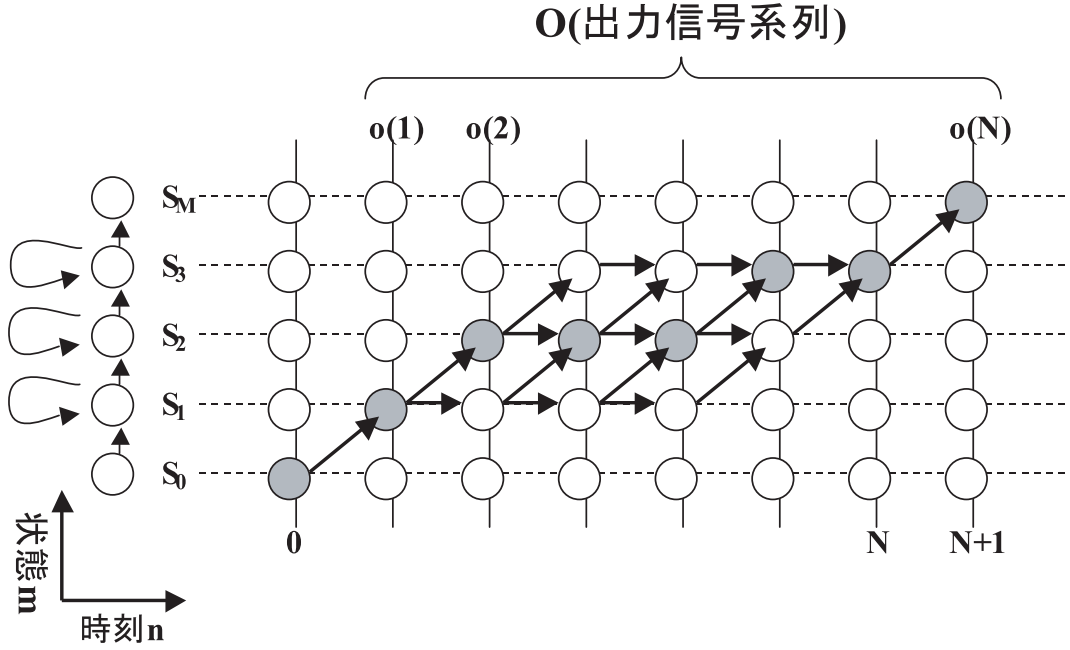


図 3.2: 状態系列と出力信号

しかし，すべての可能な状態系列の出力確率を求めているのは実時間での実行は難しくなる．そこで，以下の式により，時刻  $n$  で状態  $i$  に至る状態系列の中で最も高い確率を与える状態系列の出力確率を用いる．これをビタビアルゴリズムと呼ぶ．

$$\hat{s} = \operatorname{argmax}_S \left[ a_{0s(1)} \left\{ \prod_{n=1}^{N-1} b_{s(n)}(o(n)) a_{s(n)s(n+1)} \right\} b_{s(N)}(o(N)) a_{s(N)M} \right] \quad (3.10)$$

Julius での音響モデルは HTK(HMM Tool Kit) フォーマット [64] に準拠しており，対角共分散の混合連続分布型 HMM で構成され，音声認識に必要な音響モデルの数は，基本的に音響モデルの単位 (monophone, triphone, syllable(音節単位) など) の種類数となる．例えば，日本語の場合，monophone(無音あり) ならば 43 個，syllable(無音あり) ならば 124 個，triphone ならば約 3,000 個 (音素で  $(43)^3$  個，約 80,000 個となるが，状態共有により削減している) 必要となる．

Julius での音響モデルの学習に用いるパラメータは，16kHz，16bit でサンプリングされた音声より求められた，12 次元のメル周波数ケプストラム (MFCC)，その 1 次差分 ( $\Delta$ MFCC) と 2 次差分 ( $\Delta\Delta$ MFCC)，パワーの 1 次差分 ( $\Delta$ LogPow) と 2 次差分 ( $\Delta\Delta$ LogPow) の 38 次元を使用している．

本研究で使用している音響モデルは，モーラ単位 (平仮名 1 音) で音をモデル化したモーラモデル<sup>5</sup>，連続する 3 音素をモデル化したトライフォンモデルの 2 種類の音響モデルを利用する．

<sup>5</sup>モーラは無音を含め 133 種類．

### 3.4 言語モデル

言語モデル (Language Model : LM) とは、ある 1 単語について統計的にどの単語が後続単語として繋がる可能性が高いかを求めるために使用されるモデルである。統計的言語モデルとしては N-gram モデルが有名であり、本研究で使用する音声認識システムもこれを用いている。これは、単語、品詞などを単位とした、N-1 重マルコフモデルで実現される。例えば、単語単位の N-gram(単語 N-gram と呼ぶ) を考えた場合、ある単語列  $W = \{w_1, w_2, \dots, w_t\}$  の出現確率は以下のようになる。

$$P(W) = P(w_1) \prod_{i=2}^{N-1} P(w_i|w_1^i) \prod_{i=N}^T P(w_i|w_{i-N+1}^{i-1})$$

日本語の解析において、通常は  $N=2$  か  $N=3$  が選択される。その場合、それぞれの N-gram モデルを bigram(式 (3.11) ), trigram(式 (3.12) ) と呼ぶ。

$$P(W) = P(w_1) \prod_{i=2}^T P(w_i|w_{i-1}) \quad (3.11)$$

$$P(W) = P(w_1)P(w_2|w_1) \prod_{i=2}^T P(w_i|w_{i-2}, w_{i-1}) \quad (3.12)$$

次に、単語 bigram を例に、ある学習コーパス (単語述べ数  $M$  単語, 単語種類数  $K$  単語) が用意された時に、この bigram 確率をどのように求めれば良いかを最尤推定を用いて導く。

まず、学習コーパスの言語尤度  $L(\theta)$  は以下 (式 (3.13) ) のようになる。

$$L(\theta) = \prod_{ij} P(w_j|w_i)^{C(w_i, w_j)} \quad (3.13)$$

ここで、 $C(w_i, w_j)$  は、学習コーパス中で単語  $w_i, w_j$  が共起した回数を、 $P(w_j|w_i)$  は、単語  $w_i$  の次に単語  $w_j$  が来る確率 (bigram 確率) を示している。

同様に、対数尤度も以下のように求めることが可能である。

$$\log L(\theta) = \sum_{ij}^K C(w_i, w_j) \log P(w_j|w_i)$$

ここでは、ラグランジェの未定係数法を用いて尤度方程式を最大にする確率  $P(w_j|w_i)$  ( $K$  種類) を見つける。すべての単語  $w_i$  について、



$$\sum_j^K P(w_j|w_i) = 1$$

が成り立つのでこれを制約条件として用いると，ラグランジェ関数は次のようになる．

$$L(\theta) = \sum_{ij}^K C(w_i, w_j) \log P(w_j|w_i) + \lambda_i \{1 - \sum_j^K P(w_j|w_i)\}$$

これを変数  $P(w_j|w_i)$  で偏微分すると，

$$\frac{\partial L}{\partial P(w_j|w_i)} = \frac{C(w_i, w_j)}{P(w_j, w_i)} - \lambda_i \quad (3.14)$$

を得る．式 (3.14) を 0 とおいて， $P(w_j|w_i)$  について解くと，

$$P(w_j|w_i) = \frac{C(w_i, w_j)}{\lambda_i} \quad (3.15)$$

となる．これをすべての  $j$  について総和をとると， $\sum_{j=1}^K P(w_j|w_i) = 1$  となるので，

$$\lambda_i = \sum_{j=1}^K C(w_i, w_j) = C(w_i)$$

となる．これを再度，式 (3.15) に代入すると，

$$P(w_j|w_i) = \frac{C(w_i, w_j)}{C(w_i)} \quad (3.16)$$

となる．つまり，単語 bigram  $P(w_j|w_i)$  は，学習コーパス中に出現する単語共起  $w_i, w_j$  の回数を， $w_i$  の出現回数で割ったものとなる．

しかし，式 (3.16) をそのまま使用することは問題がある．それは，学習コーパスに出現しなかった共起における bigram は共起回数が 0 回のために確率が 0 となってしまうことである．これは，文全体の確率を bigram 単位の確率の積で求めている場合には，出現しない単語ペアが一つでもあれば文の確率が 0 となってしまう危険性を含んでいることを示している．このような問題はゼロ頻度問題と呼ばれている．

ゼロ頻度問題に対処するには，既知単語ペア (学習コーパスに存在する単語ペア) における確率和を 1 より小さくして，余った確率を，未知単語ペア (学習コーパスに存在しない単語ペア) に割り振る手段が基本となってくる．本研究で使用した言語モデル構築ツールである，CMU SLM toolkit[65] では，これを，バックオフスムージングという手

法で実現している．バックオフスムージングとは未知の N-gram の確率を，(N-1)-gram の確率から推定する手法である．

まずは最尤推定による bigram 確率

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})}$$

はバックオフスムージングにより推定される．

$$P^*(w_i|w_{i-1}) = \begin{cases} P(w_i|w_{i-1}) & \text{if } C(w_{i-1}) > 0 \\ \alpha(w_{i-1})P^*(w_i) & \text{else} \end{cases}$$

$\alpha$  は未知の bigram 確率を unigram から推定するバックオフ係数で，

$$\alpha(w_{i-1}) = \frac{1 - \sum_{C(w_{i-1}, w_i) > 0} P^*(w_i|w_{i-1})}{1 - \sum_{C(w_{i-1}, w_i) > 0} P^*(w_i)} \quad (3.17)$$

で求めることが可能である．

言語モデルの評価基準としては，パープレキシティがよく用いられる．単語系列  $W = \{w_1, w_2, \dots, w_n\}$  の出現確立を  $P(w_1, \dots, w_n)$  とする時，この言語の 1 単語あたりのエントロピーは言語のエルゴート性を仮定すると，

$$H = \frac{1}{n} \log P(w_1, \dots, w_n)$$

となる．パープレキシティとはある時点における情報理論的な予測単語数を表現する指標で，

$$PP = 2^H = p(w_1, \dots, w_n)^{-\frac{1}{n}}$$

で定義される．基本的にはパープレキシティが大きいと言語的には複雑といえる．同じ音響モデル，同じ語彙数で言語モデルを作る場合，一般的にはパープレキシティが低いほど音声認識率が高い傾向がある．

しかしながら，本研究で用いた CMU SLM toolkit で未知語を含んだ文のパープレキシティを求める時に，未知語を一つのカテゴリ (単語と同等) として扱うため，カバー率の低いほど，つまり未知語が多いほどパープレキシティが小さくなるという問題点がある．

本研究では，この N-gram が，単語 (漢字・平仮名混じり／平仮名のみ) N-gram であつたり，平仮名 N-gram であつたりと言語モデルの形態を変化させることで複数の平仮名列を出力する平仮名音声認識システムを実現している．

以下では，言語モデルの違いによる認識結果の差異と作成意図について説明する．

### 3.4.1 形態素ベース言語モデル：Word-Base Characters (WBC)

形態素ベースの trigram モデル。形態素は、漢字と英数字、平仮名、片仮名で構成されている。学習に用いた形態素数は 約 27,000 語である。

例：今回 / の / 実験 / の / 目的

WBC は本研究における STD の基準となる音声認識結果を得るための言語モデルとして作成した。形態素ベースの trigram モデルであるため、形態素の繋がりが考慮された認識結果が得られる。検索語が未知語の場合の STD 性能は期待することが出来ないが、最も発話された音声に近い音声認識結果を得ることが期待できる。

### 3.4.2 平仮名形態素ベース言語モデル：Word-Base Hiragana (WBH)

単語ベースの trigram モデル。単語はすべて平仮名で構成され、元の単語に漢字や英数字、片仮名が含まれている場合には、すべて平仮名系列に変換される。

例：こんかい / の / じっけん / の / もくてき

WBH はすべて平仮名の単語で構成されるため、音韻系列としては未知語が存在しない言語モデルとなる。形態素ベースの trigram モデルであるため、形態素の繋がりが考慮された認識結果が得られる。WBC の様に同音異義語が存在しないため、正解に近い音韻系列の音声認識結果を得ることが期待できる。

### 3.4.3 文字ベース言語モデル：Character Base (CB)

文字ベースの trigram モデル。文字はすべて平仮名によって構成されている。

例：こ / ん / か / い / の / じ / っ / け / ん / の / も / く / て / き

CB はすべて 1 文字の平仮名で構成されるため、音韻系列としては未知語が存在しない言語モデルとなる。モーラベースの trigram モデルであるため、モーラの繋がりが考慮された認識結果が得られる。モーラベースであるため、音響的な特徴に影響を受けやすい言語モデルとなっている。しかし、モーラの出現確立を学習していることから、話し言葉に適した音韻系列の音声認識結果を得ることが期待できる。

### 3.4.4 文字系列ベース言語モデル：Bi-Mora (BM)

文字系列ベースの trigram モデル。文字系列は 2 文字の平仮名によって構成されている。

例：こん / かい / のじ / っけ / んの / もく / てき

BM はすべて 2 文字の平仮名で構成されるため、音韻系列としては未知語が存在しない言語モデルとなる。文字系列ベースの trigram モデルであるため、CB より言語的な音韻系列の音声認識結果を得ることが期待できる。

### 3.4.5 文字系列ベース言語モデル：Character Sequence Base (CSB)

文字系列ベースの trigram モデル。文字系列は数文字の平仮名によって構成されている。

例：こん / かい / の / じっ / けん / の / もく / てき

CSB はすべて数文字の平仮名形態素で構成されるため、音韻系列としては未知語が存在しない言語モデルとなる。位置づけとしては、WBH と BM の中間的な言語モデルである。文字系列ベースの trigram モデルであるため、BM より言語的な音韻系列の音声認識結果を得ることが期待できる。

### 3.4.6 疑似連続音節認識用言語モデル：Non

全てのモーラの出現確率を等しくした言語モデル。全てのモーラの出現確率が等しいことで、疑似的に連続音節認識を行うことが可能となる。

疑似的に連続音節認識が行えるため、モーラや形態素の言語的な接続確立に依存することがない。このため、言語的な制約に左右されることがない、音韻系列の音声認識結果を得ることが期待できる。

## 3.5 認識用単語辞書

認識用単語辞書とは、音響モデルと言語モデルの整合性をとるために用いる。

認識用単語辞書は語彙のエントリの表記と音素記号列からなる HTK フォーマットに準拠している。音素表記は、日本音響学会の音声データベース委員会で策定されたものを標準とし、そうでない場合は、音響モデル作成者が単語のかな表記から音素表記への変換規則を用意する。

例として、「言語」という言葉を表すにはモノフォンやトライフォンの場合は、g\_e\_N\_g\_o と音素で表記するが、音節の場合は ge\_N\_go と音節で表記する。

本研究では、認識用単語辞書は音響モデル 2 種類と言語モデル 6 種類の組み合わせによる 12 種類を用意した。なお、各認識用単語辞書は言語モデルに合わせて用意したため、語彙数は音響モデルに依存しておらず、認識用単語辞書の音韻の表記が異なるのみである。

なお、これ以降では、言語モデルと単語辞書は対とし、アルファベットで表記する。

## 3.6 各モデルの学習条件

本研究において、言語モデルの Non 以外のすべてのモデルは、CSJ[46][66] のコア講演以外の講演音声を書き起こしたテキストから学習している。

表 3.1: 認識用単語辞書の語彙数

認識用単語辞書種	奇数モデル	偶数モデル
認識用単語辞書 WBC	26,693	26,693
認識用単語辞書 WBH	19,953	19,953
認識用単語辞書 CB	262	262
認識用単語辞書 BM	12,120	12,407
認識用単語辞書 CSB	15,010	15,361
認識用単語辞書 Non	146	146

ただし、本研究では第2章で述べたように CSJ のコア講演音声を対象とする STD テストコレクション [19] を使用している．STD の性能評価をオープンなデータで行うために、参考文献 [39] の音声認識条件に基づき学習および認識を行った．

ただし、言語モデルの BM と CSB は認識用単語辞書が奇数モデルと偶数モデルで異なっている．BM と CSB 以外の言語モデルでは、作成した言語モデルの性質上、奇数モデルと偶数モデルの各認識用単語辞書の語彙数は同一となり、各認識用単語辞書のエントリー数は表 3.1 に示される語彙数となる．

### 3.7 複数の音声認識システムを利用した音声認識実験と認識性能

予備実験として、用意した 12 種類の音声認識システムの出力を利用することで STD 性能が改善するかを判断するために、音声認識性能の評価を行った．

実験音声は STD の対象となる CSJ のコア講演音声である．

言語モデルに WBC を用いた場合の平均単語認識率を表 3.2 に示す．また、12 種類の音声認識システムの音節認識率と、12 種類の音声認識システムの出力を時間同期で連結させたときの音節認識率を表 3.3 に示す．表 3.3 には、各音声認識システムの 1-Best 出力の音節認識率と、10-Best 出力を時間同期で連結させた場合の音節認識率の 2 種類を掲載している．また、複数の音声認識システムの出力を組み合わせた場合の音節認識率として、10 種類の音声認識システムの 1-Best 出力を時間同期で連結させた場合の音節認識率と、10 種類の音声認識システムの 10-Best 出力 (すなわち、100 個の音声認識結果) を時間同期で連結させた場合の音節認識率を掲載している．10 種類の音声認識システムの出力の組み合わせは 2 つあり、その違いは言語モデルの組み合わせである．表 3.4 に言語モデルの組み合わせを示す．

表 3.3 では、10 種類の音声認識システムを組み合わせることによって、94% という高い Corr. を達成することができている．また、単一の音声認識システムの 10-Best 出力を組み合わせた結果と、10 種類の音声認識システムの出力を組み合わせた結果を比較すると、10 種類の音声認識システムの出力を組み合わせた結果の方が Corr. が良い．す

表 3.2: CSJ コア講演音声の平均単語認識率 [%]

LM / AM	Corr.	Acc.
WBC/Tri	76.68	71.93
WBC/Syl	67.54	64.10

表 3.3: CSJ コア講演音声の平均音節認識率 [%]

LM / AM	1-Best		10-Best	
	Corr.	Acc.	Corr.	Acc.
WBC/Tri	<b>86.46</b>	<b>83.01</b>	<b>89.96</b>	<b>44.88</b>
WBH/Tri	86.27	81.42	89.95	35.06
CB/Tri	81.83	77.42	85.99	41.74
BM/Tri	83.60	78.64	88.35	39.47
CSB/Tri	85.66	80.96	89.26	37.16
Non/Tri	71.00	51.20	74.56	21.06
WBC/Syl	79.11	76.35	84.19	35.73
WBH/Syl	79.32	75.83	84.29	29.90
CB/Syl	73.84	71.18	79.47	42.10
BM/Syl	77.89	74.42	84.60	37.26
CSB/Syl	78.58	75.36	83.55	33.03
Non/Syl	63.68	45.43	67.96	21.57
10 Systems1	94.19	-11.67	96.25	-241.04
10 Systems2	94.28	-13.78	96.47	-243.51

表 3.4: 10 種類の音声認識システムの言語モデルの組み合わせ

	言語モデルの組み合わせ
10 Systems1	WBC, WBH, CB, CSB, Non
10 Systems2	WBC, WBH, CB, BM, Non

なわち, 10 個の音声認識結果を用いるのであれば, 単一の音声認識システムの 10-Best 出力を用いるより, 異なる 10 種類の音声認識システムの 1-Best 出力を用いることで, より多くの音節をカバーすることが可能ということが示された.

### 3.8 複数の音声認識システムを利用することによる STD 性能の改善余地

一般的な未知検索語検出では、音韻 (音素や音節など) 単位での検索が行われる。本研究でも、この音韻単位での STD を行う。

これを踏まえると、単一の音声認識システムの出力より、複数の音声認識システムの出力を組み合わせた方が、特定のキーワードを見つけられる可能性が高くなる。しかし、大量の挿入誤りが発生しているため、キーワードの検索において多くの湧き出し誤りが発生する可能性が高くなる。

以上のことから、複数の音声認識システムの出力を効果的に利用することによって、高い検索性能が実現できることが期待できる。

### 3.9 まとめ

本章では、音声認識システムに必要な音響モデルや言語モデル、単語辞書といった各要素技術について述べた。また、提案する STD 性能改善に用いた複数の音声認識システムについて述べた。

複数の音声認識システムによる音声認識実験の結果より、音節単位での音声認識性能が改善されていることから、単一の音声認識システムの出力より、複数の音声認識システムの出力を組み合わせた方が、特定のキーワードを見つけられる可能性が高くなることを示した。

第 4 章では、本章で述べた複数の音声認識システムの出力を STD 用のインデックスとして用いる方法について述べる。

## 第4章 音声中の検索語検出のための検索用インデックス

本章では、複数の音声認識システムの出力を利用した STD のためのインデキシングについて述べる。

第3章では、本研究で用意した12種類の音声認識システムの出力が、音節単位での音声認識性能を改善させることを示した。このことから特定のキーワードを見つける可能性が高くなる。しかし、大量の挿入誤りが発生しているため、キーワードの検索において多くの湧き出し誤りが発生する可能性が高くなる。

本章では、この12種類の音声認識システムの出力を、どのような形で利用することが検索性能の改善につながるかを調査した。

まず、単一の音声認識システムの出力を音節単位に変換した単純なサブワードベースのインデックスと、音素単位に変換した単純なサブワードベースのインデックスを用いて検索性能を調査した。さらに、コンフュージョンネットワークを利用し、単一の音声認識システムの N-Best 出力を組み合わせたインデックスを構築し、検索性能を調査した。

次に、複数の音声認識システムの出力を利用したインデキシングを検討した。複数の音声認識システムの出力を音節単位に変換した単純なサブワードベースのインデックスと、音素単位に変換した単純なサブワードベースのインデックスを用いて検索性能を調査した。また、コンフュージョンネットワークを利用し、複数の音声認識システムの出力を組み合わせたインデックスを構築し、検索性能を調査した。

### 4.1 単一の音声認識システムの出力を利用したインデックス

単一の音声認識システムの出力を利用したインデックスは、前章で述べた12種類の音声認識システムをそれぞれ検索用のインデックスとして利用したものである。

なお、本研究ではインデックスは1発話単位で構築される。

#### 4.1.1 サブワードベースインデックス

検索語が未知語である場合には、単純な文字列検索による STD は困難となる。そこで、一般的な未知語の検索語検出では、サブワード単位での検索が行われる。サブワー



ドとは音韻系列 (半音素 [69], 音素, 音節系列などの単語より小さい単位のシンボル系列) のことを指す。

音声認識システムの出力から得られるサブワード系列を検索用のインデックスとして利用したものが, サブワードベースインデックスとなる。本研究では, このサブワードを音節系列と音素系列の2種類用意し, STD にはどちらのサブワード単位が適しているかを調査した。

本研究では, 各音声認識システムからは 10-Best の出力を得ている。この 1~10-Best までの出力をそれぞれサブワードベースのインデックス (サブインデックス) として利用し, それぞれの検出結果を統合するタイプのインデックスも用意した。

なお, 音声によっては 10-Best までの出力が得られない場合が存在する。この場合は, 出力が得られた N-Best までをサブインデックスとする。

#### 4.1.2 ネットワーク型インデックス

コンフュージョンネットワークは, シンボルの順序関係を保持しながら, 複数のシンボル系列を表現する最も効率的な方法といえる。このコンフュージョンネットワークを用いることで, 複数の音声認識結果を効率よく組み合わせることが可能となる。コンフュージョンネットワークは NULL 遷移を意味する特殊なシンボル “@” を持つ。“@” によって, Node を飛ばしてシンボル列の検索を行える場合がある。この性質を利用し, 複数の音節系列をうまく組み合わせることができると考えた。しかし, “@” の影響によりシンボル隣接性のチェックが難しくなるといった問題点が残る。

このコンフュージョンネットワークを利用したインデックスは, 単一の音声認識システムの N-Best 出力を組み合わせで構築する。この形態のインデックスをネットワーク (または CN) 型インデックスとする。

このネットワーク型インデックスを構築するサブワード単位は, 音節と音素の2種類を用意した。

音節単位でのネットワーク型インデックス (Syllable Confusion Network : SCN) のイメージと構築例を図 4.1 に示す。また, 音素単位でのネットワーク型インデックス (Phoneme Confusion Network : PCN) のイメージと構築例を図 4.2 に示す。図 4.1 と図 4.2 の例は, 単一の音声認識システムの 10-Best 出力からネットワーク型インデックスを構築している。

ネットワーク型インデックスの構築手順は ROVER の手法 [20] を利用し, 以下の手順で行われる。

- 全ての認識結果を音節 (または音素) 系列に変換
- 動的計画 (Dynamic Programming : DP) 法を用いて全ての認識結果のアライメントを取る
- アライメントが取られた音節 (または音素) 系列の 1 つの列を Arc として登録

Input voice data : Cosine  $\theta$  ( /ko sa i N shi i ta/ )

N-Best	Outputs of 10 recognition systems (all outputs are converted into syllable sequence)									
1	ko	sa	na	ni	@	shi	i	i	ka	@
2	ko	su	a	i	N	shi	ri	i	ta	a
3	ko	sa	ma	chi	@	@	i	@	ka	@
4	ko	sa	@	@	N	shi	ki	@	ta	@
5	ko	sa	@	@	N	shi	te	i	ka	@
6	do	sa	@	@	N	shi	i	@	ka	@
7	bo	sa	a	chi	mu	ri	q	@	ta	a
8	@	sa	nu	@	@	shi	i	@	ka	@
9	@	sa	mu	bi	N	shi	i	@	ta	a
10	@	sa	@	@	N	chi	i	ki	ga	@

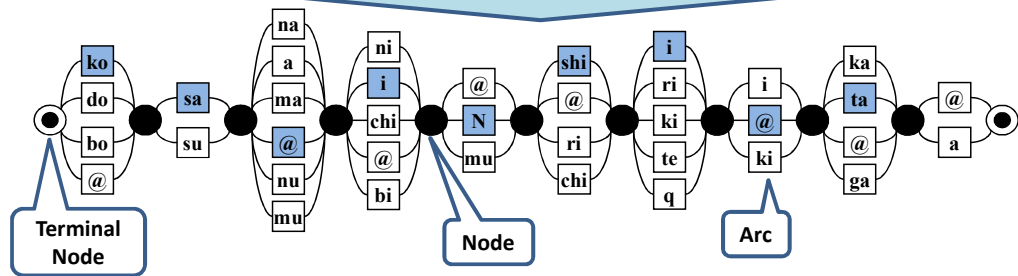


図 4.1: SCN のイメージと構築例

各 Arc に存在する音節 (または音素) に対する遷移確率などの重み付けなどは一切考慮せず、ネットワーク型インデックスに変換している。なお、図中の “@” はヌル遷移を示す。

本研究で用いる DP の傾斜制限は、図 4.3 に示すように行っている。各遷移コストは編集距離 (Edit Distance) に基づいており、一致の場合は 0、誤りの場合は置換・挿入・脱落に関わらず全て 1 としている。

なお、10-Best までの出力が得られない場合は、出力が得られた N-Best までをネットワーク型のインデックスとして構築する。

このネットワーク型インデックスから、STD を行うイメージを図 4.4 に示す。図 4.4 の例は PCN から検索を行う例である。

### 4.1.3 インデックスの種類

単一の音声認識システムの出力を利用した検索用インデックスの種類を表 4.1 に示す。表 4.1 中の SYL(1-Best) は音節

Input voice data : Cosine ( /k o s a i N/ )

N-Best	Outputs of 10 recognition systems (all outputs are converted into phoneme sequence)							
1	k	o	s	@	a	@	i	@
2	@	o	s	u	a	@	@	N
3	k	o	s	@	a	a	i	@
4	k	o	s	@	a	@	@	N
5	k	o	s	@	a	@	@	N
6	@	@	s	@	a	@	@	N
7	t	o	s	@	a	a	@	@
8	@	@	s	@	a	@	i	@
9	@	@	s	@	a	@	@	N
10	@	@	s	@	a	@	@	N

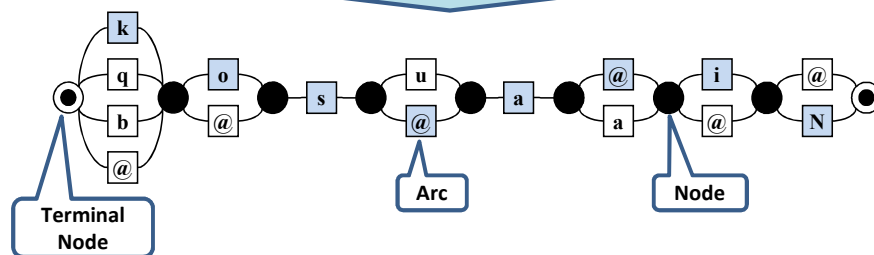


図 4.2: PCN のイメージと構築例

仮説数とは音声認識結果の数を表す. なお, 表 4.1 に示すインデックスは, 本研究で用意した 12 種類の音声認識システムそれぞれにて構築される.

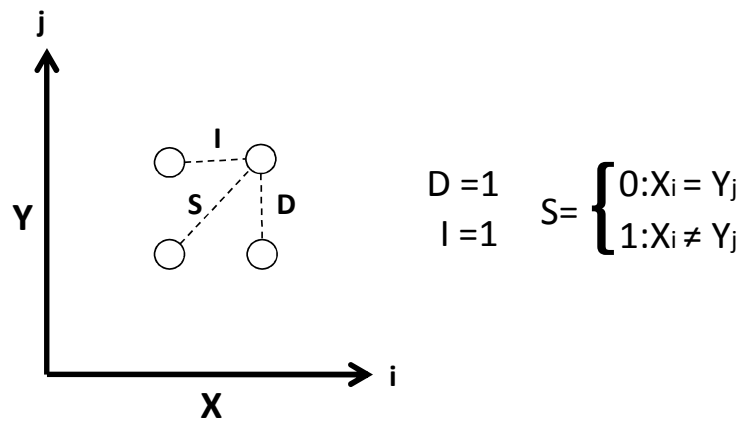


図 4.3: DP の傾斜制限と遷移コストの定義

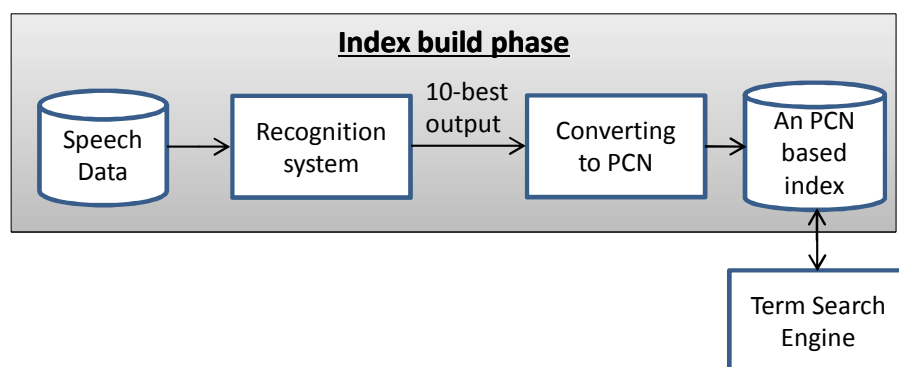


図 4.4: PCN を用いた STD の例

表 4.1: 単一の音声認識システムの出力を利用したインデックスの種類

種類	仮説数	サブインデックス の数	サブインデックスの構成
SYL(1-Best)	1	1	1-Best 出力の音節系列
PHO(1-Best)	1	1	1-Best 出力の音素系列
SYL(10-Best)	10	10	1～10-Best 出力の音節系列
PHO(10-Best)	10	10	1～10-Best 出力の音素系列
SCN	10	1	1～10-Best 出力を音節単位で CN 化
PCN	10	1	1～10-Best 出力を音素単位で CN 化

## 4.2 複数の音声認識システムの出力を利用したインデックス

複数の音声認識システムの出力を利用したインデックスは、前章で述べた 12 種類の音声認識システムの出力を組み合わせることによって、検索用のインデックスとして利用したものである。

### 4.2.1 サブワードベースインデックス

複数の音声認識システムの出力を利用したサブワードベースのインデックスは、単一の音声認識システムの出力を利用したサブワードベースインデックスを単純に組み合わせたものとなる。すなわち、各音声認識システムの 1～10-Best までの出力をそれぞれサブワードベースのインデックス (サブインデックス) として利用し、それぞれの検出結果を統合するタイプのインデックスである。

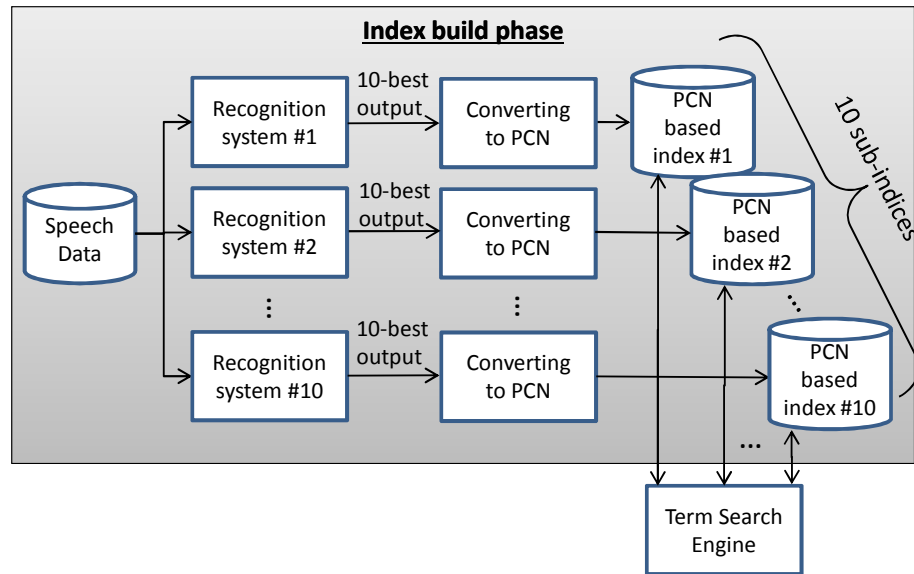


図 4.5: 複数の PCN を用いた STD の例

#### 4.2.2 ネットワークワーク型インデックス

複数の音声認識システムの出力を利用したネットワーク型のインデックスは、2種類用意した。

1つ目は複数の音声認識システムの出力を利用したサブワードベースのインデックスと同様に、単一の音声認識システムの出力を利用したネットワークワーク型インデックスを単純に組み合わせたものとなる。すなわち、各音声認識システムの1~10-Bestまでの出力からネットワークワーク型インデックスを構築し、それぞれサブインデックスとして利用し、最終的に各サブインデックスの検出結果を統合するタイプのインデックスである。この複数のネットワーク型インデックスから、STDを行うイメージを図4.5に示す。図4.5の例は10種類のPCNから検索を行う例である。この図4.5において、例えば#1から#10のいずれか1つのインデックスで検索語が含まれていれば、その検索語を検出する。

2つ目は、複数の音声認識システムのN-Best出力を組み合わせてネットワーク型インデックス構築するものである。単一の音声認識システムから構築するネットワーク型インデックスと区別するために、本研究では音節単位でのネットワーク型インデックスを音節遷移ネットワーク (Syllable Transition Network : STN)、音素単位でのネットワーク型インデックスを音素遷移ネットワーク (Phoneme Transition Network : PTN) と呼称する。

STNのイメージと構築例を図4.6に、PTNのイメージと構築例を図4.7に示す。図4.6と図4.7の例は、10種類の音声認識システムの1-Best出力からネットワーク型インデックスを構築している。

ネットワーク型インデックスの構築手順は、単一の音声認識システムの出力を利用

Input voice data : Cosine  $\theta$  ( /ko sa i N shi i ta/ )

LM/AM	Outputs of 10 recognition systems (all outputs are converted into syllable sequence)										
WBC/Tri	ko	sa	na	i	@	@	shi	i	i	ka	@
WBH/Tri	q	o	su	a	a	N	shi	ri	i	q	ta
CB/Tri	ko	sa	ma	i	chi	@	@	i	@	ka	@
CSB/Tri	ko	sa	@	@	@	N	shi	ki	@	@	ta
Non/Tri	ko	sa	@	@	@	N	shi	te	i	ka	@
WBC/Syl	@	sa	@	@	@	N	shi	i	@	ka	@
WBH/Syl	bo	sa	a	@	a	chi	ri	q	@	@	ta
CB/Syl	@	sa	bi	@	@	@	shi	i	@	ka	@
CSB/Syl	@	sa	@	@	@	N	shi	i	@	@	ta
Non/Syl	@	sa	@	@	@	N	chi	i	ki	ga	@

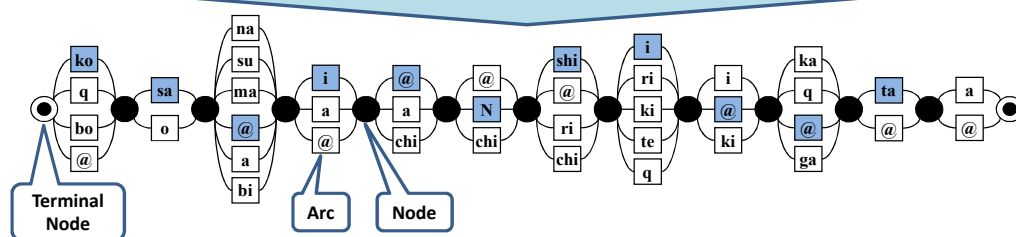


図 4.6: STN のイメージと構築例

表 4.2: STN や PTN を構築する際に用いる音声認識システムの種類と N-Best 出力の組合せ例

音声認識システムの種類	N-Best	仮説数
WBC/Tri, WBH/Tri, CB/Syl, CSB/Syl, Non/Tri	1	5
WBC/Tri, WBC/Syl	7	14
WBC/*, WBH/*, CB/*, BM/*, Non/*	1	10
WBC/*, WBH/*, CB/*, BM/*, CSB/*, Non/*	10	120

したネットワーク型インデックスの構築手順とほぼ同じであり，N-Best の部分が異なる音声認識システムの出力に変更されるのみである．

なお，STN や PTN を構築する際に用いる音声認識システムの種類と N-Best 出力は自在に変更することが可能である．例えば，表 4.2 に示すような音声認識システムの出力を組み合わせることでインデックスを構築できる．表 4.2 中の “\*” は全ての音響モデルを表す．

この PTN (または STN) から，STD を行うイメージを図 4.8 に示す．図 4.8 の例は 10 種類の音声認識システムの 1-Best 出力から PTN を構築し，検索を行う例である．

Input voice data : Cosine ( /k o s a i N/ )

LM/AM	Outputs of 10 recognition systems (all outputs are converted into phoneme sequence)								
WBC/Tri	k	o	s	@	a	@	@	i	@
WBH/Tri	q	o	s	u	a	@	a	@	N
CB/Tri	k	o	s	@	a	m	a	i	@
CSB/Tri	k	o	s	@	a	@	@	@	N
Non/Tri	k	o	s	@	a	@	@	@	N
WBC/Syl	@	@	s	@	a	@	@	@	N
WBH/Syl	b	o	s	@	a	a	a	@	@
CB/Syl	@	@	s	@	a	b	@	i	@
CSB/Syl	@	@	s	@	a	@	@	@	N
Non/Syl	@	@	s	@	a	@	@	@	N

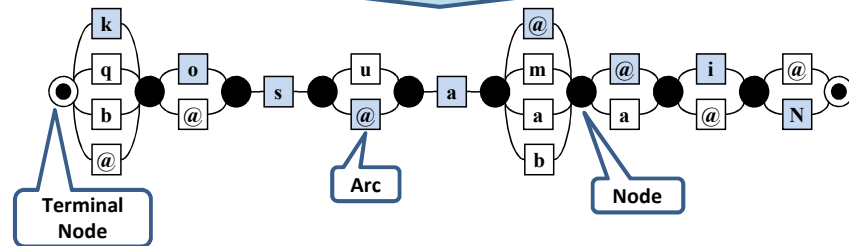


図 4.7: PTN のイメージと構築例

### 4.2.3 インデックスの種類

複数の音声認識システムの出力を利用した検索用インデックスの種類を表 4.3 に示す。表 4.3 中の“n”は音声認識システムの数を表す。

## 4.3 インデックスごとの検索性能

本研究で用意した STD のためのインデックスの種類は、表 4.1 と表 4.3 に示す 16 種類である。この 16 種類のインデックスに対して、検索性能の比較を行う。

なお、単一の音声認識システムを用いた場合の検索性能は付録 D に記載する。

本研究における検索語の検出は、検索語をテキスト形式にて用語検索エンジンに入力することで行う。日本語 STD 用テストコレクションの検索語には読み情報が付与されている。検索語はこの読み情報をもとにインデックスに適した音韻単位に変換し、用語検索エンジンに入力される。検索結果は、用語検索エンジンに設定する閾値ごとに出力される。



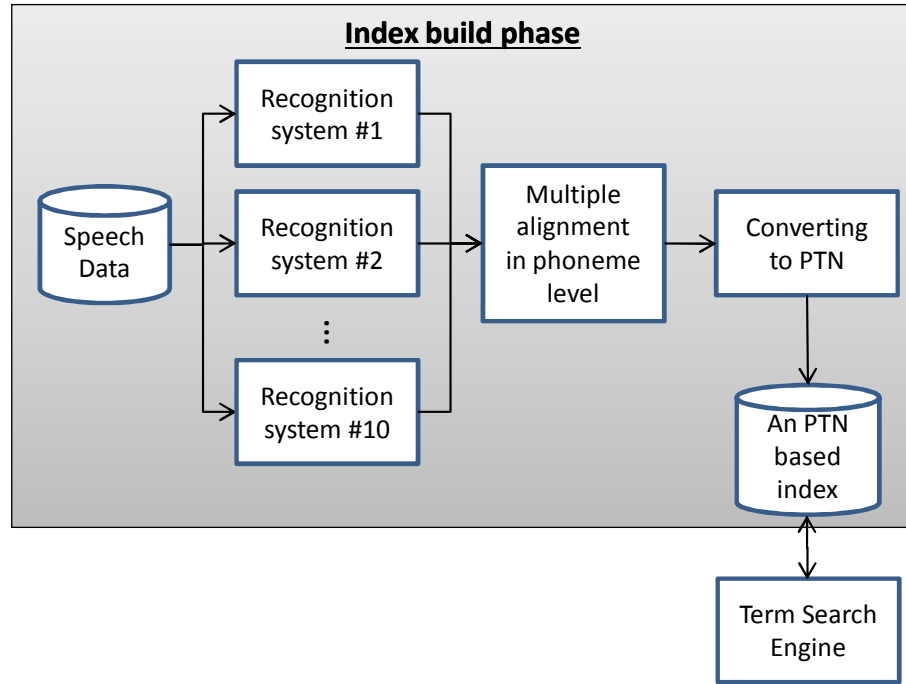


図 4.8: PTN を用いた STD の例

表 4.3: 複数の音声認識システムの出力を利用したインデックスの種類

種類	仮説数	サブインデックス の数	サブインデックスの構成
nSYLs(1-Best)	n	n	n 種の 1-Best 出力の音節系列
nPHOs(1-Best)	n	n	n 種の 1-Best 出力の音素系列
nSYLs(10-Best)	$n \times 10$	$n \times 10$	n 種の 1～10-Best 出力の音節系列
nPHOs(10-Best)	$n \times 10$	$n \times 10$	n 種の 1～10-Best 出力の音素系列
nSCNs	$n \times 10$	n	n 種の 1～10-Best 出力を SCN 化
nPCNs	$n \times 10$	n	n 種の 1～10-Best 出力を PCN 化
STN(1-Best)	n	1	n 種の 1-Best 出力を STN 化
PTN(1-Best)	n	1	n 種の 1-Best 出力を PTN 化
STN(10-Best)	$n \times 10$	1	n 種の 1～10-Best 出力を STN 化
PTN(10-Best)	$n \times 10$	1	n 種の 1～10-Best 出力を PTN 化

#### 4.3.1 動的計画法を用いた検索方法

本研究で用いる，検索語の検出アルゴリズムは DP を用いた単純な方法である．単純な検索アルゴリズムを用いた理由は，本研究の主旨が複数の音声認識システムを利用した STD 用インデックスの構築にあるためである．

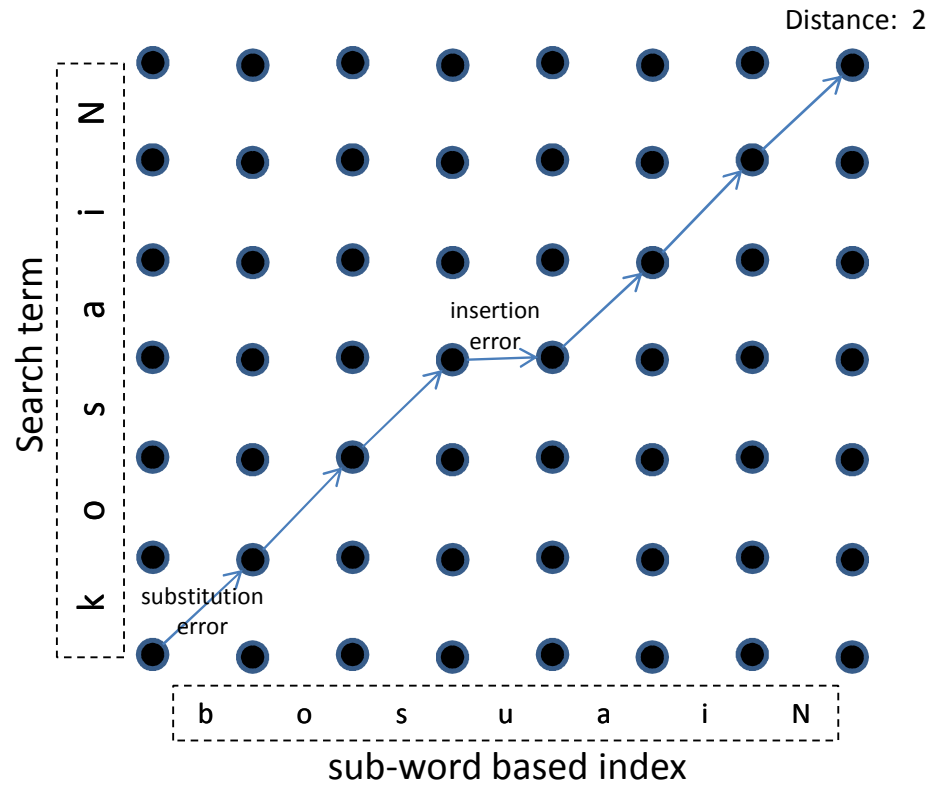


図 4.9: サブワードベースインデックスから DP を用いた検索語の検出例

本稿では、DP の傾斜制限は図 4.3 のように行っており、 $\mathbf{X}$  がインデックス、 $\mathbf{Y}$  が検索語に対応する。

用語検索エンジンに用いる DP の各遷移コストは編集距離に基づいており、一致の場合は 0、誤りの場合は置換・挿入・脱落に関わらず全て 1 としている。

ネットワーク型インデックスは 2 つの Node 間に複数の Arc を持っており、用語検索エンジンはこの複数の Arc を考慮した距離計算を行う。また、ネットワーク型インデックスには NULL 遷移が存在しており、この NULL 遷移に対するコストとして 0.1 を設定している。

最終的に、インデックスと検索語の距離が閾値以下であれば検索エンジンは検索語を検出したと判断する。

図 4.9 はサブワードベースインデックスから、DP によって “k o s a i N” を見つける例を示す。図 4.9 の例では、インデックスと検索語の距離は 2 (置換誤り 1 と挿入誤り 1) となる。

図 4.10 は PTN (または PCN) から DP によって “k o s a i N” を見つける例である。

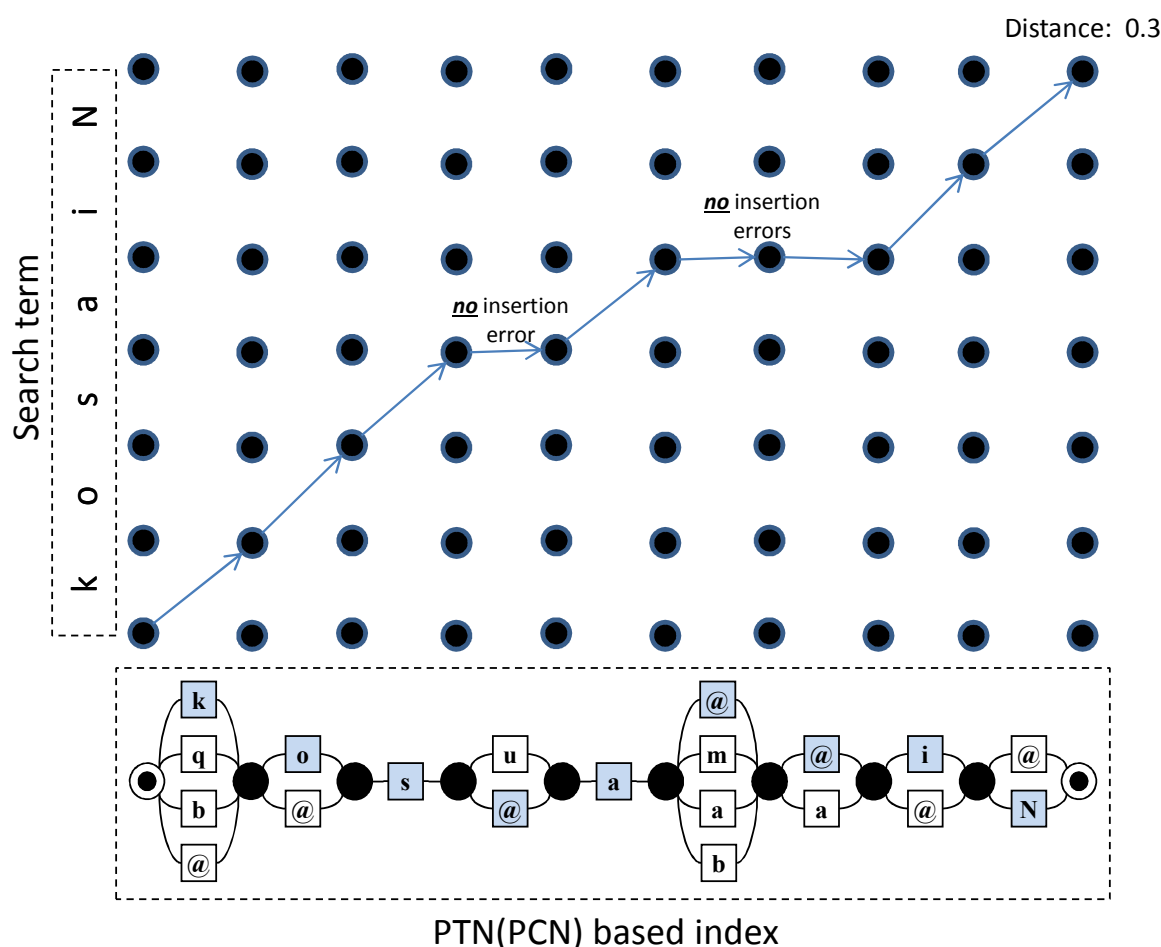


図 4.10: ネットワーク型インデックスから DP を用いた検索語の検出例

### 4.3.2 複数の音声認識システムを利用する効果

まず、単一の音声認識システムの出力を利用した場合と、複数の音声認識システムの出力を利用した場合の検索性能の違いについて比較を行う。この比較実験に用いたインデックスは、表 4.4 に示すものとなる。

単一の音声認識システムの出力を利用したインデックスにおいて、使用した音声認識システムが WBC/Tri と CB/Tri である理由としては、「情報検索システム評価用テストコレクション構築プロジェクト」(National Institute of Informatics Test Collection for IR Systems : NTCIR) の第 9 回目ワークショップでの IR for Spoken Documents(“SpokenDoc”) タスク [67] において、STD 評価用のデータとして WBC/Tri と CB/Tri の音声認識結果が採択されているためである。NTCIR とは、国立情報学研究所が 1998 年から行なっている共同研究プロジェクトのことであり、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の更なる発展を図るワークショップ型共同研究プロジェクトのことである。

表 4.4: 複数の音声認識システムを利用する効果の比較実験に用いたインデックスの種類

インデックス	インデックスの種類	音声認識システムの種類
WBC/Tri(1-Best) <sub>syl</sub>	SYL(1-Best)	WBC/Tri
CB/Tri(1-Best) <sub>syl</sub>	SYL(1-Best)	CB/Tri
WBC/Tri(1-Best) <sub>pho</sub>	PHO(1-Best)	WBC/Tri
CB/Tri(1-Best) <sub>pho</sub>	PHO(1-Best)	CB/Tri
WBC/Tri(10-Best) <sub>syl</sub>	SYL(10-Best)	WBC/Tri
CB/Tri(10-Best) <sub>syl</sub>	SYL(10-Best)	CB/Tri
WBC/Tri(10-Best) <sub>pho</sub>	PHO(10-Best)	WBC/Tri
CB/Tri(10-Best) <sub>pho</sub>	PHO(10-Best)	CB/Tri
10SYLs(1-Best)	nSYLs(1-Best)	WBC/*, WBH/*, CB/*, CSB/*, Non/*
10PHOs(1-Best)	nPHOs(1-Best)	WBC/*, WBH/*, CB/*, CSB/*, Non/*

表 4.5: 表 4.4 に示すインデックスごとの最大 F-measure と ATWV

インデックス	F-measure	ATWV
WBC/Tri(1-Best) <sub>syl</sub>	0.32	0.53
CB/Tri(1-Best) <sub>syl</sub>	0.43	0.65
WBC/Tri(1-Best) <sub>pho</sub>	0.35	0.56
CB/Tri(1-Best) <sub>pho</sub>	0.49	0.66
WBC/Tri(10-Best) <sub>syl</sub>	0.39	0.58
CB/Tri(10-Best) <sub>syl</sub>	0.53	0.70
WBC/Tri(10-Best) <sub>pho</sub>	0.37	0.62
CB/Tri(10-Best) <sub>pho</sub>	0.53	0.74
10SYLs(1-Best)	<b>0.64</b>	0.79
10PHOs(1-Best)	0.63	<b>0.80</b>

この比較実験で用いた評価尺度は、Recall-Precision カーブと F-measure, ATWV である。

表 4.5 に、表 4.4 に示すインデックスごとに Recall-Precision カーブを描いた際の最も高い F-measure と ATWV を示す。

図 4.11 に、表 4.4 に示すインデックスの種類が SYL(1-Best) と PHO(1-Best) の Recall-Precision カーブを示す。

表 4.5 と図 4.11 より、単一の音声認識システムの 1-Best 出力を利用したサブワードベースのインデックスでは、WBC/Tri と CB/Tri 共に音素単位のサブワードベースインデックスの性能が良いことがわかる。これより、音節単位より音素単位の方が STD

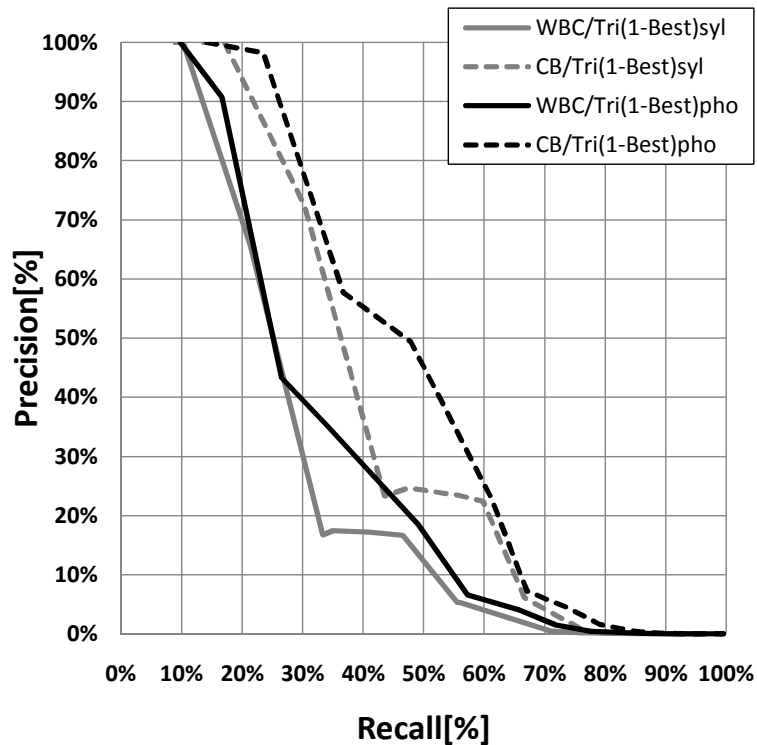


図 4.11: 単一の音声認識システムの 1-Best 出力を利用したサブワードベースインデックスの検索性能の比較

に適していることが推測される．例えば、「コサインシータ」という検索語を検出する際に、音節単位では“ko sa i N shi i ta”の7音節のサブワード系列となるが、音素単位では“k o s a i N sh i i t a”の11音素のサブワード系列となる．音声認識システムの出力では、母音または子音の片方が合っている出力が得られる可能性がある．この性質と実験結果から、音素単位の方が未知語検出により適した検索が行えていることが示された．

また、WBC/Tri と CB/Tri の音声認識結果の違いが、未知語の検出に影響していることがわかる．前章で述べたが、WBC/Tri と CB/Tri の音節単位の音声認識率では、WBC/Tri の方が高かった．しかし、未知語の検出というタスクになると、音節認識率では検索性能が測れないということが結果として得られた．

図4.12に、表4.4に示すインデックスの種類がSYL(10-Best)とPHO(10-Best), nSYLs(1-Best)とnPHOs(1-Best)のRecall-Precisionカーブを示す．

表4.5と図4.12より、単一の音声認識システムの10-Best出力を利用したサブワードベースインデックスより、複数の音声認識システムの1-Best出力を利用したサブワードベースインデックスの性能が良いことがわかる．すなわち、同じ仮説数を用いるのであれば、異なる音声認識システムの出力を用いることが有効であるということとなる．

以上より、複数の音声認識システムの出力を利用することが、STDに有効であることが示された．また、サブワードベースインデックスのサブワードの単位は、音節よ

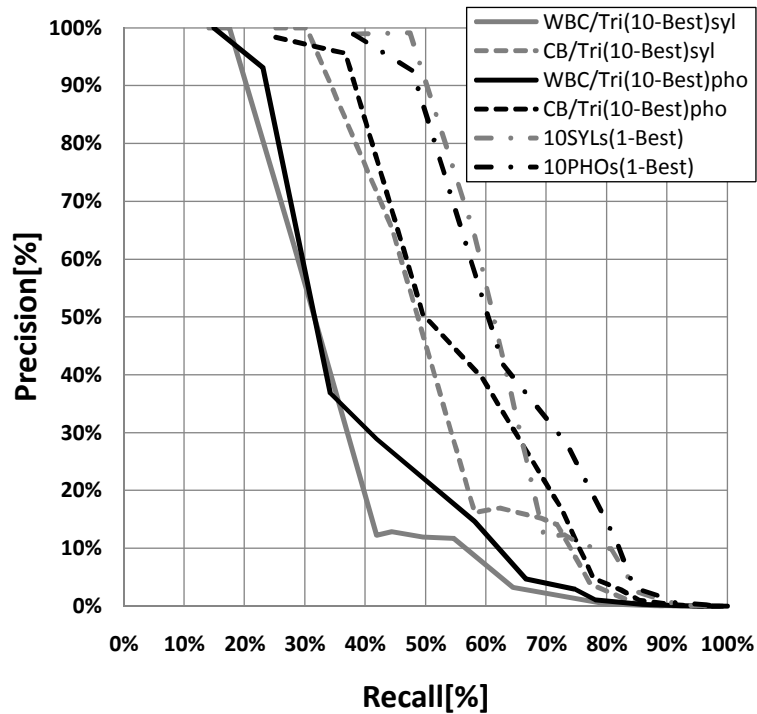


図 4.12: 10 個の音声認識結果を利用したサブワードベースインデックスの検索性能の比較

り音素が適していることが示された。

### 4.3.3 インデックスの形態ごとの評価

続いて、インデックスの形態による STD の性能比較を行う。この比較実験に用いたインデックスは、表 4.6 に示すものとなる。

この比較実験で用いた評価尺度は、Recall-Precision カーブと F-measure, ATWV である。

表 4.7 に、表 4.6 に示すインデックスごとに Recall-Precision カーブを描いた際の最も高い F-measure と ATWV を示す。

図 4.13 に、表 4.6 に示す WBC/Tri のみを用いたインデックスの Recall-Precision カーブを示す。

表 4.7 と図 4.13 より、単一の音声認識システムの出力を利用したインデックスでは、10-Best 出力を利用した音素単位のサブワードベースインデックスが ATWV では最も良く、F-measure では音節単位のネットワーク型インデックスが良いことがわかる。また、Recall-Precision カーブではサブワードベースのインデックスとネットワーク型インデックスでは同程度の性能を示しており、どの形態のインデックスが STD に適しているかを判断することは難しい。図 4.14 に、表 4.6 に示すインデックスの種類が SYL(10-Best)

表 4.6: インデックスの形態による効果の比較実験に用いたインデックスの種類

インデックス	インデックス の種類	音声認識システムの種類
WBC/Tri(1-Best) <sub>syl</sub>	SYL(1-Best)	WBC/Tri
WBC/Tri(10-Best) <sub>syl</sub>	SYL(10-Best)	WBC/Tri
WBC/Tri(SCN)	SCN	WBC/Tri
WBC/Tri(1-Best) <sub>pho</sub>	PHO(1-Best)	WBC/Tri
WBC/Tri(10-Best) <sub>pho</sub>	PHO(10-Best)	WBC/Tri
WBC/Tri(PCN)	PCN	WBC/Tri
10SYLs(1-Best)	nSYLs(1-Best)	WBC/*, WBH/*, CB/*, CSB/*, Non/*
STN(1-Best)	STN(1-Best)	WBC/*, WBH/*, CB/*, CSB/*, Non/*
10SCNs	nSCNs	WBC/*, WBH/*, CB/*, CSB/*, Non/*
10PHOs(1-Best)	nPHOs(1-Best)	WBC/*, WBH/*, CB/*, CSB/*, Non/*
PTN(1-Best)	PTN(1-Best)	WBC/*, WBH/*, CB/*, CSB/*, Non/*
10PCNs	nPCNs	WBC/*, WBH/*, CB/*, CSB/*, Non/*

表 4.7: 表 4.6 に示すインデックスごとの最大 F-measure と ATWV

インデックス	F-measure	ATWV
WBC/Tri(1-Best) <sub>syl</sub>	0.32	0.53
WBC/Tri(10-Best) <sub>syl</sub>	0.39	0.58
WBC/Tri(SCN)	0.43	0.59
WBC/Tri(1-Best) <sub>pho</sub>	0.35	0.56
WBC/Tri(10-Best) <sub>pho</sub>	0.37	0.62
WBC/Tri(PCN)	0.43	0.57
10SYLs(1-Best)	0.64	0.79
STN(1-Best)	0.67	0.78
10SCNs	<b>0.68</b>	0.68
10PHOs(1-Best)	0.63	0.80
PTN(1-Best)	0.64	<b>0.82</b>
10PCNs	0.62	0.73

と PHO(10-Best), nSYLs(1-Best) と nPHOs(1-Best) の Recall-Precision カーブを示す.

表 4.7 と図 4.14 より, 複数の音声認識システムの出力を利用したインデックスでは, PTN が ATWV では最も良く, F-measure では 10SCNs が良いことがわかる. また, Recall-Precision カーブではネットワーク型インデックスが全体的に高い性能を示している.

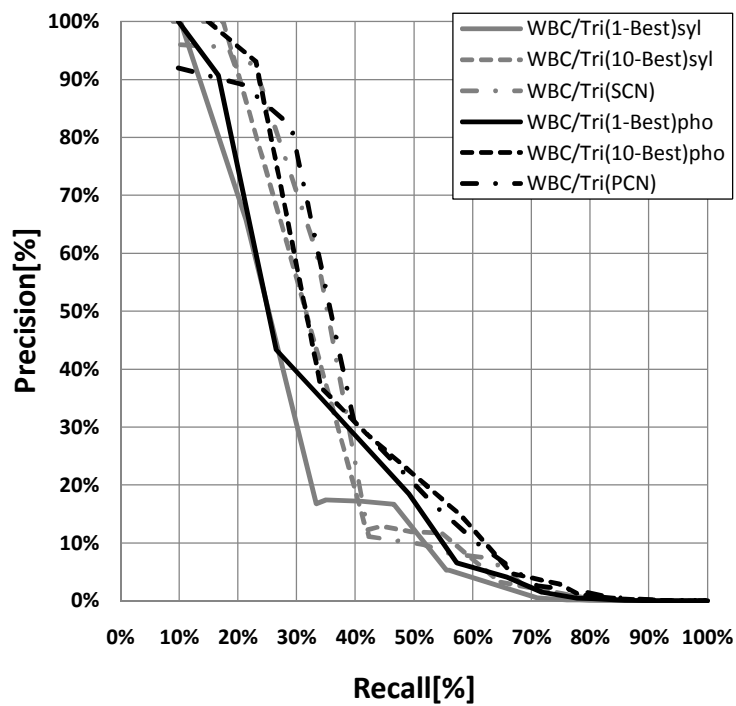


図 4.13: 単一の音声認識システムの出力を利用したインデックスの検索性能の比較

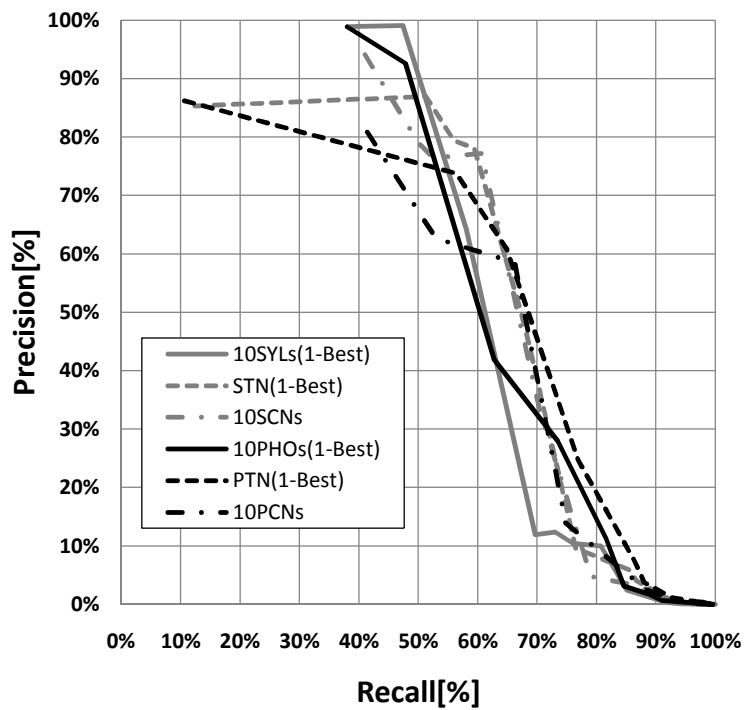


図 4.14: 10 種類の音声認識システムの出力を利用したインデックスの検索性能の比較



表 4.8: インデックスを構成する仮説数による効果の比較実験に用いたインデックスの種類

インデックス	インデックスの種類	音声認識システムの種類
WBC/Tri(10-Best) <sub>pho</sub>	PHO(10-Best)	WBC/Tri
WBC/Tri(PCN)	PCN	WBC/Tri
10PHOs(1-Best)	nPHOs(1-Best)	WBC/*, WBH/*, CB/*, CSB/*, Non/*
PTN(1-Best)	PTN(1-Best)	WBC/*, WBH/*, CB/*, CSB/*, Non/*
10PHOs(10-Best)	nPHOs(10-Best)	WBC/*, WBH/*, CB/*, CSB/*, Non/*
10PCNs(10-Best)	nPCNs	WBC/*, WBH/*, CB/*, CSB/*, Non/*
PTN(10-Best)	PTN(10-Best)	WBC/*, WBH/*, CB/*, CSB/*, Non/*

表 4.9: 表 4.8 に示すインデックスごとの最大 F-measure と ATWV

インデックス	F-measure	ATWV
WBC/Tri(10-Best) <sub>pho</sub>	0.37	0.62
WBC/Tri(PCN)	0.43	0.57
10PHOs(1-Best)	0.63	0.80
PTN(1-Best)	0.64	<b>0.82</b>
10PHOs(10-Best)	<b>0.72</b>	0.80
10PCNs	0.62	0.73
PTN(10-Best)	0.34	0.75

以上の結果より、単一の音声認識システムでは、インデックスの形態によって STD の性能が大きく変わることはないことが示された。しかし、複数の音声認識システムの出力を利用する場合には、ネットワーク型インデックスを用いることで、検索性能が改善されていることが示された。よって、ネットワーク型インデックスを用いることが、本研究において有効であることが示された。

#### 4.3.4 インデックスを構成する仮説数の評価

次に、インデックスを構成する仮説数の違いによる STD の性能比較を行う。この比較実験に用いたインデックスは、表 4.8 に示すものとなる。

この比較実験で用いた評価尺度は、Recall-Precision カーブと F-measure, ATWV である。

表 4.9 に、表 4.8 に示すインデックスごとに Recall-Precision カーブを描いた際の最も高い F-measure と ATWV を示す。

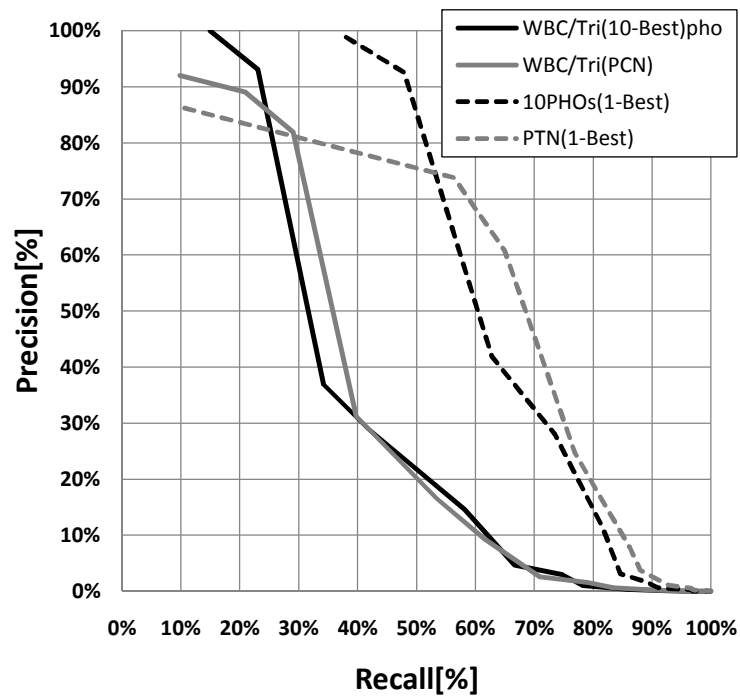


図 4.15: 10 個の仮説数を利用したインデックスの検索性能の比較

図 4.15 に、表 4.8 に示す仮説数が 10 個の場合のインデックスの Recall-Precision カーブを示す。

表 4.9 と図 4.15 より、仮説数が 10 個の場合のインデックスでは、Recall-Precision カーブ、F-measure と ATWV とともに PTN(1-Best) が最も良い性能を示していることがわかる。

図 4.16 に、表 4.8 に示す仮説数が 100 個の場合のインデックスの Recall-Precision カーブを示す。

表 4.9 と図 4.16 より、仮説数が 100 個の場合のインデックスでは、Recall-Precision カーブ、F-measure と ATWV 全てにおいて、10PHOs(10-Best) が最も良い性能を示していることがわかる。

以上の結果より、ネットワーク型インデックスでは探索の幅が広がり、検索語が検出され易くなっている。これにより、誤検出が多く発生してしまうことが示された。また、仮説数が多くなる場合には、サブワードベースインデックスを用いた方が良いことが示された。ただし、ATWV においては PTN(1-Best) が最も良い検索性能を示している。

よって、PTN(1-Best) において、誤検出を抑制することによって最も良い STD が行える可能性が示された。

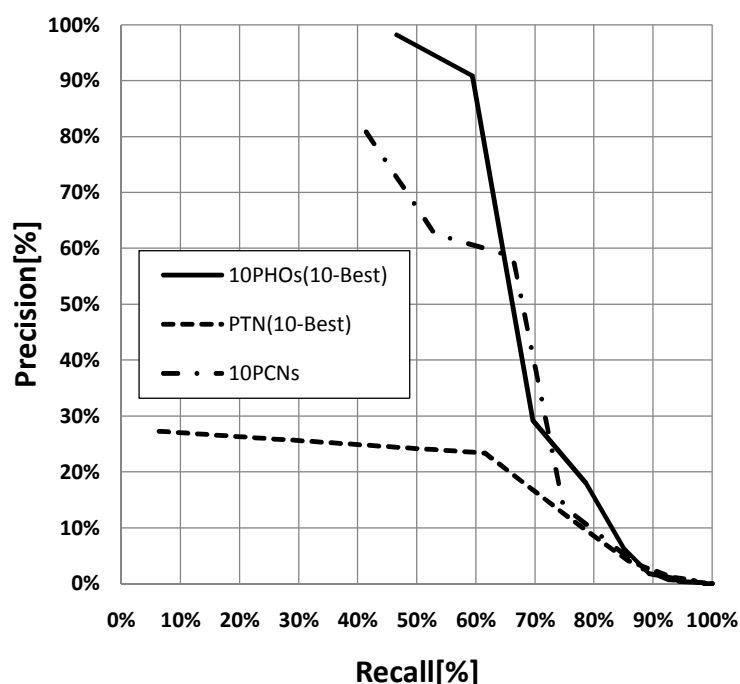


図 4.16: 100 個の仮説数を利用したインデックスの検索性能の比較

#### 4.3.5 インデックスを構成する音声認識システム数の評価

次に、インデックスを構成する音声認識システム数の違いによる STD の性能比較を行う。この比較実験に用いたインデックスは、表 4.10 に示すものとなる。

この比較実験で用いた評価尺度は、Recall-Precision カーブと F-measure, ATWV である。

表 4.11 に、表 4.10 に示すインデックスごとに Recall-Precision カーブを描いた際の最も高い F-measure と ATWV を示す。

図 4.17 に、表 4.11 に示すサブワードベースインデックスの Recall-Precision カーブを示す。

表 4.11 と図 4.17 より、サブワードベースインデックスの場合では、Recall-Precision カーブ、F-measure と ATWV とともに 10PHOs が最も良い性能を示していることがわかる。すなわち、サブワードベースインデックスでは、音声認識システムの数が多いほど、検索性能が高くなることが示された。

図 4.18 に、表 4.11 に示す nPCNs の Recall-Precision カーブを示す。

表 4.11 と図 4.18 より、nPCNs の場合では、Recall-Precision カーブ、F-measure と ATWV とともに 6PCNs の WBC を用いない場合と WBH を用いない場合が良い性能を示していることがわかる。すなわち、単一の音声認識システムの 10-Best 出力から構築される WBC と WBH のネットワーク型インデックスの検出性能が悪く、nPCNs の検索性能を低下させていたことが示された。

表 4.10: インデックスを構成する音声認識システム数による効果の比較実験に用いたインデックスの種類

インデックス	インデックスの種類	音声認識システムの種類
10PHOs	nPHOs(1-Best)	WBC/*, WBH/*, CB/*, CSB/*, Non/*
8PHOs	nPHOs(1-Best)	WBC/*, WBH/*, CB/*, Non/*
6PHOs(unWBC)	nPHOs(1-Best)	WBH/*, CB/*, Non/*
6PHOs(unWBH)	nPHOs(1-Best)	WBC/*, CB/*, Non/*
6PHOs(unCB)	nPHOs(1-Best)	WBC/*, WBH/*, Non/*
6PHOs(unNON)	nPHOs(1-Best)	WBC/*, WBH/*, CB/*
10PCNs	nPCNs	WBC/*, WBH/*, CB/*, CSB/*, Non/*
8PCNs	nPCNs	WBC/*, WBH/*, CB/*, Non/*
6PCNs(unWBC)	nPCNs	WBH/*, CB/*, Non/*
6PCNs(unWBH)	nPCNs	WBC/*, CB/*, Non/*
6PCNs(unCB)	nPCNs	WBC/*, WBH/*, Non/*
6PCNs(unNON)	nPCNs	WBC/*, WBH/*, CB/*
PTN(All)	PTN(1-Best)	WBC/*, WBH/*, CB/*, CSB/*, Non/*
PTN(unCSB)	PTN(1-Best)	WBC/*, WBH/*, CB/*, Non/*
PTN(unCSB+unWBC)	PTN(1-Best)	WBH/*, CB/*, Non/*
PTN(unCSB+unWBH)	PTN(1-Best)	WBC/*, CB/*, Non/*
PTN(unCSB+unCB)	PTN(1-Best)	WBC/*, WBH/*, Non/*
PTN(unCSB+unNON)	PTN(1-Best)	WBC/*, WBH/*, CB/*

図 4.19 に、表 4.11 に示す PTN の Recall-Precision カーブを示す。

表 4.11 と図 4.19 より、nPCNs の場合では、Recall-Precision カーブ、F-measure と ATWV において、Non 以外の言語モデルを用いない場合において検索性能が向上していることがわかる。特に、ATWV においては、CSB と CB を用いないことによって 0.85 という高い検索性能が示されている。すなわち、多くの音声認識システムを用いるより、適度な音声認識システムの種類を用いた方が、検索性能が改善されることが示された。

## 4.4 まとめ

本章では、複数の音声認識システムの出力をどのような形のインデックスとして利用することが、STD 性能の改善につながるかについて述べた。

単一の音声認識システムの出力を利用した場合では、仮説数が多くなるほど検索性能が向上し、ネットワーク型のインデックスを構築することで Recall が 30% から 40% の間では検索性能が良くなることが示された。また、10PHOs(1-Best) の結果に示され

表 4.11: 表 4.10 に示すインデックスごとの最大 F-measure と ATWV

インデックス	F-measure	ATWV
10PHOs	0.63	0.80
8PHOs	0.62	0.78
6PHOs(unWBC)	0.61	0.77
6PHOs(unWBH)	0.54	0.77
6PHOs(unCB)	0.57	0.76
6PHOs(unNON)	0.55	0.72
10PCNs	0.62	0.73
8PCNs	0.62	0.72
6PCNs(unWBC)	0.64	0.75
6PCNs(unWBH)	0.63	0.75
6PCNs(unCB)	0.60	0.70
6PCNs(unNON)	0.60	0.69
PTN(All)	0.64	0.82
PTN(unCSB)	0.67	0.84
PTN(unCSB+unWBC)	0.68	0.83
PTN(unCSB+unWBH)	0.68	0.82
PTN(unCSB+unCB)	0.69	0.85
PTN(unCSB+unNON)	0.65	0.77

るように、複数の音声認識システムの出力を利用することで高い検索性能が示され、PTN(1-Best) においては Recall が 60%以上で最も良い検索性能となった。以上から、複数の音声認識システムの出力を CN 化することの有用性が示された。

しかし、多くの仮説を用いてネットワーク型のインデックスを構築しても、大量の湧き出し誤りが検出されてしまい、検索性能が低下した。この原因としては、ネットワークの Node や Arc が多くなり過ぎてしまい、DP を用いた単純な検索方法では多くの情報を生かしきれていないということが考えられる。また、STD に用いる音声認識システムの N-Best 出力や音声認識システムの出力を変更することによって検索性能が改善されることが示された。すなわち、適切な音声認識システムの N-Best 出力や音声認識システムの出力を選別することによって、STD 性能が改善される可能性が示されたこととなる。しかし、この検索語検出のタスクが変更された場合に、最適な N-Best 出力や音声認識システムの種類が変わる可能性がある。

これらの問題に対応するためにも、検索エンジンの改善が必要となる。また、PTN(1-Best) においても、Recall が低い場合においては 10PHOs(1-Best) などと比較すると誤検出が発生している。

次章では、この誤検出を抑制するための誤検出抑制パラメータと検索エンジンの改善について述べる。

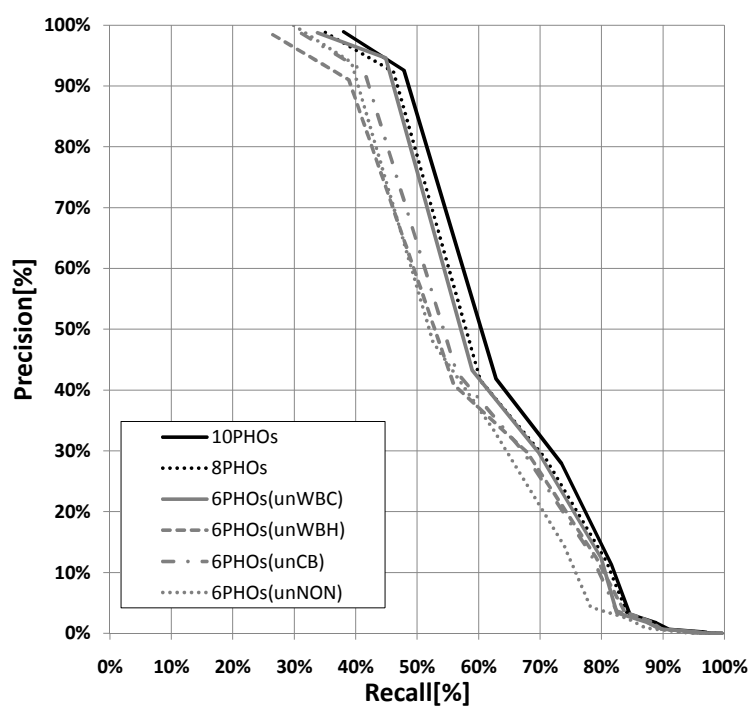


図 4.17: サブワードベースインデックスの検索性能の比較

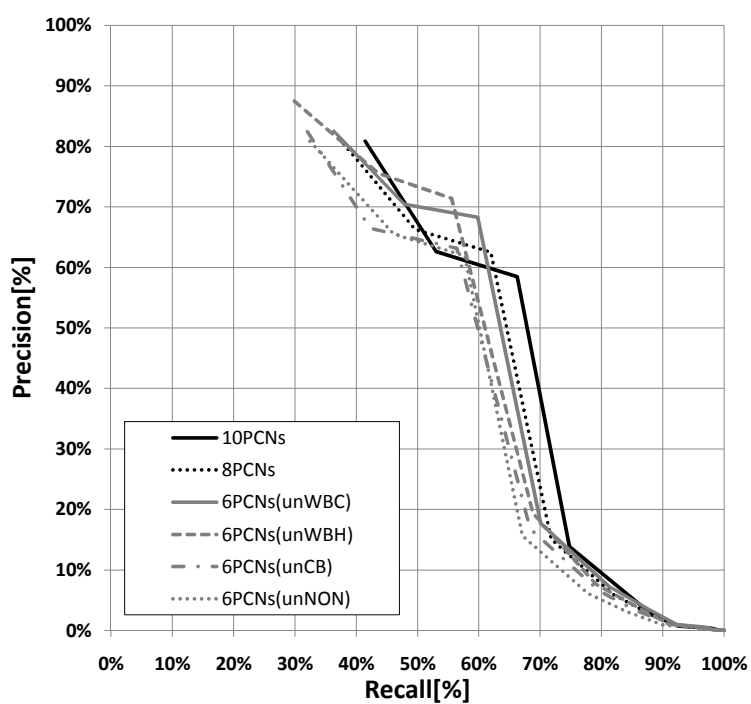


図 4.18: nPCNs の検索性能の比較

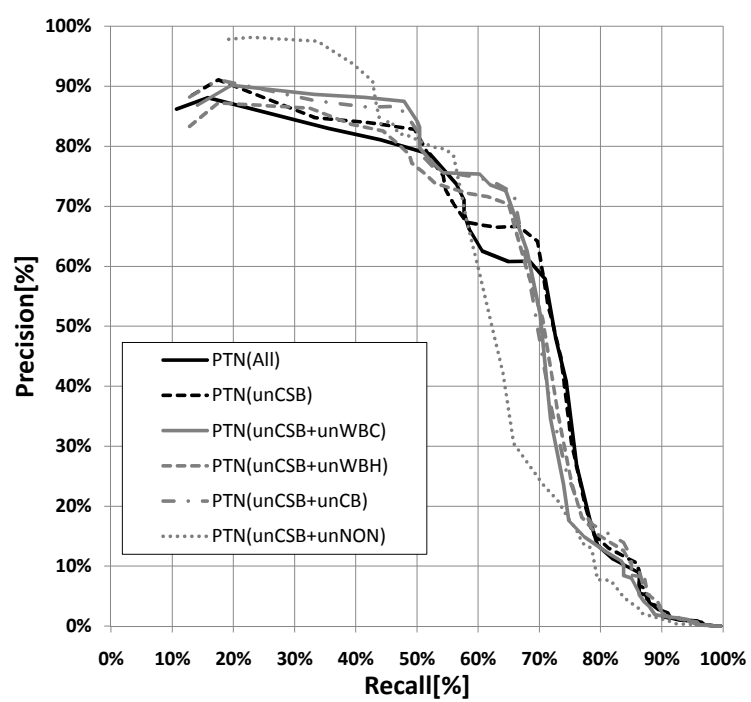


図 4.19: PTN の検索性能の比較

## 第5章 音声中の検索語検出のための検索方法の改善

本章では、複数の音声認識システムの出力を利用したネットワーク型インデックスに対する、検索語の検出方法について述べる。

第4章では、本研究で用意した12種類の音声認識システムの出力をネットワーク型のインデックスとして利用することによって、STD性能が改善されることを述べた。しかし、ネットワーク型インデックスでは、Recallが低い場合においてサブワードベースインデックスより多くの誤検出が発生してしまい、検索性能が低下する傾向にあった。これは、多くの仮説を用いてネットワーク型インデックスを構築した場合に大量の湧き出し誤りが検出されたことから、ネットワーク型インデックスの表現力の高さが影響していることが考えられる。

湧き出し誤りを抑制するために、複数の音声認識システムの出力を利用したネットワーク型インデックスを構築する際に得られる情報が利用することが可能ではないかと考えた。

本章では、複数の音声認識システムの出力を利用したネットワーク型インデックスを構築する際に得られる情報を、誤検出を抑制するパラメータとして利用する手法について述べる。

### 5.1 誤検出抑制パラメータ

複数の音声認識システムの出力を利用したネットワーク型インデックスを構築する際に得られる情報を、語検出を抑制するためのパラメータとして検討した。

誤検出を抑制するために用いたパラメータは次に示す3種類である。

**Voting** : 同じ音素を認識していた音声認識システムの数。

多くの音声認識システムで認識されているほど、その音素の信頼性が高くなる可能性がある

**ArcWidth** : 2 Node間に存在するArcの数。

Arcの数が少なくなるほど、そのNode間の認識結果が信頼性が高くなる可能性がある



**CM スコア**：音素単位の Confusion Matrix (CM).

音声認識における、ある音素が正解・挿入・脱落している確率を用いることによって、その音素の信頼性が類推できる可能性がある

音素単位の CM は、CSJ のコア以外の講演音声で、Non/Tri を用いて認識した音素認識結果から求めた。この音声認識は、CSJ の STD 用テストコレクション [19] の音声認識条件に基づいている。つまり、CSJ のコア以外の講演を奇数講演と偶数講演に分けて学習し、奇数講演は偶数講演で学習したモデルによって認識し、偶数講演は奇数講演で学習したモデルによって認識しているため、オープンな音声認識となっている。

この音素単位の CM から、次に示す 3 種類の情報を誤検出抑制パラメータとして用いる。

$CM_{Del}$ ：ある音素が脱落している確率

$CM_{Ins}$ ：ある音素が挿入している確率

$CM_{Cor}$ ：ある音素が正解している確率

## 5.2 編集距離ベースの誤検出抑制パラメータの組合せによる検索性能 (1)

誤検出抑制パラメータの導入は、編集距離に基づく距離の計算に抑制パラメータに基づくスコアを加味することで、誤検出を考慮した検索エンジンを実現する。

### 5.2.1 誤検出抑制パラメータの導入方法 (1)

各スコアは、式 (5.2) から式 (5.8) に示すように算出され、式 (5.1) に示すように適用される。なお、各スコアは単体または組み合わせて使用することが可能であり、その際は式 (5.1) の該当項が適用されなくなる。

$$D(i, j) = \min \left\{ \begin{array}{l} D(i, j-1) + 1.0 + Cm_{Del}(j) \\ D(i-1, j) + Null(i) \\ \quad + Cm_{Ins}(i, j) \\ D(i-1, j-1) + Match(i, j) \\ \quad + Vot(i, j) + Acw(i) \\ \quad + Cm_{Cor}(i, j) \end{array} \right. \quad (5.1)$$

$$Match(i, j) = \begin{cases} 0.0 : Query(j) \in PTN(i) \\ 1.0 : Query(j) \notin PTN(i) \end{cases} \quad (5.2)$$

$$Null(i) = \begin{cases} 0.1 : NULL \in PTN(i) \\ 1.0 : NULL \notin PTN(i) \end{cases} \quad (5.3)$$

$$Vot(i, j) = \begin{cases} \alpha \div Voting(p) \\ : \exists p \in PTN(i), \\ p = Query(j) \\ 1.0 : Query(j) \notin PTN(i) \end{cases} \quad (5.4)$$

$$Acw(i) = \beta \times ArcWidth(i) \quad (5.5)$$

$$Cm_{Del}(j) = 1.0 - P(\phi, Query(j)) \quad (5.6)$$

$$Cm_{Ins}(i, j) = \begin{cases} 1.0 - P(p, \phi) \\ : \exists p \in PTN(i), \\ p = Query(j) \\ 0.0 : Query(j) \notin PTN(i) \end{cases} \quad (5.7)$$

$$Cm_{Cor}(i, j) = \begin{cases} 1.0 - P(p, Query(j)) \\ : \exists p \in PTN(i), \\ p = Query(j) \\ 0.0 : Query(j) \notin PTN(i) \end{cases} \quad (5.8)$$

$D(i, j)$  は DP 格子上の  $(i, j)$  の位置に至るまでの距離である。

$Query(j)$  は検索語の  $j$  番目の音素を表し、 $PTN(i)$  は PTN の  $i$  番目の Node が持つ Arc の集合を表す。また、 $p$  は PTN の  $i$  番目の Node が持つ、ある Arc の音素を表す。

式 (5.4) の  $\alpha$  は NULL 遷移よりコストを低くするために 0.5 を設定した。 $Voting(p)$  は  $Query(j)$  と一致する  $p$  を認識していた音声認識システムの数を表す。

式 (5.5) の  $\beta$  も NULL 遷移よりコストを低くするために 0.01 を設定した。 $ArcWidth(i)$  は  $PTN(i)$  の Arc の数を表す。

$P(i, j)$  は CM の確率を表し、 $\phi$  は空文字を表す。つまり  $P(i, j)$  において  $i = j$  のとき正解率を表し、 $P(\phi, j)$  のとき  $j$  が脱落する確率、 $P(i, \phi)$  のとき  $i$  が挿入する確率を表す。

### 5.2.2 抑制パラメータの組合せ

検索性能の比較のためのインデックスは、10 種類の音声認識システムの 1-Best 出力を音素単位でネットワーク型インデックスとして構築した PTN である。この PTN は表 5.1 に示す内容で構築されている。

誤検出抑制パラメータを適用させる組合せは、表 5.2 に示す組み合わせとした。

表 5.1: 誤検出抑制パラメータを導入する PTN の構成内容

音声認識システムの種類	N-Best	仮説数
WBC/*, WBH/*, CB/*, BM/*, Non/*	1	10

表 5.2: 誤検出抑制パラメータの組み合わせ (1)

検索方法	誤検出抑制パラメータ
Only EditDist	—————
+ Voting Cost	“Only EditDist” + Voting
+ CM Cost	“Only EditDist” + CM <sub>Cor</sub> + CM <sub>Del</sub> + CM <sub>Ins</sub>
+ ArcWidth Cost	“Only EditDist” + ArcWidth
+ CM <sub>Cor</sub>	“Only EditDist” + CM <sub>Cor</sub>
+ CM <sub>Del</sub>	“Only EditDist” + CM <sub>Del</sub>
+ CM <sub>Ins</sub>	“Only EditDist” + CM <sub>Ins</sub>
+ Vot+Acw Cost	“Only EditDist” + Voting + ArcWidth
+ Vot+CM Cost	“Only EditDist” + Voting + CM <sub>Cor</sub> + CM <sub>Del</sub> + CM <sub>Ins</sub>
+ Acw+CM Cost	“Only EditDist” + ArcWidth + CM <sub>Cor</sub> + CM <sub>Del</sub> + CM <sub>Ins</sub>
+ All Cost	All Parameters

表 5.3: 誤検出抑制パラメータの組み合わせによる検索性能の比較 (1)

検索方法	F-measure	MAP	MRP
Only EditDist	0.64	0.81	0.75
+ Voting Cost	0.71	<b>0.86</b>	<b>0.81</b>
+ CM Cost	0.48	0.71	0.66
+ ArcWidth Cost	0.63	0.82	0.78
+ CM <sub>Cor</sub>	0.62	0.81	0.76
+ CM <sub>Del</sub>	0.64	0.78	0.74
+ CM <sub>Ins</sub>	0.53	0.72	0.63
+ Vot+Acw Cost	<b>0.74</b>	0.85	0.79
+ Vot+CM Cost	0.48	0.74	0.68
+ Acw+CM Cost	0.48	0.71	0.65
+ All Cost	0.48	0.75	0.71

### 5.2.3 評価実験

この比較実験で用いた評価尺度は、Recall-Precision カークと F-measure, MAP, MRP である。

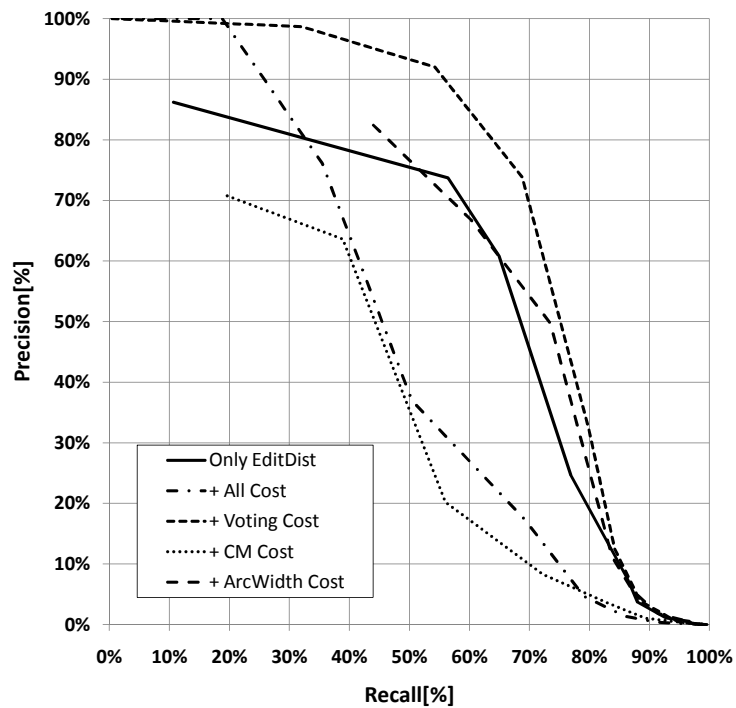


図 5.1: 1 種類の誤検出抑制パラメータを導入した検索性能の比較

表 5.3 に，表 5.2 に示す誤検出抑制パラメータの組み合わせごとの検索性能を示す．

図 5.1 は，編集距離のみに対して 1 種類の誤検出抑制パラメータを加えた場合の Recall-Precision カーブを表す．図 5.2 は，編集距離のみに対して CM から得られる 3 種類の誤検出抑制パラメータをそれぞれ加えた場合の Recall-Precision カーブを表す．図 5.3 は，編集距離のみに対して 2 種類の誤検出抑制パラメータを加えた場合の Recall-Precision カーブを表す．

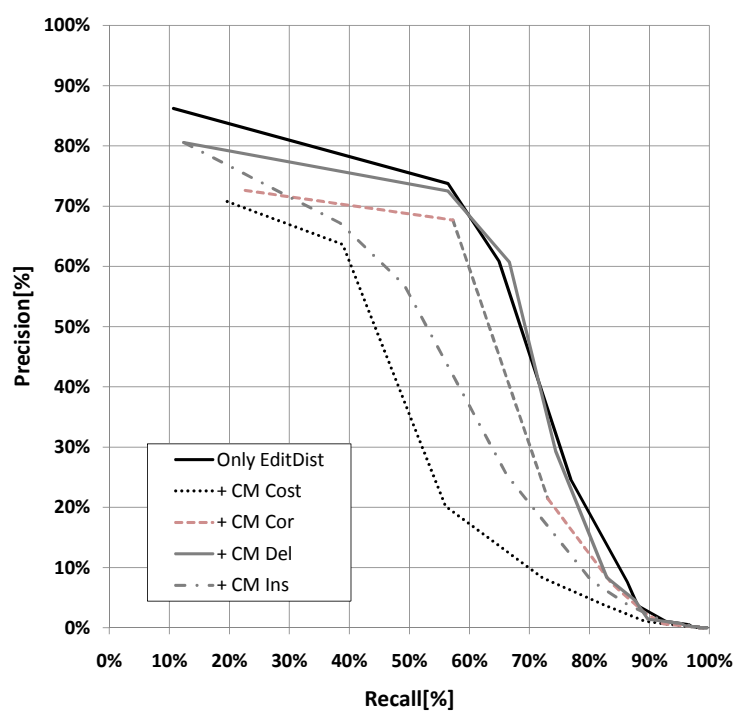


図 5.2: CM スコアを導入した検索性能の比較

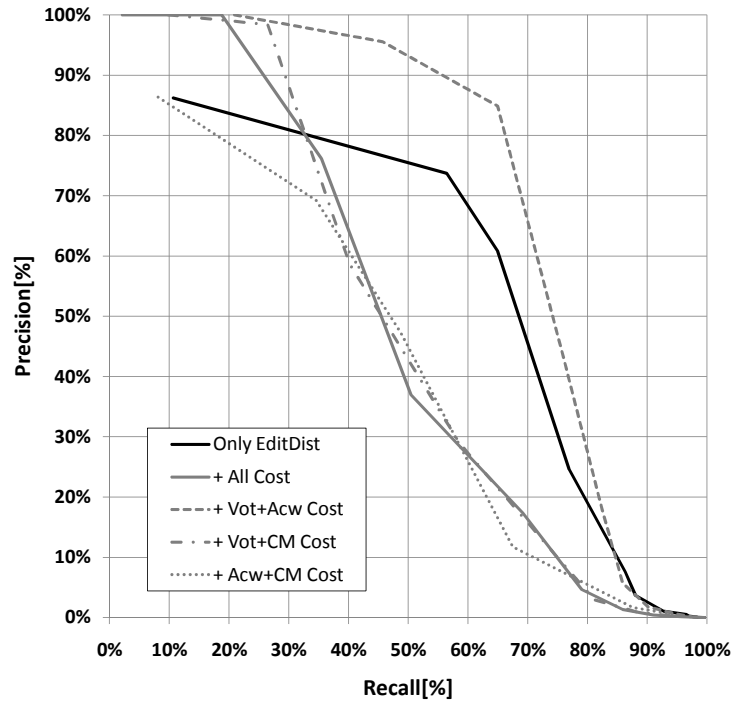


図 5.3: 複数の誤検出抑制パラメータを導入した検索性能の比較

実験結果から示されるように、編集距離と Voting を用いた場合に最も高い検索性能を得ることができた。また、CM スコア以外のパラメータを導入することで検索性能が向上していることから、これらのパラメータを利用することで誤検出が抑制されることが示された。

CM スコアを導入することで検索性能が低下している原因として、距離計算への適用が適切ではなかったことが考えられる。本論文では、編集距離に  $Cm_{Cor}$ ,  $Cm_{Del}$ ,  $Cm_{Ins}$  を加算する形で CM スコアを適用している。 $Cm_{Del}$  と  $Cm_{Ins}$  は  $Cm_{Cor}$  に比べてとても小さい値となり、編集距離に加算する場合には図 4.3 の I, D のコストが高くなりすぎてしまい、PTN と検索語間で適切な距離計算が行えていなかったと考えられる。

この CM スコアの適用を、定数である 1.0 から  $Cm_{Del}$  と  $Cm_{Ins}$  を減算するのではなく、 $Cm_{Del}$  と  $Cm_{Ins}$  をそのまま利用することで、検索性能の向上が期待できる。

## 5.3 編集距離ベースの誤検出抑制パラメータの組合せによる検索性能 (2)

前節の実験結果を踏まえ、CM スコアの導入方法を再検討した。

### 5.3.1 誤検出抑制パラメータの導入方法 (2)

各スコアの基本的な導入方法は変わらず、式 (5.1) に示すように適用される。CM スコアを利用する際の計算式である、式 (5.6) から式 (5.8) を、式 (5.9) から式 (5.11) に示すように変更した。

$$Cm_{Del}(j) = P(\phi, Query(j)) \quad (5.9)$$

$$Cm_{Ins}(i, j) = \begin{cases} P(p, \phi) \\ \quad : \exists p \in PTN(i), \\ \quad \quad p = Query(j) \\ 0.0 : Query(j) \notin PTN(i) \end{cases} \quad (5.10)$$

$$Cm_{Cor}(i, j) = \begin{cases} 1.0 - P(p, Query(j)) \\ \quad : \exists p \in PTN(i), \\ \quad \quad p = Query(j) \\ 0.0 : Query(j) \notin PTN(i) \end{cases} \quad (5.11)$$

### 5.3.2 抑制パラメータの組合せ

検索性能の比較のためのインデックスは前節と同様に、10 種類の音声認識システムの 1-Best 出力を音素単位でネットワーク型インデックスとして構築した PTN である。誤検出抑制パラメータを適用させる組合せは、表 5.4 に示す組み合わせとした。

### 5.3.3 評価実験

この比較実験で用いた評価尺度は、Recall-Precision カーブと F-measure, MAP, MRP である。

表 5.5 に、表 5.4 に示す誤検出抑制パラメータの組み合わせごとの検索性能を示す。

図 5.4 は、編集距離のみに対して 1 種類の誤検出抑制パラメータを加えた場合の Recall-Precision カーブを表す。図 5.5 は、編集距離のみに対して CM から得られる 3 種類の誤検出抑制パラメータをそれぞれ加えた場合の Recall-Precision カーブを表す。図 5.6 は、編集距離に Voting を加味した距離計算に対して、CM スコアを加えた場合の Recall-Precision カーブを表す。図 5.7 は、編集距離に ArcWidth を加味した距離計算に対して、CM スコアを加えて場合の Recall-Precision カーブを表す。図 5.8 は、編集距離に

表 5.4: 誤検出抑制パラメータの組み合わせ (2)

検索方法	誤検出抑制パラメータ
Only EditDist	—————
+ Voting Cost	“Only EditDist” + Voting
+ CM Cost	“Only EditDist” + CM <sub>Cor</sub> + CM <sub>Del</sub> + CM <sub>Ins</sub>
+ ArcWidth Cost	“Only EditDist” + ArcWidth
+ CM <sub>(cor)</sub>	“Only EditDist” + CM <sub>Cor</sub>
+ CM <sub>(del)</sub>	“Only EditDist” + CM <sub>Del</sub>
+ CM <sub>(ins)</sub>	“Only EditDist” + CM <sub>Ins</sub>
+ Vot+CM Cost	“Only EditDist” + Voting + CM <sub>Cor</sub> + CM <sub>Del</sub> + CM <sub>Ins</sub>
+ Vot+CM <sub>(cor)</sub>	“Only EditDist” + Voting + CM <sub>Cor</sub>
+ Vot+CM <sub>(del)</sub>	“Only EditDist” + Voting + CM <sub>Del</sub>
+ Vot+CM <sub>(ins)</sub>	“Only EditDist” + Voting + CM <sub>Ins</sub>
+ Acw+CM Cost	“Only EditDist” + ArcWidth + CM <sub>Cor</sub> + CM <sub>Del</sub> + CM <sub>Ins</sub>
+ Acw+CM <sub>(cor)</sub>	“Only EditDist” + ArcWidth + CM <sub>Cor</sub>
+ Acw+CM <sub>(del)</sub>	“Only EditDist” + ArcWidth + CM <sub>Del</sub>
+ Acw+CM <sub>(ins)</sub>	“Only EditDist” + ArcWidth + CM <sub>Ins</sub>
+ Vot+Acw Cost	“Only EditDist” + Voting + ArcWidth
+ Vot+Acw+CM <sub>(cor)</sub>	“Only EditDist” + Voting + ArcWidth + CM <sub>Cor</sub>
+ Vot+Acw+CM <sub>(del)</sub>	“Only EditDist” + Voting + ArcWidth + CM <sub>Del</sub>
+ Vot+Acw+CM <sub>(ins)</sub>	“Only EditDist” + Voting + ArcWidth + CM <sub>Ins</sub>
+ All Cost	All Parameters

Voting と ArcWidth を加味した距離計算に対して 1 種類の誤検出抑制パラメータを加えた場合の Recall-Precision カーブを表す。図 5.9 は、編集距離のみに対して 2 種類の誤検出抑制パラメータを加えた場合の Recall-Precision カーブを表す。

実験結果から示されるように、CM スコアを利用する計算式を変更したことによって、検索性能の低下が軽減された。また、MRP においては、編集距離のみの場合と比較して、検索性能が僅かではあるが向上していることがわかる。

以上より、CM スコアの導入方法を検討することによって、更なる検索性能の改善が行える可能性が示された。



表 5.5: 誤検出抑制パラメータの組み合わせによる検索性能の比較 2

検索方法	F-measure	MAP	MRP
Only EditDist	0.64	0.81	0.75
+ Voting Cost	0.71	<b>0.86</b>	<b>0.81</b>
+ CM Cost	0.60	0.80	0.76
+ ArcWidth Cost	0.63	0.82	0.78
+ CM <sub>(cor)</sub>	0.62	0.81	0.76
+ CM <sub>(del)</sub>	0.64	0.80	0.76
+ CM <sub>(ins)</sub>	0.63	0.80	0.75
+ Vot+CM Cost	0.61	0.82	0.77
+ Vot+CM <sub>(cor)</sub>	0.64	0.82	0.77
+ Vot+CM <sub>(del)</sub>	0.71	0.85	0.79
+ Vot+CM <sub>(ins)</sub>	0.71	<b>0.86</b>	0.79
+ Acw+CM Cost	0.58	0.81	0.76
+ Acw+CM <sub>(cor)</sub>	0.58	0.81	0.77
+ Acw+CM <sub>(del)</sub>	0.63	0.82	0.78
+ Acw+CM <sub>(ins)</sub>	0.63	0.82	0.76
+ Vot+Acw Cost	<b>0.74</b>	0.85	0.79
+ Vot+Acw+CM <sub>(cor)</sub>	0.64	0.82	0.78
+ Vot+Acw+CM <sub>(del)</sub>	0.73	0.85	0.79
+ Vot+Acw+CM <sub>(ins)</sub>	0.72	0.85	0.78
+ All Cost	0.62	0.82	0.79

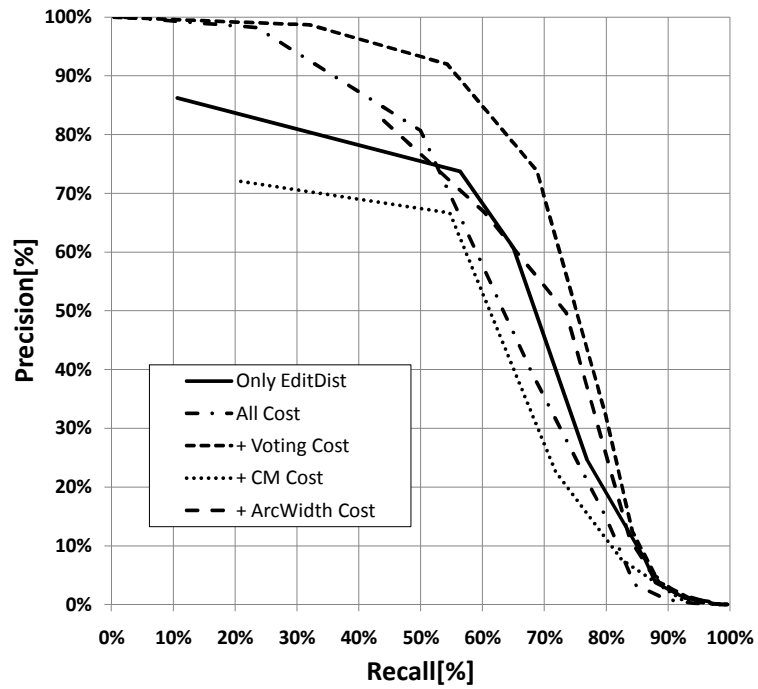


図 5.4: 1 種類の誤検出抑制パラメータを導入した検索性能の比較

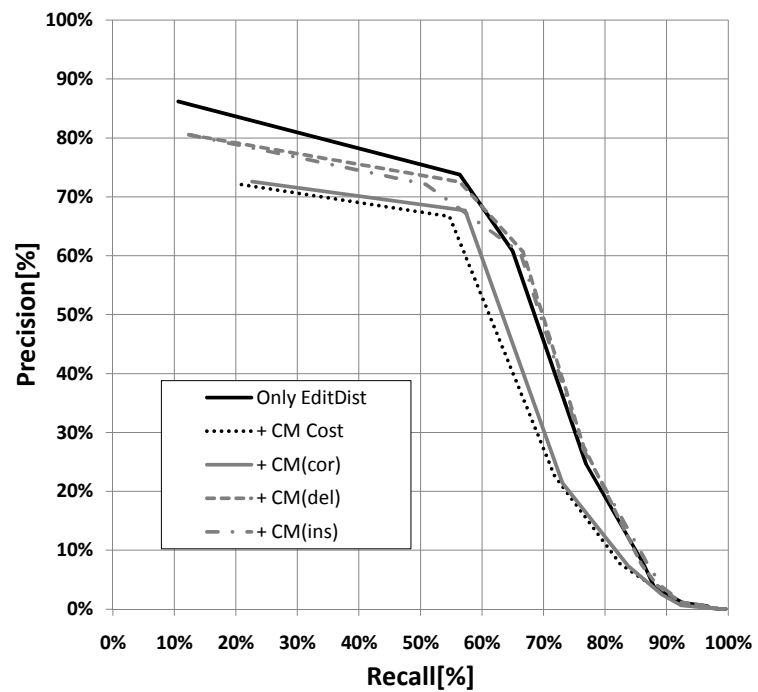


図 5.5: CM スコアを導入した検索性能の比較

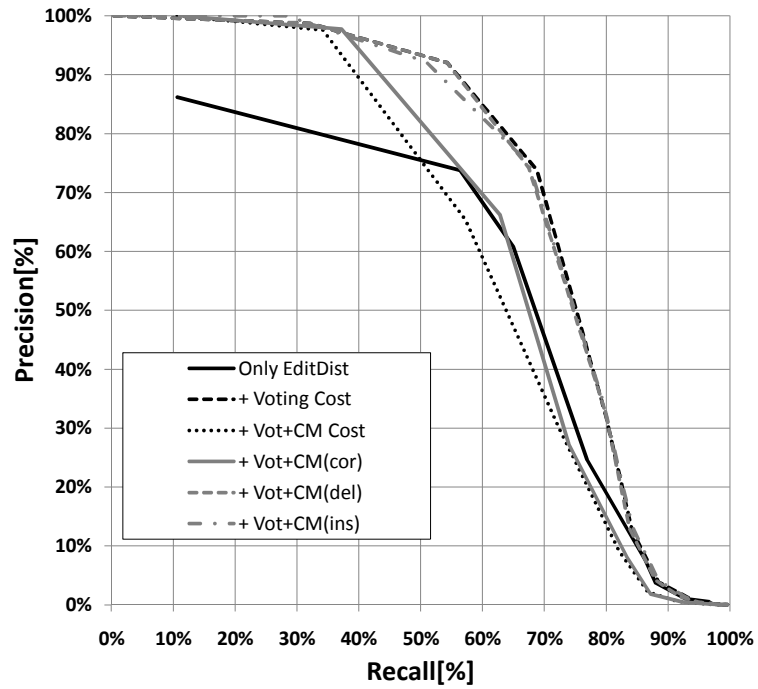


図 5.6: Voting に CM スコアを導入した検索性能の比較

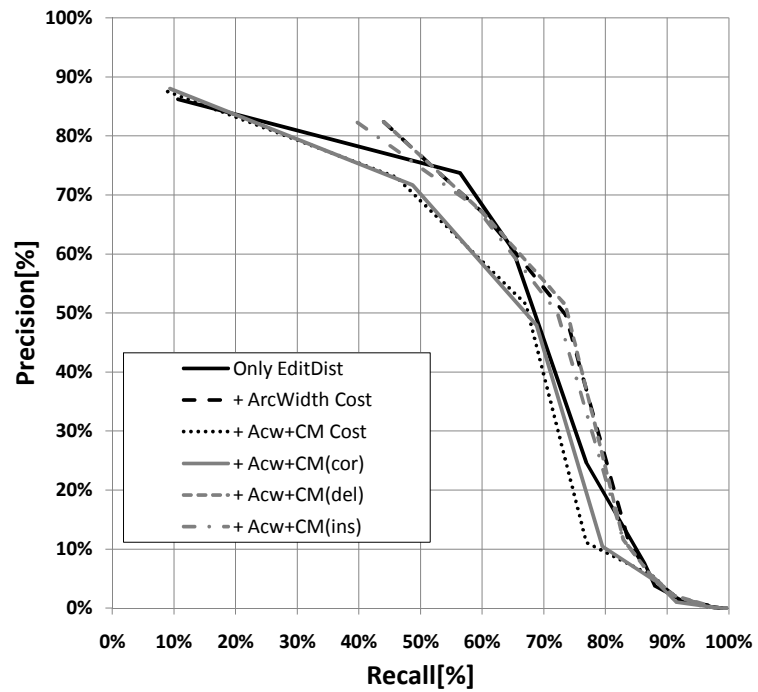


図 5.7: ArcWidth に CM スコアを導入した検索性能の比較

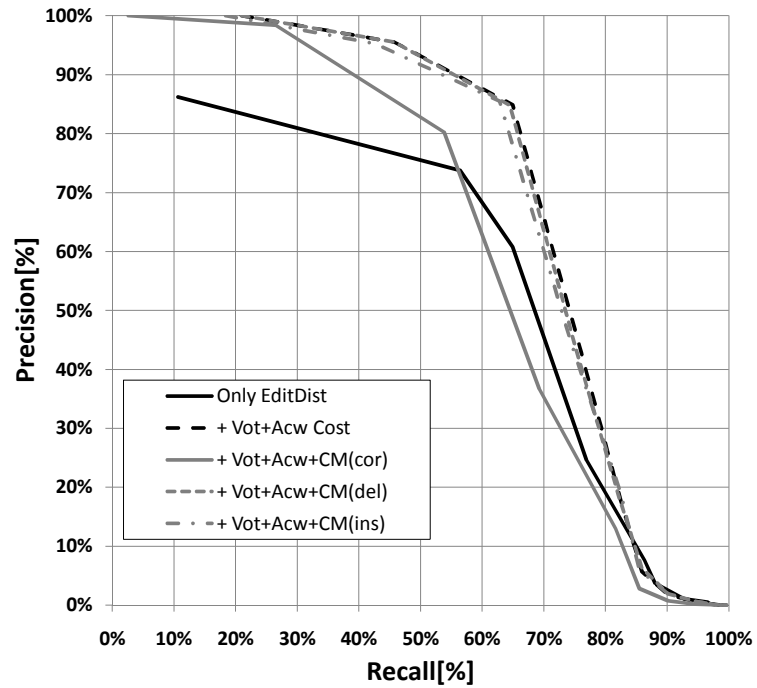


図 5.8: Voting と ArcWidth に CM スコアを導入した検索性能の比較

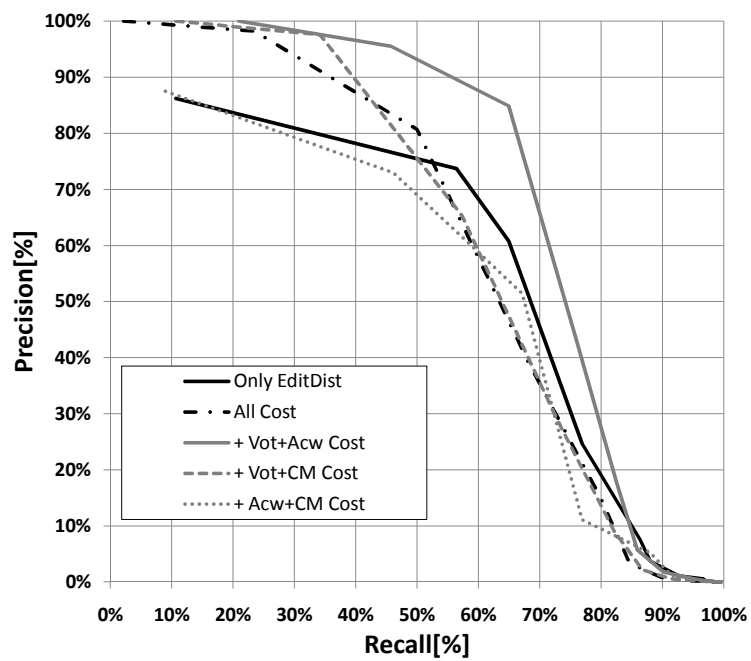


図 5.9: 複数の誤検出抑制パラメータを導入した検索性能の比較

## 5.4 まとめ

本章では，複数の音声認識システムの出力を利用したネットワーク型インデックスに対する，誤検出を抑制した検索語の検出方法について述べた．

誤検出を抑制するパラメータとして，3種類の情報を利用し，導入の方法によって計5種類の誤検出抑制パラメータを検討した．これらの誤検出抑制パラメータを，DPの距離計算式に導入することによって，誤検出が抑制されることが実験結果より示された．

特に，音素を認識した音声認識システムの数である Voting を導入することによって，大幅に検索性能が改善された．他のパラメータにおいても，編集距離のみを用いた DP の距離計算式を用いた場合より誤検出が抑制され，検索性能が改善された．しかし，CM スコアに関しては，導入する方法を再検討した結果，MRP において僅かに改善された程度であった．

これらの誤検出抑制パラメータの導入方法をさらに検討することによって，STD の性能がより改善されることが期待できる．

本手法は，2011 年 12 月に開催された NTCIR-9 (Proceedings of NTCIR-9 Workshop Meeting) の STD タスクにおいて最も優れた検索性能を示した [47]．

## 第6章 音声中の検索語検出のための誤検出を改善する手法

本章では、検索語長や複数の音声認識システムの出力を利用したネットワーク型インデックスの複雑さに着目した、検索語の検出方法について述べる。

第5章では、複数の音声認識システムの出力を利用したネットワーク型インデックスを構築する際に得られる情報を、誤検出を抑制するパラメータとして利用する方法について述べた。これにより、誤検出を抑制することが可能となり、さらに高い検索精度を実現することができた。

本章では、誤検出を抑制するパラメータのより効果的な利用方法について述べる。まず、検索語の音素長による誤検出の傾向を調査した結果について述べる。次に、複数の音声認識システムの出力を利用したネットワーク型インデックスの複雑さに着目した、検索語の検出方法について述べる。

また、音声中の検索語検出のタスクの一つである iSTD タスクに対して本提案手法が有効であるか評価を行い、その結果について述べる。

### 6.1 検索語長の誤検出傾向に着目した検索語の検出方法

第5章では、複数の音声認識システムの出力を利用したネットワーク型インデックスを構築する際に得られる情報を、誤検出を抑制するパラメータとして利用する方法について述べた。このパラメータは、検索語の音素長に関わらず一定の割合で誤検出抑制パラメータを適応していた。

しかし、検索語の特性として音素長が短い検索語は検出され易く誤検出が多く、また音素長が長い検索語は誤検出が少ないことが予測される。そこで、編集距離のみの STD 性能に着目し、検索語の音素長による検索性能を比較する。

#### 6.1.1 検索語の音素長による検索性能

表 6.1 は、第5章で行った評価実験結果の内、編集距離のみを用いた結果を検索語を構成する音素長が 10 以上 / 10 未満で分類したときの検索性能を示している。また、表 6.2 は検索語の出現数と STD によって検索語を検出した数を示している。式 (5.1) を用いて DP スコアを計算したときに、検索語の音素長で正規化した DP スコアが 0.10,

表 6.1: “Only EditDist” における音素長別の STD 性能

閾値	音素長	Recall	Precision	F-measure
0.10	10 以上	0.68	0.89	0.77
	10 未満	0.64	0.44	0.52
	合計	0.66	0.61	0.63
0.15	10 以上	0.80	0.66	0.72
	10 未満	0.76	0.16	0.27
	合計	0.78	0.27	0.40

表 6.2: “Only EditDist” における音素長別の STD 性能

閾値	音素長	正解出現数	検出数	正解数	誤検出数
0.10	10 以上	123	93	83	10
	10 未満	110	158	70	88
	合計	233	251	153	98
0.15	10 以上	123	148	98	50
	10 未満	110	514	83	431
	合計	233	662	181	481

0.15 以下の場合 (これは Recall-Precision カーブでの Recall が 0.6~0.8 の F-measure が最も良くなる領域の閾値となる), 検索語が検出されたと判断している.

表 6.2 より, クエリの音素長が 10 未満になると, 誤検出数が明らかに増加していることが分かる. 特に検出の閾値を 0.10 → 0.15 と緩くすることで, 音素長が短いクエリの誤検出が 88 → 431 と大幅に増加している.

PTN 型インデックスを用いた場合は, その表現能力の高さが悪影響を及ぼしていると考えられる. 短い音素長の検索語を検索することを考えた場合, 対象の検索語と全く同じ, もしくは数音素だけが異なる音素列パターンがネットワーク上に存在している状況では, その音素列パターンの箇所が誤検出されてしまう.

### 6.1.2 検索語の音素長に対する遷移コストの適応

検索語の音素長による検索性能の比較結果より, 音素長の長さに応じて, 以下のよう DP の遷移コストを加味することで誤検出を抑制することを検討した.

挿入・脱落・置換コストを変動させる

NULL 遷移に対する遷移コストを変動させる

短い音素長の検索語を検出する場合は、より高い遷移コストを与えることで、完全一致に近い場合に限り検出を許可する。また、NULL 遷移についても、短い音素列の検索語においては悪影響が大きいことから、これに対する遷移コストを変動させる。

NULL 遷移を低コストで許すことで、PTN はより高い音素列パターンを表現することが可能となるが、検索語の音素長が短い場合には逆に誤検出の原因となりうる例が多い。

そこで、NULL 遷移のコストに対して多数決の要素を含めることを検討した。すなわち、より多くの認識システムが NULL 判定をするほど、その NULL 遷移は信頼性が高いと判定し NULL 遷移コストを低く設定する。反対に、NULL 判定の認識システムが少なければ、NULL 遷移は信頼できないとし、コストを大きく設定する。

以上をまとめると、検索語の音素長 (10 以上 / 10 未満) に応じて、DP の脱落・挿入・置換の遷移コスト、NULL 遷移を変化させることで、特に音素長が短いクエリについての検索性能を改善させる。

この処理を導入した用語検索エンジンにおける DP コストの計算式を式 (6.1)～式 (6.7) に示す。なお、第 5 章で述べた誤検出抑制パラメータ “CM スコア” については、誤検出を抑制する効果が薄いことから、本章では考慮しない。

$$D(i, j) = \min \begin{cases} D(i, j-1) + Del \\ D(i-1, j) + Null(i) + Null_V(i) \\ D(i-1, j-1) + Match(i, j) + Vot(i, j) + Acw(i) \end{cases} \quad (6.1)$$

$$Match(i, j) = \begin{cases} 0.0 : Query(j) \in PTN(i) \\ 1.0 : Query(j) \notin PTN(i), J \geq 10 \\ 1.5 : Query(j) \notin PTN(i), J < 10 \end{cases} \quad (6.2)$$

$$Del = \begin{cases} 1.0 : J \geq 10 \\ 1.5 : J < 10 \end{cases} \quad (6.3)$$

$$Null(i) = \begin{cases} 0.1 : NULL \in PTN(i) \\ 1.0 : NULL \notin PTN(i), J \geq 10 \\ 1.5 : NULL \notin PTN(i), J < 10 \end{cases} \quad (6.4)$$

$$Null_V(i) = \begin{cases} \frac{\alpha}{Voting(NULL)} : NULL \in PTN(i), J \geq 10 \\ \frac{\beta}{Voting(NULL)} : NULL \in PTN(i), J < 10 \\ 1.0 : NULL \notin PTN(i), J \geq 10 \\ 1.5 : NULL \notin PTN(i), J < 10 \end{cases} \quad (6.5)$$



$$Vot(i, j) = \begin{cases} \frac{\gamma}{Voting(p)} : \exists p \in PTN(i), p = Query(j) \\ 0.0 : Query(j) \notin PTN(i) \end{cases} \quad (6.6)$$

$$Acw(i) = \delta \times ArcWidth(i) \quad (6.7)$$

ここで、 $D(i, j)$  は格子点  $(i, j)$  に至るまでのトータルコスト、 $Del$  は脱落誤りコスト、 $Null(i)$  は NULL 遷移コスト、 $Match(i, j)$  は置換誤りコスト、 $Vot(i, j)$  および  $Acw(i)$  は誤検出抑制パラメータである。  $D(i, j)$  は DP 格子上の  $(i, j)$  の位置に至るまでの距離である。

$J$  は検索語の音素長を表し、 $Query(j)$  は検索語の  $j$  番目の音素を表す。また、 $PTN(i)$  は PTN の  $i$  番目のノードが持つアークの集合を表し、 $p$  は PTN の  $i$  番目のノードが持つ、あるアークの音素を表す。  $Voting(NULL)$  は NULL 遷移を認識していた音声認識システムの数を表す。式 (6.5) における  $\alpha$  と  $\beta$  は、テストセットに応じて最適な値を設定する。

脱落・置換誤りに対してコストを 1、正解の場合は 0 と設定したときの編集距離を  $ED_f$  と示す。

クエリの音素長によって遷移コストを変化させるため、式 (6.2)、式 (6.3) に示すように、音素長が 10 未満の場合の脱落・置換誤りコストを 1.5 に設定し、10 以上の場合は 1 に設定する。この場合の編集距離を  $ED_d$  と記す。

挿入誤りは NULL 遷移に基づく。第 4 章、第 5 章では、NULL 遷移コストは 0.1 に固定していたが (これを  $NULL_f$  と示す)、検索語の音素長によってコストを変化させる ( $NULL_d$  と示す)。また、NULL 遷移にも多数決の概念を導入し、複数の音声認識システムの NULL 判定に応じて NULL 遷移コストを変化させる。これを  $Null_v(i)$  とする (式 (6.5))。なお、式 (6.1) において、 $Null(i)$  と  $Null_v(i)$  は排他的に利用する。すなわち、どちらかのパラメータを利用するともう一方は 0 になる。

誤検出抑制パラメータについては、式 (6.6)、式 (6.7) が該当する。式 (6.6) の  $\gamma$  は、音素長が 10 以上の検索語において、5 つ以上の音声認識システムが同じ音素を認識していた場合に、0.1 よりコストを低くするために 0.5 を設定した。  $Voting(p)$  は  $Query(j)$  と一致する  $p$  を認識していた音声認識システムの数を表す。式 (6.7) の  $\delta$  は、音素長が 10 以上の検索語において、全ての音声認識システムが異なる音素を認識していた場合を除き、0.1 よりコストが低くなるように 0.01 を設定した。  $ArcWidth(i)$  は  $PTN(i)$  のアークの数を表す。

STD 性能の比較を行うパラメータの組み合わせを表 6.3 に示す。Voting は、Voting によってコストを決定することを示す。表 6.3 の Only EditDist をベースラインとする。

### 6.1.3 評価実験

図 6.1 に検索語の音素数に応じたパラメータ適応後の Recall-Precision カーブを、表 6.4 にグラフ上において最も高い F-measure と MAP 値を示す。

表 6.3: 探索パラメータの組み合わせ

検索方法	探索パラメータ
Only EditDist	$ED_f + NULL_f$
Voting1	$ED_f + NULL_f + \text{Voting}$
Voting2	$ED_d + NULL_f + \text{Voting}$
Voting3	$ED_d + NULL_d + \text{Voting}$
Vot+Acw1	$ED_f + NULL_f + \text{Voting} + \text{ArcWidth}$
Vot+Acw2	$ED_d + NULL_f + \text{Voting} + \text{ArcWidth}$
Vot+Acw3	$ED_d + NULL_d + \text{Voting} + \text{ArcWidth}$

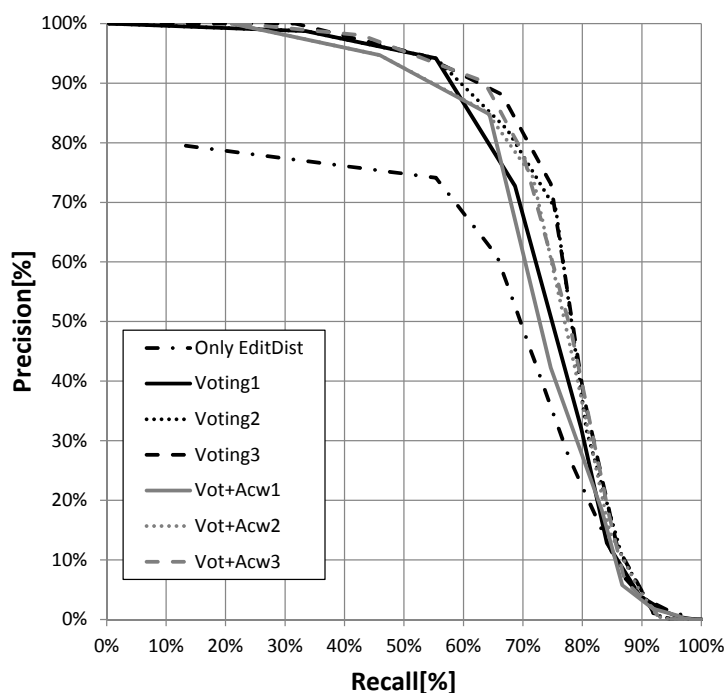


図 6.1: 検索語の音素長に応じたパラメータ適応による検索性能の比較 (Recall-Precision カurve)

Only EditDist に比べて Voting1 と Vot+Acw1 の検索性能が向上していることから、誤検出抑制パラメータを使用することで、Recall が 80% 以下の領域において大幅に検索性能が改善していることが確認できる。また、Voting1 や Vot+Acw1 に比べて Voting2 や Vot+Acw2 の検索性能が向上していることから、音素長が 10 未満の検索語に対する編集距離のコストを高くすることで、湧き出し誤検出を抑制し、Recall が 70% 前後で検索性能が改善していることが確認できる。さらに、Voting2 や Vot+Acw2 に比べて Voting3 や Vot+Acw3 の検索性能が向上していることから、検索性能が改善してい

表 6.4: 検索語の音素長に応じたパラメータ適応による検索性能の比較 (F-measure と MAP)

パラメータ	全体		10 音素未満		10 音素以上	
	F-measure	MAP	F-measure	MAP	F-measure	MAP
Only EditDist	0.63	0.80	0.59	0.60	0.77	0.89
Voting1	0.71	0.87	0.70	0.80	0.79	0.90
Voting2	0.74	0.87	0.70	0.80	0.79	0.90
Voting3	0.75	0.87	0.72	0.81	0.78	0.89
Vot+Acw1	0.73	0.86	0.71	0.80	0.79	0.89
Vot+Acw2	0.73	0.86	0.70	0.80	0.79	0.89
Vot+Acw3	0.75	0.87	0.72	0.81	0.77	0.89

ることが確認できる。これは、NULL 遷移のコストを Voting によって決定することで、個々の NULL 遷移の信頼度に応じたコストを与えることができたためであると考えられる。また、音素長が 10 未満の検索語に対する NULL 遷移のコストを高くすることで、F-measure が最大となる閾値を高くすることが可能であることが確認できる。また、音素長が 10 以上の検索語における F-measure が最大となる閾値に近づけることが可能であることが確認できる。

図 6.2, 図 6.3 に、それぞれ音素長が 10 未満, 10 以上の検索語について求めた Recall-Precision カーブを示す。この結果を明らかであるように、音素長が 10 未満の検索語について、誤検出抑制パラメータ (Voting) を導入することによって STD 性能が大幅に改善されている。Voting1 → Voting3 の比較や Vot+Acw3 の結果より、誤検出抑制パラメータや NULL 遷移のコストを調整することで、更なる改善が得られていることが確認できる。

一方で、音素長が 10 以上の検索語についても、F-measure 等の性能改善は達成されたが、短い音素長がの検索語と比べ改善の幅は僅かである。

以上より、誤検出抑制パラメータを使用することで誤検出が減少し、検索性能が改善することが示された。また、音素長が 10 未満の検索語に対して編集距離のコストを高くすることで、検索性能が改善することが示された。さらに、NULL 遷移のコストを Voting によって決定することで、NULL 遷移の信頼度に応じたコストを与えることが可能であり、検索性能が改善することが示された。このとき、音素長が 10 未満の検索語に対する NULL 遷移のコストを高く設定することで、F-measure が最大となる閾値を高くすることが可能となり、音素長が 10 以上の検索語における F-measure が最大となる閾値に近づけることで、検索性能が改善することが示された。

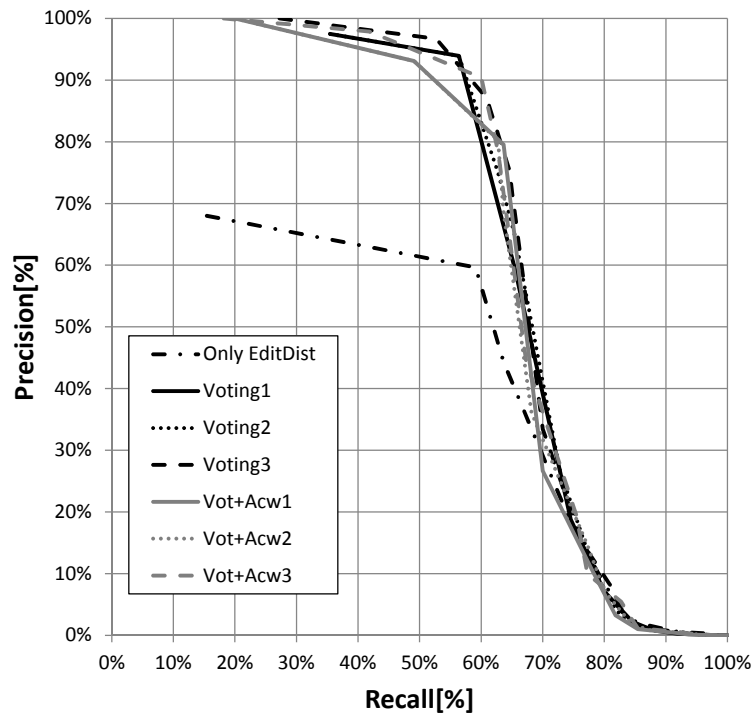


図 6.2: 音素長が 10 未満の検索語に対する検索語の音素長に応じたパラメータ適応による検索性能の比較 (Recall-Precision カーブ)

## 6.2 ネットワーク型インデックスの複雑さに着目した検索語の検出方法

誤検出を抑制するパラメータとして、ArcWidth(Arc の幅) を利用している。このパラメータは 2 Node 間に存在する Arc の数を指している。この Arc の数が少なくなるほど、その Node 間の認識結果が信頼性が高くなる可能性がある。

しかし、単純に Arc の幅のみではネットワーク型インデックスの複雑さを有効活用できていないことが考えられる。これは、第 5 章の誤検出抑制パラメータを単体で用いた場合に、Voting と ArcWidth の STD 性能がほぼ同等であることや Voting+ArcWidth の STD 性能が突出して高くないことから推測される。

そこで、ネットワーク型インデックスの「複雑さ」に着目し、誤検出を抑制することが可能ではないかと検討した。複数の音声認識システムの出力から信頼度を得る手法は、単語の事後確率がしばしば利用される。他の研究では、Varadarajan らは認識器から得られたそれぞれの発話の単語ラティスエントロピーを利用している [68]。また、濱中らは複数の音声認識システムのエントロピーを利用することで、音声認識性能の向上を試みている [70]。

ネットワーク型インデックスにおいてもエントロピーを利用することが可能と考え、

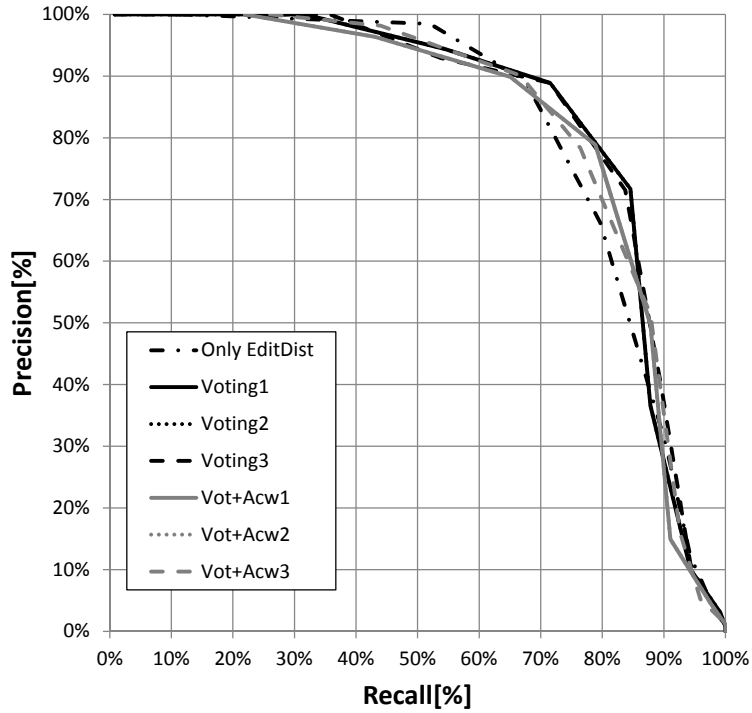


図 6.3: 音素長が 10 未満の検索語に対する検索語の音素長に応じたパラメータ適応による検索性能の比較 (Recall-Precision カーブ)

エントロピーを分析し、その結果を利用することで更なる検索精度向上が期待できるのではないかと仮説を立てた。

### 6.2.1 ネットワーク型インデックスのエントロピー

PTN のエントロピーは、任意の 2 ノード間に存在する音素の数と事後確率を用いて求められる。音素の事後確率は、その音素を出力した音声認識システムの数に基づいて計算する。これは前述した誤検出抑制パラメータに基づいている。

PTN のエントロピーは次の式で計算する。

$$VE_i = - \sum_{j=1}^{J_i} \frac{Voting(p_{ij})}{R} \log_2 \frac{Voting(p_{ij})}{R} \quad (6.8)$$

$$PE = \frac{1}{I-1} \sum_{i=1}^{I-1} VE_i \quad (6.9)$$

式 (6.8) で PTN の任意のノード間 ( $i$  番目と  $i+1$  番目、本稿では便宜上  $i$  番目と記す) のエントロピー (Voting Entropy: VE) を求め、PTN 全体のエントロピー (PTN Entropy:

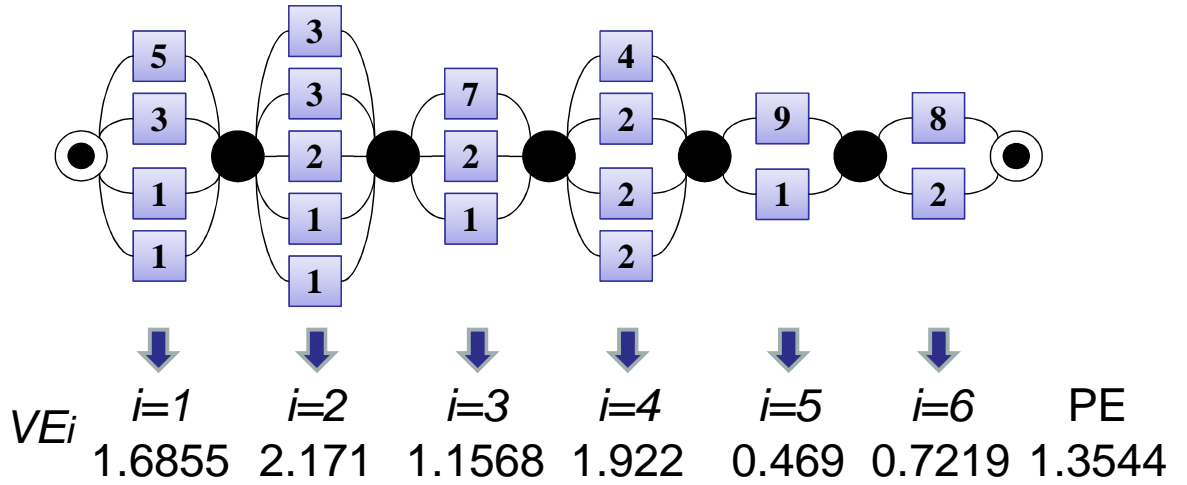


図 6.4: PTN のエントロピーのイメージ

PE) は式 (6.9) で算出する．ここで， $p_{ij}$  は， $i$  番目のノードの  $j$  番目の音素を示し， $J_i$  は  $i$  番目のノードの音素数を表す． $I$  は PTN が持つノード数である． $Voting(p_{ij})$  は，音素  $p_{ij}$  を出力した音声認識システムの数を示す． $R$  は，PTN を作成するために用いた音声認識システムの総数である．式 (6.8) と式 (6.9) のイメージを図 6.4 に示す．

なお，式 (6.9) では音声 1 発話分の音声認識結果から構成した PTN のエントロピーを計算していることになるが，これを，

$$PE = \frac{1}{T-1} \sum_{i=t_s}^{t_e-1} VE_i \quad (6.10)$$

とすることで，ある検索語  $t$  が含まれる区間のみのエントロピーを計算することができる．ここで， $t_s$  は，検索語  $t$  が検出されたときの先頭のノード， $t_e$  は末尾のノードを表わす． $T$  は  $t$  の検出された音素数である．式 (6.10) のイメージを図 6.5 に示す．

### 6.2.2 検索語が含まれる区間のエントロピー

STD のテストセットに含まれる検索語が存在する区間に対して PTN のエントロピーを調査した．使用したテストセットは，日本語 STD テストコレクション [19] のうち 50 検索語から成るコア講演用未知語テストセット (CORE) と，NTCIR-9 SpokenDoc のフォーマルランテストセット [47] である．NTCIR-9 のテストセットは，未知語 (NTCIR\_OOV) と既知語 (NTCIR\_IV) に分けて分析する．

各テストセットに含まれる検索語が存在する区間に対して PTN のエントロピーを集計したものを表 6.5 に示す．

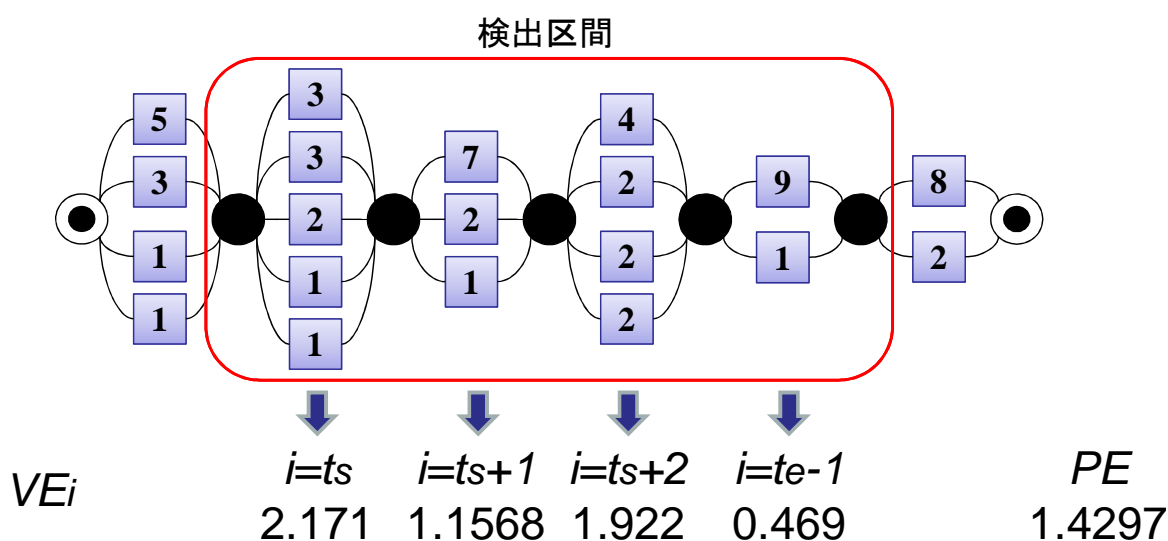


図 6.5: PTN のエントロピーのイメージ (検索語検出区間)

表 6.5: 検索語が存在する区間の PTN エントロピー

テストセット	検索語数	PE(平均)
CORE	233	0.63
NTCIR_OOV	195	0.60
NTCIR_IV	167	0.48

表 6.5 において, CORE および NTCIR\_OOV と比較すると NTCIR\_IV の PE は低くなっている. これは, 未知語が含まれる区間は音声認識システム間の出力結果に揺れが大きいことに起因する. すなわち, 未知語は単語認識ができないためアーク数の多い PTN が構築され易く, 既知語は単語認識が可能であるため未知語と比べるとアーク数の少ない PTN が構築されやすいためだと考えられる. 従って, 未知語と既知語では, 既知語を含む発話から構成される PTN の方が PE が小さくなり, 情報量の観点から見ても, 未知語より既知語の方が検出しやすいという結果が導き出せる.

そこで, 未知語の検索語が含まれる区間のエントロピーが高くなる事実を踏まえ, 未知語が検出された際にその区間のエントロピーをチェックし, エントロピーがある設定閾値よりも低いようであれば誤検出であるという仮説を立て, 誤検出抑制に利用できないかどうかを検証した.

### 6.2.3 評価実験

検索性能の評価には, Recall-Precision カーブ, カーブ上での最大の F-measure を用いた. 図 6.6 に, CORE テストセット, NTCIR-9 のフォーマルランセットを対象とし

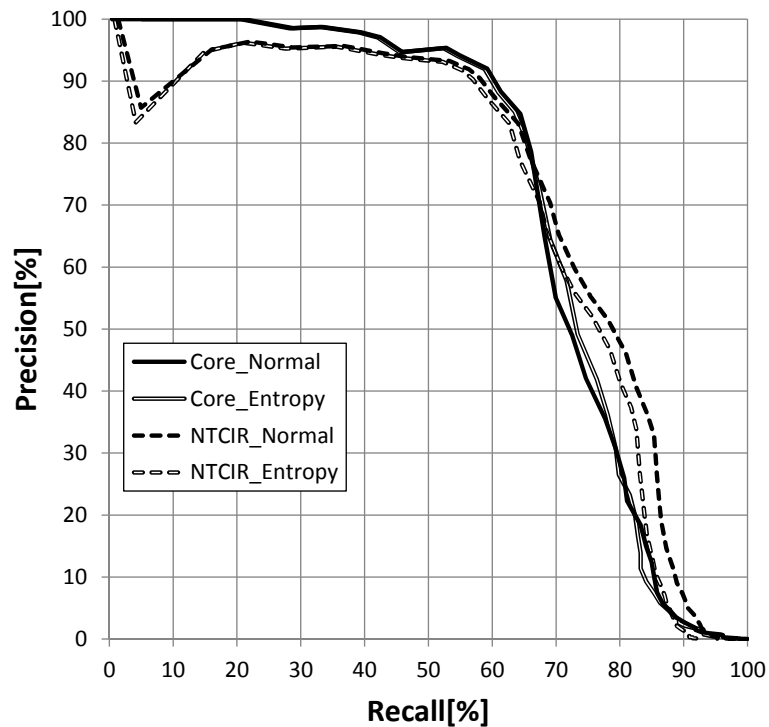


図 6.6: エントロピーを導入した際の検索性能の比較 (Recall-Precision カーブ)

た際の、エントロピーを用いて誤検出抑制を行った場合 (w/ Entropy), 行わなかった場合 (w/o Entropy) の Recall-Precision カーブを示す。

実験では、STD 検出コストの閾値に連動する形で、足切りするエントロピーの閾値を変化させた。図 6.7 は、CORE テストセットにおいて、正しく検出された検索語の STD 検出コストとエントロピーの関係を散布図で表したものである。横軸は STD 検出コスト<sup>1</sup>、縦軸がエントロピーである。図 6.7 で一次直線が引いてあるが、これが足切りに利用するエントロピーの閾値を示している。すなわち、検索語が検出された箇所がこの一次直線より上に位置する場合、その検出箇所は誤検出であると見なす。

CORE テストセットに対しては、図 6.6 を見ても明らかであるように、最大の F-measure が得られる辺りにおいては、正解検出が誤ってリジェクトされてしまったため、Recall が若干低下している。結果として、最大の F-measure が 0.732 から 0.727 へと低下した。しかし、Recall が 65~80% 辺りにおいては、エントロピーによる誤検出の抑制によって若干の精度改善が見受けられる。一方、NTCIR-9 フォーマルランセットにおいては、未知語の検索語のみにエントロピーの足切りを施したが、全体的に Recall が低下してしまう結果となった。

STD の検出コストが 0.5 以上になると Recall は 90% を超える一方で、多くの誤検出が発生する。図 6.8 に誤検出の分布を掲載する。STD 検出コスト 0.5 以上では、多くの誤検出においてエントロピーが低下している。実験結果より、Recall 率が限りなく

<sup>1</sup>最大 1、最小 0 で、低い方が検出されやすい



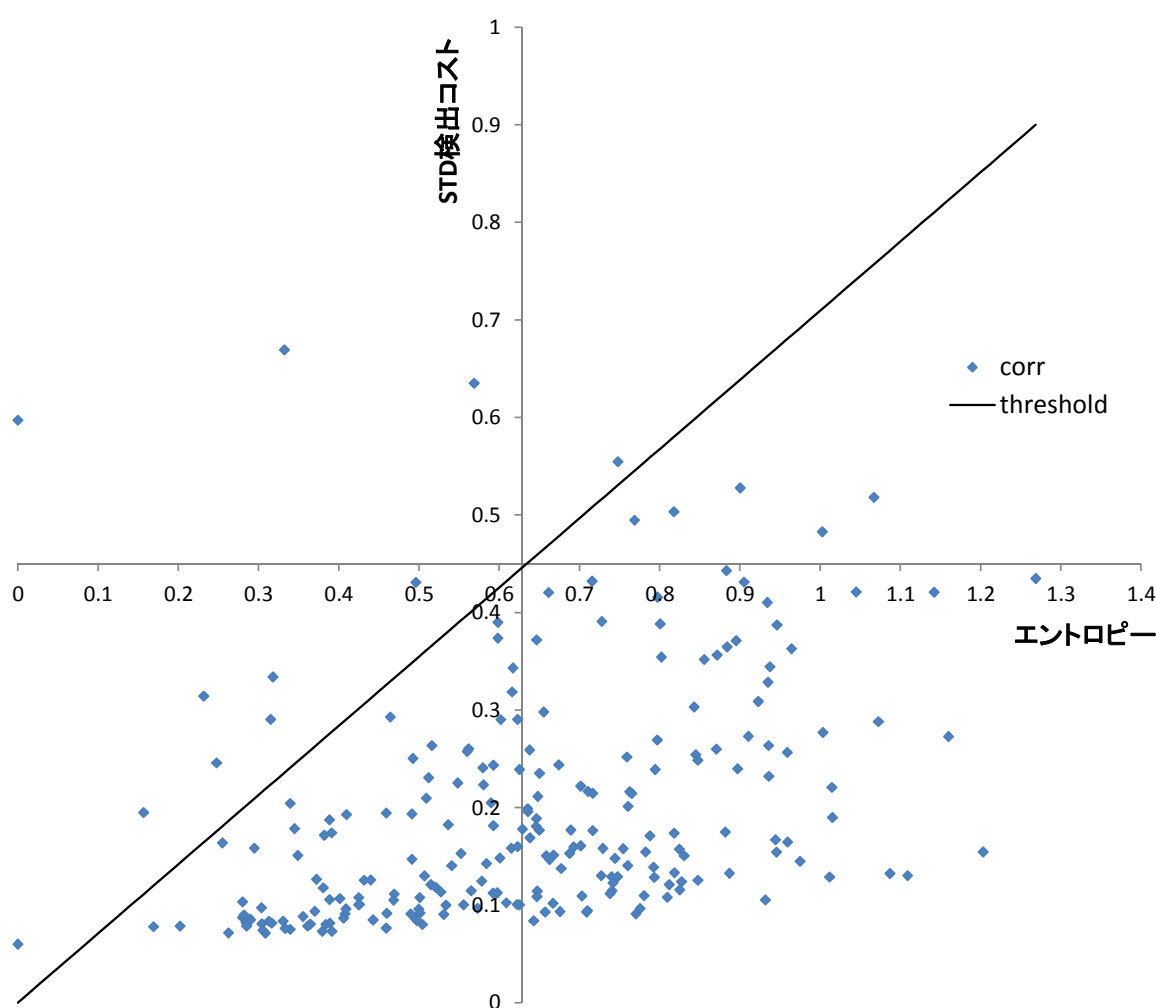


図 6.7: STD の検出コストとエントロピーの関係図

100%に近いところでの誤検出を半分以下に抑えることができることを確認できたが、最大の F-measure が得られるような STD 検出コスト (コストが 0.1~0.2 あたり) では、低いエントロピーの誤検出は非常に少なかった。

以上より、未知語検出における低エントロピーの検出は信用できないという仮説によって、一定のエントロピーの閾値で足切りする効果が有る傾向が見られた。一方、検索語のセットによっては足切りの悪影響が出現することが判明した。

#### 6.2.4 最良の STD 性能時のエントロピー

前節の評価実験結果から示されたように、単純なエントロピーの利用では、STD の性能改善に効果が薄い。そこで、F-measure が最大となる STD コストの閾値における、PTN のエントロピーを調査した。表 6.6 に調査結果を示す。なお、表 6.6 では、検索語が検出された区間のエントロピーとそれが含まれる発話全体のエントロピーを掲載し

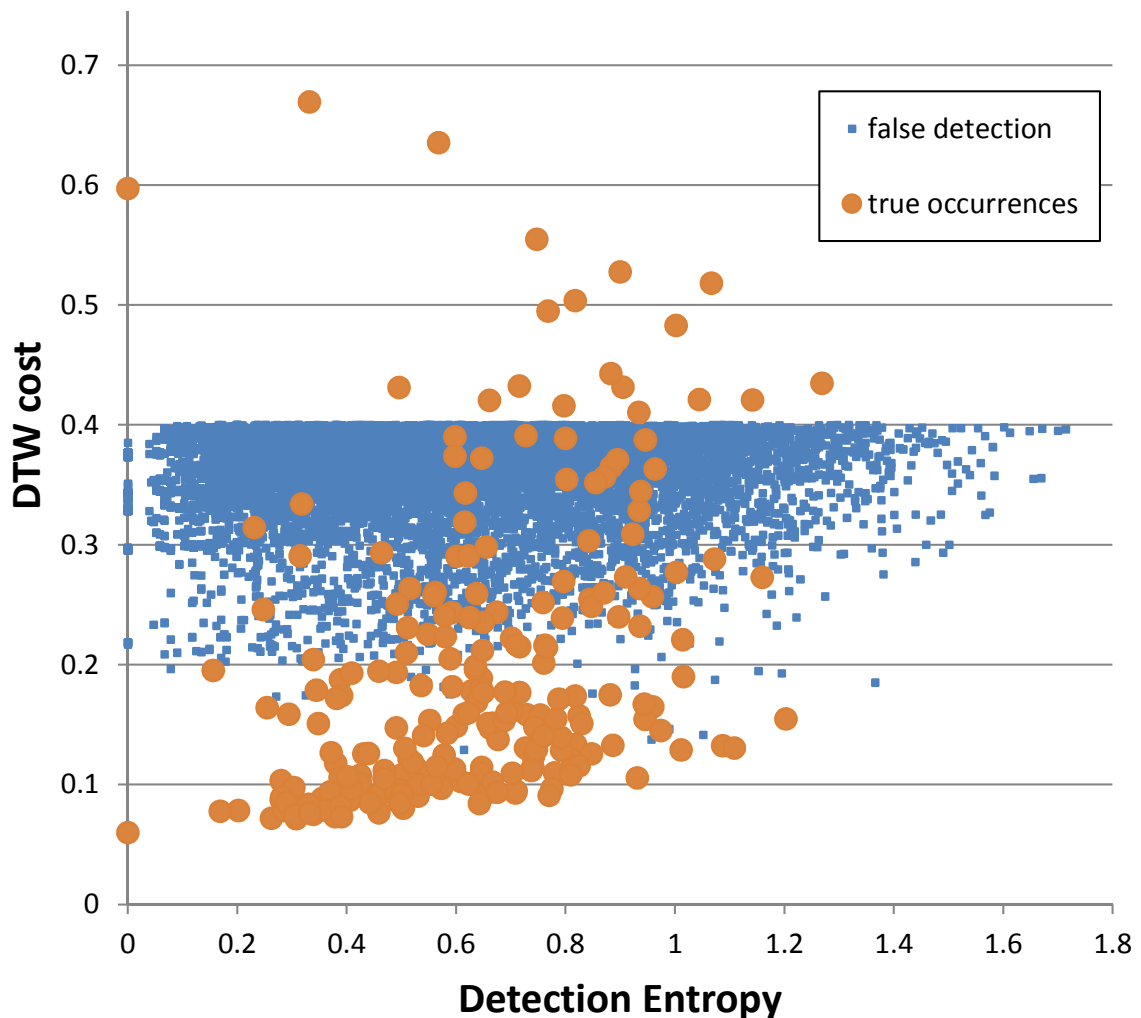


図 6.8: 誤検出を含む STD の検出コストとエントロピーの関係図

ている。

表 6.6 の各テストセットの検出結果を比較すると、正解検出、誤検出、不検出の順にエントロピーが高い。正解検出のエントロピーが低いことから、検索語が含まれる発話において、音声認識システム間の揺れが小さい場合は、検出がしやすい(すなわち厳しい閾値でも検出可能)と導き出せる。

一方で、不検出だった検索語は正解検出された検索語よりも(特に発話の)エントロピーが高いことから、音声認識システム間の揺れが大きい PTN の場合では、そこに含まれている検索語の検出は難しいことが分かる(検出閾値を固定した場合)。この場合、閾値を緩めることで不検出發話を検出することが可能となるが、誤検出が増加する問題がある。

表 6.6: 最大の検出性能 (F-measure) 時の PTN のエントロピー

テストセット Recall, Precision, F-measure	検出結果	発話数	平均 PE (検出区間)	平均 PE (発話全体)
CORE R:64.4%, P:84.7% F:73.2%	正解検出	150	0.459	0.573
	誤検出	27	0.508	0.715
	不検出	83	0.525	0.729
NTCIR_OOV R:53.8%, P:86.8% F:66.3%	正解検出	105	0.438	0.512
	誤検出	16	0.517	0.586
	不検出	90	0.528	0.701
NTCIR_IV R:70.7%, P:91.5% F:79.8%	正解検出	118	0.398	0.394
	誤検出	11	0.434	0.429
	不検出	49	0.485	0.684

以上より、未知語・既知語に限らず誤検出を低く抑えようとする、エントロピーが低い発話に含まれる検索語しか検出することができない。例えば、NTCIR\_IV の場合で Recall が約 70% であり、残りの約 30% の検索語は音声認識が難しい発話に含まれている。

STD の性能を改善するための一つのアプローチとして、エントロピーが高い発話に対して何らかの音声認識上の対策を講じることが考えられる。例えば、未知語の場合は、検出の閾値を緩くすることによって Recall を改善することは可能である。そこで発生する多くの誤検出については、誤検出の方が不検出よりもエントロピーが高い傾向にあることから、検出された区間のエントロピーが低い検索語に対する何らかのフィルタリングが効果的であると推察される。しかし、前節でも述べたように、単純な足切りでは効果が薄いため、単純に足切りを行うのではなく、より厳密な音響マッチング等を施すことによって、検出／リジェクトの判定を行うことが検討できる。

また、誤検出よりも不検出の方がエントロピーが高いことから、誤検出された発話には、検索語と類似している音素列が複数の音声認識システムで認識されている場合があると考えられる。この場合、単純な音素系列のマッチングである DP では誤検出かどうかの判断が難しく、検索語が含まれていると判断してしまうため、何らかの対策が必要となる。また、エントロピーとは関係ないが、既知語の誤検出 11 個については、11 個中 8 個の誤検出が「東京都」というクエリに対する誤検出であった。これは「東京と○」(○には地名が入る)という発話が誤って検出されており、「東京都」と「東京と」の誤りである。こういった同音異義語の誤りについては、コンテキスト情報等を使うことによって解決することが可能である。

## 6.3 iSTD タスクにおける PTN の性能

これまで、音声中の検索語検出というタスクは検索語の検出性能を重点に評価が行われてきた。これは STD 技術の評価する上で有用な評価手段であった。

一方で、実環境で STD 技術を利用するにあたって、ある特定の検索語がどの音声アーカイブ内にも存在していないことを発見する技術の要求もある。NTCIR-10 SpokenDoc-2 STD サブタスクでは、“inexistent Spoken Term Detection”(iSTD) タスクが新たに設定された [37]。

この iSTD タスクに対して、本研究で提案した手法が有効であるかを評価する。

### 6.3.1 iSTD タスク

iSTD タスクは、ある与えられた検索語が音声アーカイブ内に存在する／しないを検査し、その結果を返すタスクである。iSTD タスクは、既存の STD タスクと異なり、クエリセットをまとめて一つの評価をすること、音声ドキュメント集合全体に対する検索語の出現／非出現のみを評価することが特徴である。

NTCIR-10 SpokenDoc-2 STD サブタスクにおける iSTD の評価音声は、音声ドキュメント処理ワークショップ (Spoken Document Processing Workshop : SDPWS) の学会講演音声 (全 104 講演) が対象となる。

検索語は、SDPWS のいずれかの講演内で 1 回以上発話されている検索語の集合 (集合 ( $\in$ )) と、1 度も発話されていない検索語の集合 (集合 ( $\notin$ )) から構成される。この 1 度も発話されていない検索語の集合をどこまで検出されなかったかが評価されるタスクである。NTCIR-10 SpokenDoc-2 STD サブタスクにおける iSTD の検索語は、検索語 100 件 + ダミー検索語 100 件の計 200 件である。

iSTD タスクの評価は、以下によって行われる。

- Recall-Precision カーブ
- Recall-Precision カーブにおける最大の F-measure
- detection=“no” 判定に限定した Recall と Precision

### 6.3.2 評価実験

PTN の iSTD タスクにおける検索語の検出方法は STD タスクに用いた PTN からの検出方法とほぼ同一であるが、2 ステップの検出手法となる。iSTD スコアは検出された候補に対して、STD エンジンによって計算された最も低いスコアとみなしたものである。これは、STD タスクにおいてはスコアが高いもの＝すなわち検出できた、スコアが低いもの＝すなわち検出し難いものとしているためである。

iSTD における STD エンジンの初回ステップは、DP ベースの計算に基づいて iSTD スコアを出力する。第 2 ステップでは、初回ステップで算出された iSTD スコアと検出

表 6.7: PTN を用いた iSTD タスク性能

	Rank 100* <sup>1</sup>			Maximum* <sup>2</sup>			
	Rec. [%]	Prec. [%]	F. [%]	Rec. [%]	Prec. [%]	F. [%]	Rank
Base Line	75.00	75.00	75.00	90.00	68.18	77.59	132
エントロピーなし	79.00	79.00	79.00	84.00	78.50	81.16	107
エントロピーあり	82.00	82.00	82.00	85.00	80.19	82.52	106

候補のエントロピー値を組み合わせで算出し、最終的な iSTD スコアを出力する。初回ステップの iSTD スコアは、次式により算出される。

$$\text{iSTD score (at first pass)} = 1 - \text{"DTW cost"} \quad (6.11)$$

第 2 ステップで適用するエントロピーは、前節で行ったエントロピーの分析結果より決定した。STD プロセスの第 2 ステップへの適用は、線形関数  $y = ax + b$  ( $x$  はエントロピー、 $y$  は DP コスト) を用いて検出を分離することを試みる。

パラメータ  $a$  及び  $b$  は、STD の性能の最大化に寄与するように設定した。この iSTD タスクでは、 $a$  及び  $b$  はそれぞれ 0.014, 0 を設定した。

NTCIR-10 SpokenDoc-2 STD サブタスクに対して、エントロピーを用いない iSTD エンジンとエントロピーを適用した評価実験を行った。

実験結果を表 6.7 に示す。表中の Rank 100\*<sup>1</sup> は上位 100 件の値で計算した性能を表し、Maximum\*<sup>2</sup> は上位  $N$  件の値で計算した性能を表す ( $N$  は Recall Precision カーブにおいて F-measure が最大となる件数を指す)。

実験結果より、検出候補に対してエントロピーを適用することで検索語リストの上位 100 位の値で計算された F-measure において 3.0% の改善を達成した。また、本手法は 2013 年 6 月に開催された NTCIR-10 SpokenDoc-2 iSTD サブタスクにおいて、最も優れた検索性能を示した [37]。

## 6.4 まとめ

本章では、誤検出を抑制するパラメータのより効果的な利用方法について述べた。まず、検索語の音素長による誤検出の傾向を調査した結果について検討を行い、音素長に応じて誤検出抑制パラメータの適用を変えることで検索性能が改善することが示された。

評価実験より、音素長が 10 未満の検索語に対して編集距離のコストを高くすることで、検索性能が改善することが示された。さらに、NULL 遷移のコストを Voting によって決定することで、NULL 遷移の信頼度に応じたコストを与えることが可能であり、検索性能が改善することが示された。このとき、音素長が 10 未満の検索語に対する NULL 遷移のコストを高く設定することで、F-measure が最大となる閾値を高くす

ることが可能となり、音素長が 10 以上の検索語における F-measure が最大となる閾値に近づけることで、検索性能が改善することが示された。

次に、複数の音声認識システムの出力を利用したネットワーク型インデックスの複雑さに着目した、検索語の検出方法について述べた。

評価実験より、エントロピーを用いることで、若干の検索性能の改善が見られた。しかし、単純なエントロピーの利用では、STD の性能改善に効果が薄いことが示された。エントロピーの調査結果から、未知語・既知語に限らず誤検出を低く抑えようとすると、エントロピーが低い発話に含まれる検索語しか検出することができないことが示された。STD の性能を改善するための一つのアプローチとして、エントロピーが高い発話に対して何らかの音声認識上の対策を講じることが必要であることが示された。

また、音声中の検索語検出のタスクの一つである iSTD タスクに対して本提案手法が有効であるか評価を行い、その結果について述べた。iSTD タスクに対して本手法を用いることで、高い検出性能が得られることが示された。また、エントロピーを適用することで、iSTD タスクにおいては検索性能が向上することが示された。

## 第7章 音声中の検索誤検出の応用

本章では，第6章までで提案した，複数の音声認識システムの出力を利用したネットワーク型インデックスによる音声中の検索語の検出方法の応用について述べる．

第4章と第5章では，複数の音声認識システムの出力を利用したネットワーク型インデックスと，ネットワーク型インデックスを構築する際に得られる情報を，誤検出を抑制するパラメータとして利用する方法について述べた．第6章では，誤検出を抑制するパラメータのより効果的な利用方法や，複数の音声認識システムの出力を利用したネットワーク型インデックスの複雑さに着目した検索語の検出方法について述べた．また，音声中の検索語検出タスクの一つである iSTD タスクにおいて提案する手法の効果を検証した．

本章では，提案した音声中の検索語検出手法を応用することが可能であるかを考察する．提案した音声中の検索語検出手法を，大語彙連続認識システムで用いる言語モデルの学習データ選別や認識単語の選別に用いることで，音声認識性能を向上させることが可能かを考察した．また，電子ノート作成支援システム [38] に提案した STD 手法を利用し，その効果を考察した．

### 7.1 音声認識の語彙推定への利用

提案した音声中の検索語検出手法を，大語彙連続認識システムで用いる言語モデルの学習データ選別や，認識単語の選別に用いることで，音声認識性能を向上させることが可能かを考察する．

#### 7.1.1 音声認識の語彙推定

汎用の音声認識システムでは，講義・講演を音声認識する際に高い認識率 (単語正解率，単語正解精度，名詞正解率が得られていない．これは，講義・講演では話題が限られており，特定の単語，特定の言い回しが多いことが理由に挙げられる．

講義認識用の言語モデルにそれ以前の講義音声の書き起こしデータを用いる方法がある [71]．これは，講義の連続性 (多くの大学で1科目あたり15コマの授業が連続的に開講されている) を利用し，以前の講義音声の書き起こしを用いて適応化を行っている．しかし，講義の書き起こしの作成は非常に高コストであり，現実問題として書き起こしを用意することは難しい．そこで，講義で使われたテキストや Switchboard コーパス，授業で用いる教科書や講義で使用したパワーポイント等の電子スライド情報を利用す

る方法が提案されている [72][73]. しかし, これらの手法はスライドを利用している講義音声を認識する場合のみに利用できる. 現在でも講師の多くは黒板を用いた講義を実施しており, この場合は当然スライド情報を用いることができない. そこで, 小暮らは [74], 大学では学生向け (電子) シラバスが用意されていることに着目し, これを利用することで言語モデル適応化を行うためのドキュメントを収集する方法を提案している. この方法では, 講義の前に言語モデルを適応化することができるので, リアルタイムで講義音声を比較的精度よく認識することが可能となる. リアルタイムで認識する際は, 話題に適応化された言語モデルを利用し, かつ言語モデルがコンパクトである方が認識処理速度も高速になる. 一方で, 授業シラバスのような事前情報が利用できない状況と考えた場合, Web を利用することが有効である. 梶原らは [75], Web ドキュメントを用いた講演音声認識のための反復適応化手法を提案している. これらのように, 様々な適応化手法が提案されているが, これらは適応化によりモデルサイズが増加する.

コンパクトな言語モデルを構築するための手法は, 踊堂ら [76] や Stolck [77] が提案している. これらの手法は, エントロピー等の指標により N-gram パラメータ数の削減を図っている. また, A. Sethy ら [78] は集めてきた大量の WEB データから, 音声認識対象のデータに類似した学習テキストを相対エントロピー基準で選択する方法を提案している.

### 7.1.2 STD を利用した語彙推定

ここでは, これらの手法と異なり, 発話毎に認識辞書に登録する語彙を STD により推定することを検討し評価を行う. すなわち, 発話毎に語彙集合を形成することで, より話題に特化した辞書を作成する.

PTN を用いた語彙推定処理を図 7.1 に示す. 提案手法では, 認識対象音声を 2 度認識するため, リアルタイム用途ではない. アーカイブ作成等で応用できる技術であると考えている. PTN による語彙推定では, まず CSJ 講演集合から言語モデルを 5 種類作成する. 作成した言語モデルと音響モデル 2 種類を用いて, 講義音声の 10 種類の音声認識結果を得る. 10 種類の認識結果から PTN を作成し, vocabulary に登録されている単語をクエリとして単語検索を行う. STD を行うことにより, どの単語がどの発話に含まれているのかが分かる. この結果を用いて各発話に対する認識辞書を構築する. そして, 発話毎の辞書を用いて再度音声認識することで認識率の改善を図る.

### 7.1.3 評価実験

認識対象の音声には, 山梨大学工学部コンピュータ・メディア工学科コンピュータサイエンスコースで開講された 3 講義と CSJ の評価データ用テストセットから 3 講演 (講演 ID:A01M0007, A01M0035, A01M0074) の音声を用いた.



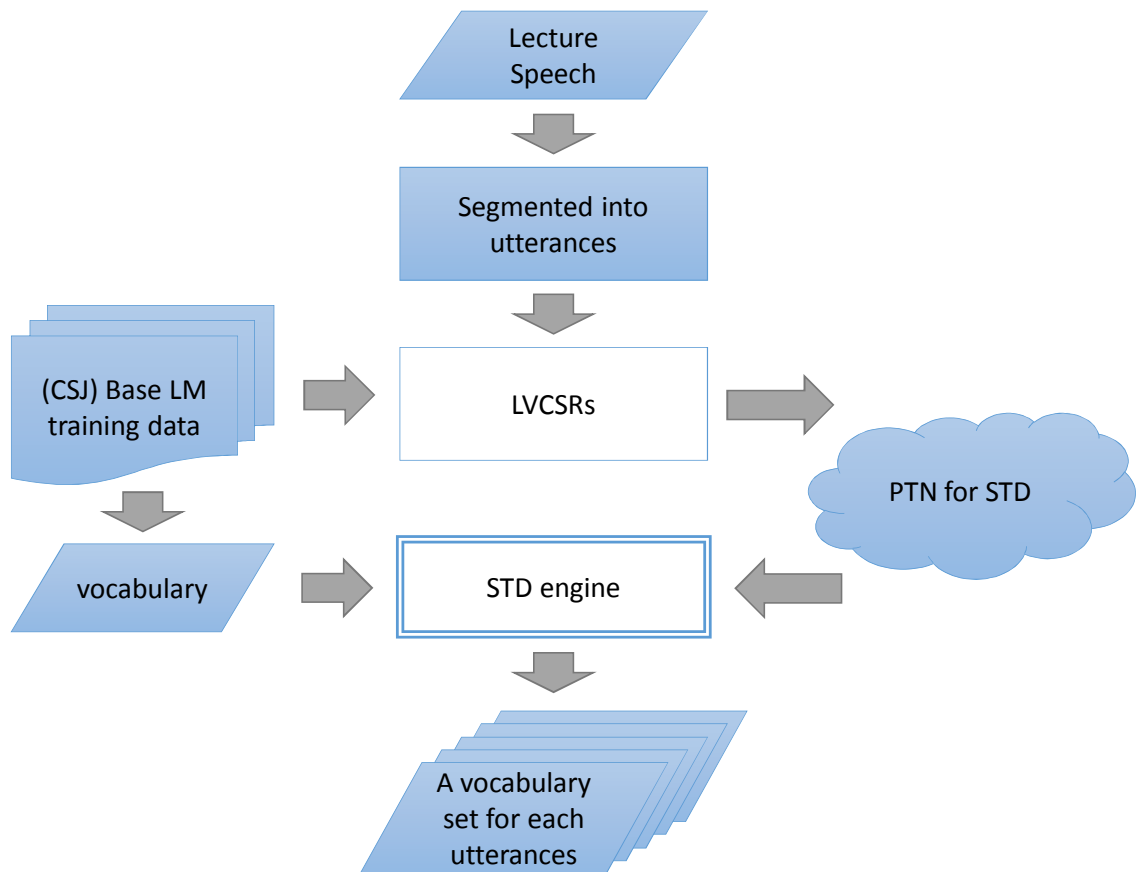


図 7.1: PTN による STD を利用した語彙推定の流れ

ベースとなる言語モデルは CSJ に含まれる 3,286 講演 (評価データを除く学会講演・模擬講演・読み上げ・対話をすべて含む, 約 123M バイト) から学習した語彙数 20,000 の単語 trigram である。ベースラインの認識辞書には, 言語モデル学習時に利用した語彙数 20,000 のものを利用している。

音響モデルの学習に用いるパラメータは, 16kHz, 16bit でサンプリングされた音声より求められた, 12次元のメル周波数ケプストラム (MFCC), その1次差分 ( $\Delta$ MFCC) と2次差分 ( $\Delta\Delta$ MFCC), パワーの1次差分 ( $\Delta$ LogPow) と2次差分 ( $\Delta\Delta$ LogPow) の38次元を使用している。音響モデルには CSJ に収録されているコアを除く学会・模擬講演全 2525 講演の男女混合話者から学習した総状態数約 3,000 の 64 混合 triphone を用いた。

表 7.1 に語彙推定に STD を用いることにより作成した認識辞書を用いて音声認識した結果と講義・講演毎の未知語率と語彙数を示す。“Base” は, 語彙推定を行っていない結果, “STD” は STD による語彙推定を利用したものである。“STD” の vocabulary size は発話毎の辞書の語彙サイズの平均である。STD を用いた語彙推定処理により, 認識辞書の語彙を大幅に削減することが示された。それに伴い, すべての講演・講義で音声認識率が改善していることが示された。しかし, 改善幅はわずかである。原因と

表 7.1: 語彙推定による音声認識率の比較結果

	Lecture1		Lecture2		Lecture3	
	Base	STD	Base	STD	Base	STD
Corr.[%]	59.67	60.43	41.92	43.23	45.66	46.21
Acc.[%]	54.76	55.86	31.18	33.56	33.13	34.87
N Corr.[%]	47.64	48.47	32.37	33.29	34.55	34.88
OOV Rate[%]	7.82	16.33	5.97	25.50	7.44	22.29
vocabulary size	20000	2155	20000	1056	20000	1050
	A01M0007		A01M0035		A01M0074	
	Base	STD	Base	STD	Base	STD
Corr.[%]	82.37	82.39	70.42	70.90	83.06	83.87
Acc.[%]	78.87	79.16	66.99	67.54	79.51	80.59
N Corr.[%]	85.81	85.81	67.30	68.12	83.07	85.11
OOV Rate[%]	8.10	13.70	9.92	15.67	6.15	9.16
vocabulary size	20000	540	20000	2095	20000	866

して、必要な語彙が削られたことによる未知語率の悪化、言語モデルを再学習をしていないこと(学習データの選択を含む)、が挙げられる。特に、STD 技術が完全ではないことから、未知語率が大幅に悪化してしまったことが大きい。STD は短い単語検出に弱く(湧き出し誤検出が大量に発生してしまう)、これが語彙推定精度を大きく下げている。しかし、実験結果から STD を用いた語彙推定処理が有効であることが実証された。

## 7.2 音声電子ノート作成支援システムへの応用

STD の研究の多くは検出性能向上を目的とするものであり、実環境下での有効性の評価を目的とするものは少ない。

STD を応用した既存のシステムには、音声・動画検索ソフトウェア [79][80] や報道番組の書き起こしシステムのキーワード検索機能 [81] 等がある。これらの利用シーンとして、映画やドラマなどの動画からの特定のシーン抽出、コールセンターでの録音音声からの発話抽出等が想定されている。そのため、これらの分野での利用は可能であると考えられる。これらの他にも講義音声の聞き直しや議事録の検索など様々な分野での利用に期待が持てるが、実際に有効であるかの評価はされていない。

また、講義音声の聞き直しを対象とした STD 技術の応用の先行研究として、放送大学の講義音声を検索対象としたキーワード検索 [34] がある。この研究によって、放送大学の講義音声は STD 用の評価テストコレクション [82] のデータと比較して良好な検索性能が得られることが示されている。このことから、講義音声を対象に STD が有効に利用できる場合には、ノート見直し作業の速度が向上し、学生の学習効率向上が期

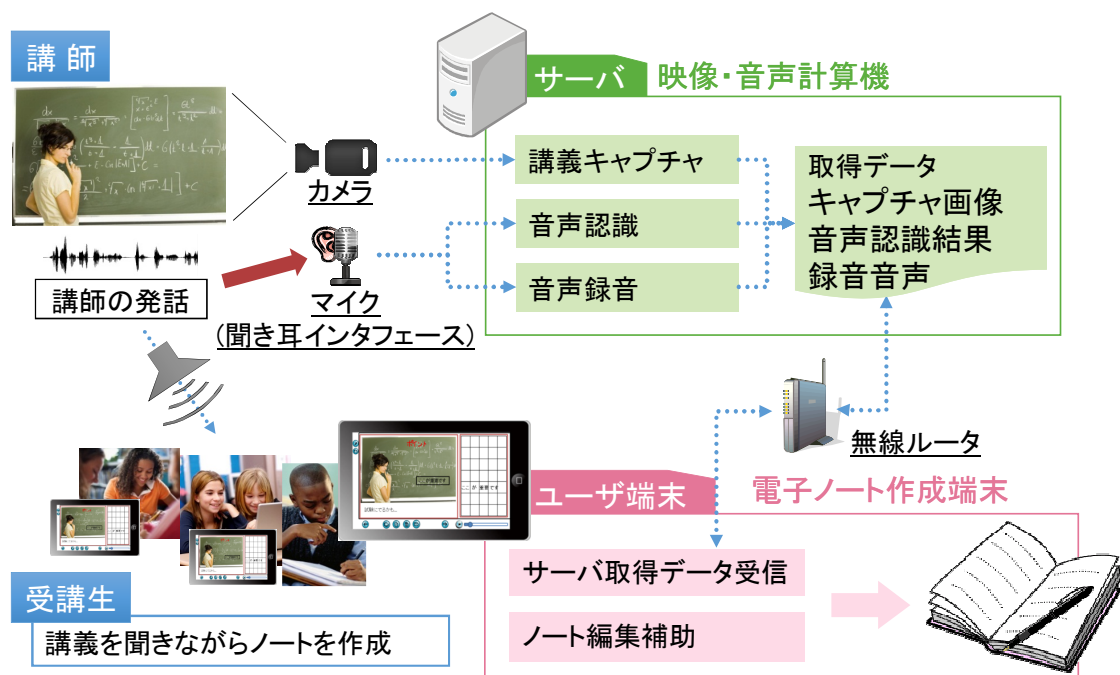


図 7.2: 電子ノート作成支援システムの構成と利用概要

待できると思われる．そこで，STD 使用者と不使用者の電子ノート見直し作業にかかる時間を比較する被験者実験を行うことで，STD の有効性評価を行う．

### 7.2.1 電子ノート作成支援システム

現在，大学の講義において受講生の理解が追いつかないという問題がある．この原因の一つとして，講義スライドの展開速度が速いことが挙げられる．講義の展開が速い場合，スライドや板書の書き逃しや講師の話の聞き逃しが多く起こる．その結果，受講生は講義内容の理解が難しくなる．このような問題を解決するために，マルチメディア情報を利用した電子ノート作成支援システムを開発中である [38]．このシステムでは，スライドや黒板のキャプチャ静止画や音声認識した文字列，キーボード・手書きによる書き込み等の機能を利用してノートを作成することができる．また，講義終了後には録音音声の再生を利用することができるため，聞き逃しにも対応可能である．この録音音声に対して，STD を利用し話し手の話した言葉を精度よく検索できるようになれば，より高速なノートの見直しが期待される．

講義において電子ノート作成支援システムを利用した場合のシステムの構成を 7.2 に示す．このシステムは，映像・音声計算機（以下，サーバという）と，電子ノート作成・閲覧端末（以下，ユーザ端末という）の 2 つから構成される．

サーバは，カメラにより撮影されたスライド投影や黒板の映像を静止画として保存

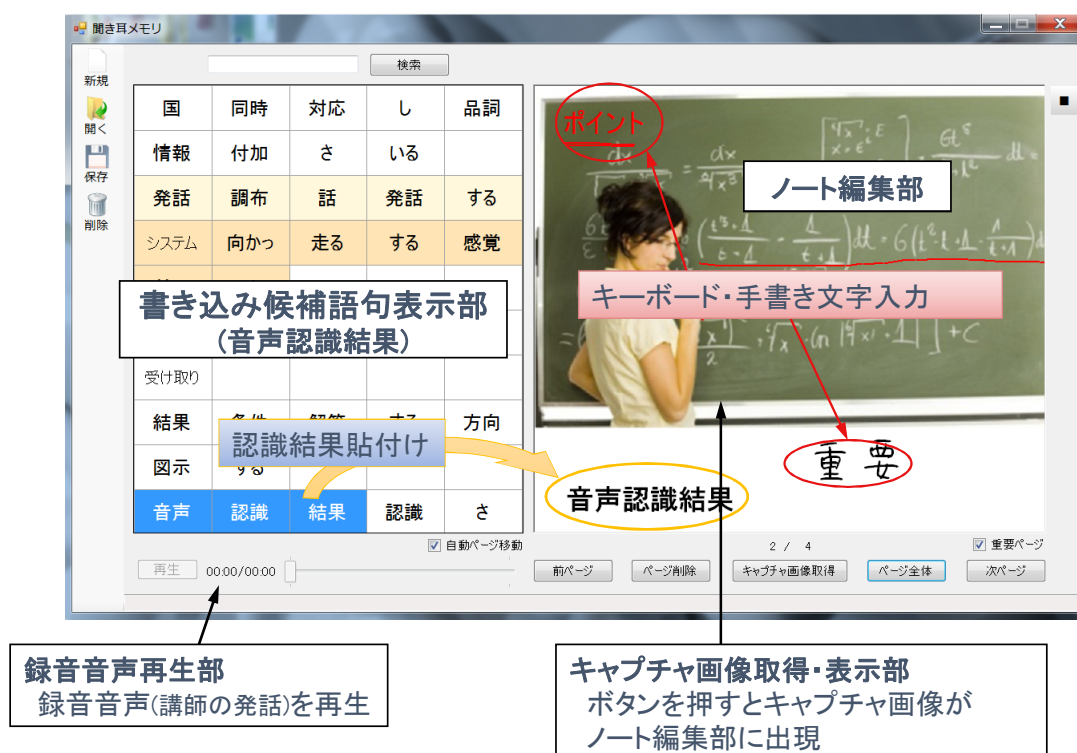


図 7.3: 電子ノート作成支援システムのユーザ端末画面イメージと使用例

する．同時に，講師が装着しているマイクから入力された音声を認識・録音（以下，録音音声という）する．静止画は受講生がユーザ端末を通し要求することでユーザ端末に送信され，ノート編集部に配置される．音声認識結果は，随時ユーザ端末に送信される．この支援システムでは，音声認識を全面的には利用せず，補助的に利用している．認識誤りが発生し書き込みたい語句がユーザ端末の画面上に表示されなかった場合でも，手書き入力で対応できる．これにより，音声の誤認識の影響により，システムに決定的な不具合が生じることが回避されている．以上の構成で，講師は普段通り講義を行い，受講生は講師の話を聞きながら，受信できる静止画と音声認識結果を利用しノートを作成することが可能となる．

電子ノート作成支援システムのユーザ端末の使用例と画面イメージを7.3に示す．ユーザ端末は，主として以下の機能を有する．

1. 講義音声の録音
2. 講義音声の認識結果とキーボード・手書きによる書き込み
3. 黒板やスライド等のキャプチャ静止画の貼り付け
4. 録音音声の再生

## 5. 認識語句の検索

## 6. 検索語の該当する発話の検索・頭出し再生 (STD の利用)

これらの機能のうち、電子ノート作成時には、第1項目および第2項目、第3項目が利用される。それ以外の機能は、作成した電子ノートを見直しの際に利用される。ユーザ端末は、ノート編集部と音声認識による書き込み候補語句表示部の2画面から構成される。ノート編集部には、サーバから取得するスライド等のキャプチャ静止画が貼り付けられる。その画像上、もしくは空白部分に、手書きで文字や図形を自由に書き込むことができる。キーボード入力による書き込みにも対応している。ノート編集部に表示されている情報がノートの1ページとなり、ページを加えていきノートを作成する。書き込み候補語句表示部には、サーバに保存される音声認識結果の単語列が表示される。ユーザは単語を選択することで、選択した単語をノート編集部に配置することができる。講義の終了後、録音音声サーバからユーザ端末に送信され、ユーザ端末で再生できるようになる。作成したノートを開覧する際、ノートに配置した語句を選択することで、その語句が講師により発話された時点から録音音声を頭出し再生できる。録音音声の再生位置は、シークバー操作での調整も可能である。また、STDを利用することで、録音された講師の発話からユーザが指定した文字列が発声された箇所を検索・頭出し再生も可能である。

### 7.2.2 電子ノート作成支援システムへのSTDの適用

電子ノート作成支援システムへのSTDの適用は、ノート見直し作業を対象としている。ユーザ端末は、音声からのキーワード検出機能(STD)を持つ。任意のキーワードを入力し検索を行うと、ユーザ端末の録音音声再生部に検索結果が表示される。STDによる検索結果の表示例を7.4に示す。検索結果は(1)リスト形式と(2)シークバーの対応位置の2種類の表示が可能である。リスト形式の場合は、発話箇所が該当する時間、検索語句、ANDやOR等のマルチワード検索のオプションの種類を表示する。シークバーの対応位置の場合は、発話箇所をシークバーに対応した位置に丸で表示する。検索結果は、リスト形式とシークバー対応位置の両方を表示、もしくはシークバー対応位置のみの表示が可能である。

### 7.2.3 被験者実験

STDの有効性を評価するために、STDの使用者と不使用者の電子ノート見直し作業にかかる時間の比較実験を行った。

被験者実験では、被験者10名によるノート見直し作業にかかる時間を測定した。被験者は大学生・大学院生の10名である。作業内容は、一か月前に被験者が作成した講義内容の電子ノートを参照しながら試験問題に解答するというものである。この解答

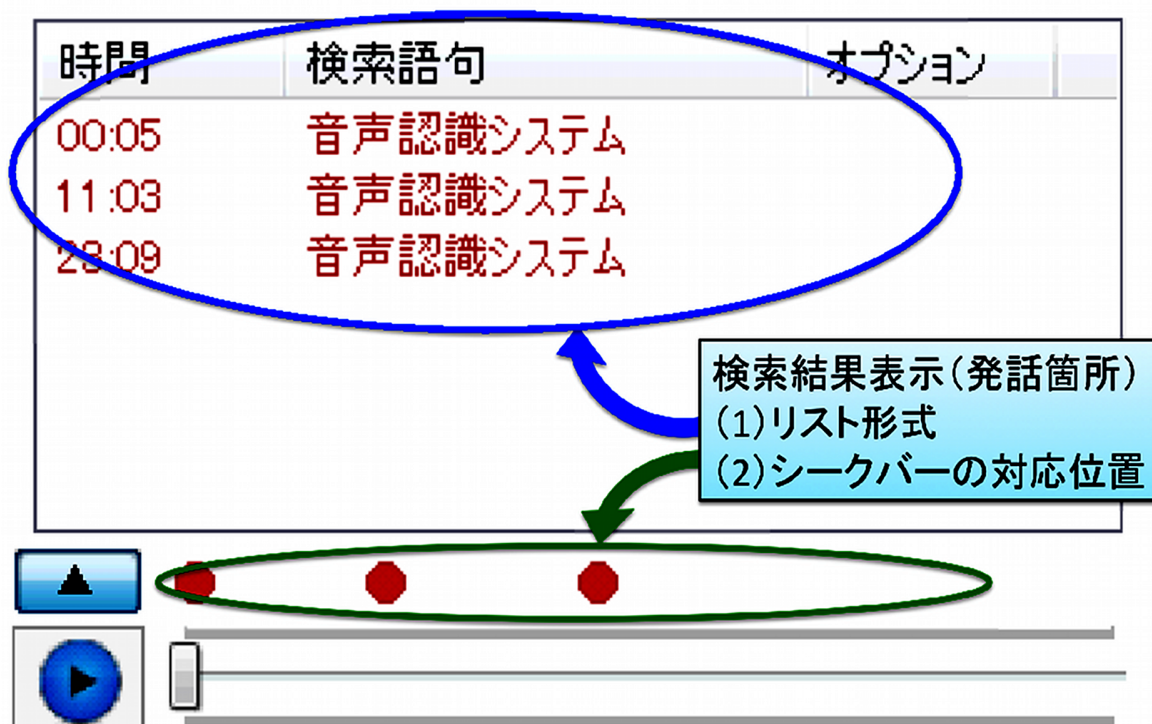


図 7.4: STD による検索結果の表示例

表 7.2: 実験で使用した STD の性能

STD 性能	F-measure= 約 0.17 Recall= 約 61%, Precision= 約 10%
検索語延べ数	108 語
検索語種類数	49 語
平均検索語数	22 語／人
検索時間	約 10 秒／検索語 (内, 約 7 秒はインデックス構築時間)
平均音素数	9 音素 (最短 4, 最長 20)

がすべて正答となるまでの時間を計測した。なお、ノート参照の際、全被験者の内半数には STD を使用する。

ノート見直し作業時の STD 性能を表 7.2 に示す。この講義音声の単語認識率は約 26%であったが、STD の Recall は約 61 しかし、Precision は約 10 これは、検索語として“実”や“ダル”等の音素数が短い検索語の湧き出し誤りが原因である。これら 2 語を除いた際の STD 性能は、Recall が約 67%, Precision が約 20%, F-measure が約 30%であった。

STD を使用した 5 名と不使用の被験者 5 名の正答時間の平均値と標準偏差を表 7.3 に示す。STD の使用者 5 名と不使用者 5 名の設問ごとの平均正答時間を表 7.4 に示す。

表 7.3: STD 使用者と不使用者の正答時間の平均値と標準偏差 [分' 秒"]

	不使用者	使用者
平均値	40'58"	35'25"
標準偏差	14'34"	6'17"

表 7.4: STD 使用者と不使用者の設問ごとの正答時間の平均値 [分' 秒"]

設問	不使用者	使用者	差
1	2'27"	2'46"	0'19"
2	6'33"	1'28"	<b>-5'05"</b>
3	5'11"	6'18"	1'07"
4	2'43"	3'49"	1'06"
5	14'33"	11'55"	<b>-2'38"</b>
6	3'02"	4'03"	1'01"
7	3'55"	2'05"	<b>-1'50"</b>
8	2'34"	3'00"	0'26"

表 7.3 の平均値から、STD の使用者は不使用者に比べ、5 分程速く解答できている (ただし、危険率 5% で有意差なし)。標準偏差では、STD の使用者は不使用者に比べ正答時間の個人差が小さいことが確認できる。

STD 使用者には、電子ノート見直し作業の後に STD に関するアンケートに回答して頂いた。5 段階評価で STD の必要性が 4.2 と高い評価を得られた。

また、全ての被験者には、電子ノート見直し作業の後に自由記述のアンケートに回答して頂いた。STD 不使用者のアンケートの結果では

- 認識ができていない単語が多かった
- 認識ができていない場合、広い範囲の音声聞くことになり、解答に必要な箇所を見つけ出すのに苦労した

との回答があった。

一方、STD 使用者のアンケート結果では

- 認識結果にない単語も、STD を使用することで検索ができたため役立った
- 誤りがある場合でも聞くべき範囲を視覚的に特定できたため、解答に必要な箇所を見つけ出せた (シークバーの位置に対応した検索結果表示)
- 検索速度が遅いと感じた

との回答があった。

以上より，電子ノート見直し作業において，STD は有効である可能性があるということが示された。

しかし，アンケート結果に“検索速度が遅いと感じた”という回答があったことから検索速度の向上が必要であることが示された。

## 7.3 まとめ

本章では，提案した音声中の検索語検出手法をシステムソリューションに用いることが可能であるかを考察した．まず，提案した音声中の検索語検出手法を，大語彙連続認識システムで用いる言語モデルの学習データ選別や，認識単語の選別に用いることで，音声認識性能を向上させることが可能かを考察した．STD を用いた語彙推定処理により，認識辞書の語彙を大幅に削減することが可能となり，それに伴い，すべての講演・講義で音声認識率が改善することが示された。

また，電子ノート作成支援システム [38] に提案した STD 手法を利用し，その効果を考察した．結果として，電子ノート見直し作業において，STD は有効である可能性があるということが示された．しかし，検索速度の向上が必要であることが課題として明らかになった．この検索速度が遅いことについては，NTCIR-9[47]，NTCIR-10[37] においても明らかである．これは，提案手法が検索性能に重点を置いているためである．



## 第8章 結論

本論文では、複数の音声認識システムの出力を利用することによって、STD 性能を改善させる手法について述べた。また、本研究で提案した未知検索語に頑健な STD 手法を用いたシステムソリューションについて考察した。

第3章では、複数の音声認識システムの出力を用いることで、音節単位での音声認識性能が改善されることを複数の音声認識システムによる音声認識実験の結果から示した。この結果から、単一の音声認識システムの出力より、複数の音声認識システムの出力を組み合わせた方が、特定のキーワードを見つけられる可能性が高くなることを示した。

第4章では、複数の音声認識システムの出力をどのような形態のインデックスとして利用することが、STD 性能の改善につながるかについて述べた。単一の音声認識システムの出力を利用した場合では、仮説数が多くなるほど検索性能が向上し、ネットワーク型のインデックスを構築することで Recall が 30 から 40%の間では検索性能が良くなることが示された。また、10PHOs(1-Best) の結果に示されるように、複数の音声認識システムの出力を利用することで高い検索性能が示され、PTN(1-Best) においては Recall が 60%以上で最も良い検索性能となった。以上から、複数の音声認識システムの出力を CN 化することの有用性が示された。しかし、多くの仮説を用いてネットワーク型のインデックスを構築しても、大量の湧き出し誤りが検出されてしまい、検索性能が低下した。この原因としては、ネットワークの Node や Arc が多くなり過ぎてしまい、DP を用いた単純な検索方法では多くの情報を生かしてきていないということが考えられる。また、STD に用いる音声認識システムの N-Best 出力や音声認識システムの出力を変更することによって検索性能が改善されることが示された。すなわち、適切な音声認識システムの N-Best 出力や音声認識システムの出力を選別することによって、STD 性能が改善される可能性が示されたこととなる。しかし、この検索語検出のタスクが変更された場合に、最適な N-Best 出力や音声認識システムの種類が変わる可能性がある。

これらの問題を解決し、音声中の検索語検出性能を改善させるために、第5章では誤検出を抑制するための誤検出抑制パラメータと検索エンジンの改善について第5章で述べた。誤検出を抑制するパラメータとして、3種類の情報を利用し、導入の方法によって計5種類の誤検出抑制パラメータを検討した。これらの誤検出抑制パラメータを、DP の距離計算式に導入することによって、誤検出が抑制されることが実験結果より示された。特に、音素を認識した音声認識システムの数である Voting を導入することによって、大幅に検索性能が改善された。他のパラメータにおいても、編集距離の

みを用いた DP の距離計算式を用いた場合より誤検出が抑制され、検索性能が改善された。しかし、CM スコアに関しては、導入する方法を再検討した結果、MRP において僅かに改善された程度であった。

第6章では、さらなる音声中の検索語検出性能改善のため、検索語の音素長に応じた誤検出抑制パラメータの適用法について述べた。音素長が10未満の検索語に対して編集距離のコストを高くすることで、検索性能が改善することが示された。さらに、NULL 遷移のコストを Voting によって決定することで、NULL 遷移の信頼度に応じたコストを与えることが可能であり、検索性能が改善することが示された。このとき、音素長が10未満の検索語に対する NULL 遷移のコストを高く設定することで、F-measure が最大となる閾値を高くすることが可能となり、音素長が10以上の検索語における F-measure が最大となる閾値に近づけることで、検索性能が改善することが示された。

また、他の誤検出抑制法としてネットワーク型インデックスの複雑さに着目したエントロピーを検討した。しかし、単純なエントロピーの利用では、STD の性能改善に効果が薄いことが示された。エントロピーの調査結果から、未知語・既知語に限らず誤検出を低く抑えようとする、エントロピーが低い発話に含まれる検索語しか検出することができないことが示された。STD の性能を改善するための一つのアプローチとして、エントロピーが高い発話に対して何らかの音声認識上の対策を講じることが必要であることが示された。

また、音声中の検索語検出のタスクの一つである iSTD タスクに対して本提案手法が有効であるか評価を行い、その結果について述べた。本手法を iSTD タスクに用いることで、高い検出性能が得られることが示された。さらに、エントロピーを適用することで、iSTD タスクにおいては検索性能が向上することが示された。

STD ならびに iSTD において、複数の音声認識システムの出力を利用することと、それらの出力をネットワーク型のインデックスとして利用することが有効であることが示された。また、複数の音声認識システムの出力から得られる情報を利用することが、誤検出を抑制した検索語の検出に有効であることが示された。以上より、本研究の目標である未知検索語に対して頑健な STD 手法を提案することは達成された。

未知検索語に対して頑健な STD 手法に対する今後の課題とその解決案として、以下が挙げられる。

1 点目として、複数の音声認識システムの出力の厳密なアライメントを検討する必要がある。濱中らの実験結果 [70] より、複数の音声認識システムの厳密なアライメントとエントロピーを用いることによって音声認識性能が向上することが示されている。本手法の複数の音声認識システムのアライメントは、ROVER[20] の手法と同様のベースとなる音素列と他の音素列を1つずつペアワイズアライメントしていくことでアライメントを行っている。このアライメント手法はアライメントの制度自体には注力しておらず、アライメントの順序によって結果が異なるという問題がある。この解決策として、アライメントをプログレッシブ法を用いることが挙げられる。

また、アライメントに厳密な音響マッチングを導入することも挙げられる。音声認識結果には、認識した音素または音節の発声フレームが出力される。このフレーム情報に基づいてアライメントを行うことで、アライメント精度の改善が図れる可能性が

ある。

2点目として、ネットワーク型インデックスを構築する音声認識システムの組み合わせを検討する必要がある。機械学習などを使って最適な認識システムの組合せを選ぶことで、検索性能の改善が図れる可能性がある。

3点目として、さらなる誤検出抑制パラメータの検討と検索語とインデックスの距離計算方法の検討が挙げられる。本論文ではエントロピーの指標を用いることで検索性能の改善を図ったが、その効果は僅かであった。また、本論文では、編集距離ベースの検索語とインデックスの距離計算に基づいて、検索語の検出を行った。実験結果から、編集距離を用いることで高い検索性能が得られることが示されたが、エントロピーの指標に基づく距離計算や、CM スコアをベースとした距離計算を行うことで、本論文で示した検索結果とは異なる結果が得られる可能性がある。このエントロピーベースの距離計算や、CM スコアベースの距離計算に、Voting などの誤検出抑制パラメータを導入することによって、検索性能が改善される可能性がある。

また、これらの編集距離ベースの検出結果と、エントロピーベースやCM スコアベースの検出結果を統合することによって、検索性能が改善される可能性がある。

第7章では、提案した音声中の検索語検出手法をシステムソリューションなどに用いることが可能であるかを考察した。まず、提案した音声中の検索語検出手法を、大語彙連続認識システムで用いる言語モデルの学習データ選別や、認識単語の選別に用いることで、音声認識性能を向上させることが可能かを考察した。結果として、STD を用いた語彙推定処理により、認識辞書の語彙を大幅に削減することが可能となり、それに伴い、すべての講演・講義で音声認識率が改善することが示された。

また、電子ノート作成支援システムに提案した STD 手法を利用し、その効果を考察した。結果として、電子ノート見直し作業において、STD は有効である可能性があるということが示された。

以上より、本研究で提案した未知検索語に頑健な STD 手法はシステムソリューションへの応用が可能であることが示された。しかし、検索速度の向上が必要であることが課題として明らかになった。また、本提案手法では複数の音声認識システムを利用している。このため、リアルタイムな処理を行う場合には、多くの計算リソースを必要とする。

検索速度の改善については、マルチスレッド／マルチプロセス化や GPGPU を用いた並列処理によって改善することが可能と考えられる。また、計算機上での PTN の表現方法を見直すことによってインデックスの構築、並びに検索語の検出速度の向上が図れると考えられる。

今後の展望として、本研究で提案した未知検索語に頑健な STD 手法をリアルタイム性が必要となるシステムソリューションへの適用課題は多々存在する。しかし、大量の音声ドキュメントから検索語を検出するタスクにおいては有用である。例えば、コールセンターなどで録音された大量の音声データから、オペレータが顧客に対して発してはならない NG ワードを発話していないか、また、顧客満足度の高いオペレータと低いオペレータではどのような発話の違いがあるのかなどを分析するツールとして有用であると考えられる。

# 謝辞

本研究を遂行し学位論文をまとめるにあたり，終始暖かい激励とご指導，ご鞭撻を頂いた，指導教官である関口芳廣教授ならびに西崎博光助教に心より感謝申し上げます．関口教授，西崎助教には筆者の山梨大学工学部コンピュータ・メディア工学科及び専攻在学中より，音声情報処理に関してご指導頂きました．研究を進めるための環境を整備頂き，幾度と音声情報処理研究の道に導いて頂いたことに心より感謝申し上げます．

本論文をまとめるにあたり，有益な御助言を賜りました山梨大学工学部コンピュータ理工学科 福本文代教授，山梨大学工学部情報メカトロニクス工学科 宗久知男教授，同 鈴木良弥教授，同 小谷信司教授，同 丹沢勉准教授に心より感謝申し上げます．

社会人学生として，研究と仕事の両立を支援して頂いた，東京エレクトロンTS株式会社 佐野聡氏，小島伸二氏，中矢哲氏，アライメントチームの皆様に心より感謝申し上げます．

博士課程在学中，共同研究者として，古屋裕斗氏，中込大生氏，米倉千冬氏，鈴木和将氏，澤田直輝氏に多大なご協力を頂きました．厚く御礼申し上げます．また，共に切磋琢磨し研究に挑んだ関口・西崎研究室の方々に感謝します．諸氏との交友により，充実した日々を過ごすことができました．ここに記して謝意を表します．

最後になりますが，これまで私を暖かく応援してくれた両親へ心から感謝します．そして，社会人学生として博士課程への入学を快く承諾し，どのような状況においても応援してくれました素晴らしい婚約者 智恵美に心から感謝します．

## 参考文献

- [1] Petr Motlicek, Fabio Valente, Philip N. Garner, “English Spoken Term Detection in Multilingual Recordings,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 206–209, 2010.
- [2] Chun-an Chan, Lin-shan Lee, “Unsupervised Spoken-Term Detection with Spoken Queries Using Segment-based Dynamic Time Warping,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 693–696, 2010.
- [3] Dong Wang, Simon King, Nicholas Evans, Raphael Troncy, “CRF-based Stochastic Pronunciation Modeling for Out-of Vocabulary Spoken Term Detection,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 1668–1669, 2010.
- [4] 栗城吾央, 伊藤慶明, 小嶋和徳, 石亀昌明, 田中和世, 李時旭, “未知語音声クエリにおける音声中の検索語検出—Web を利用した拡張辞書とサブワードの認識結果の統合—”, 日本音響学会 2009 年春季講演発表会講演論文集, pp. 197–200, 2009.
- [5] 岩見圭祐, 藤井康寿, 山本一公, 中川聖一, “距離つきトライグラムアレイによる未知語音声の超高速検索”, 日本音響学会 2009 年春季講演発表会講演論文集, pp. 203–206, 2009.
- [6] 澤田心太, 桂田浩一, 新田恒雄, 入部百合絵, 手島茂樹, “大規模音声ドキュメントからの高速キーワード検索法の提案とその評価”, 日本音響学会 2009 年春季講演発表会講演論文集, pp. 69–70, 2009.
- [7] 松永徹, 趙國, 山下洋一, “音声ドキュメント検索語検索における音響情報を用いた再評価”, 日本音響学会 2009 年春季講演発表会講演論文集, pp. 71–72, 2009.
- [8] X. Liu, M. J. F. Gales and P. C. Woodland, “Language Model Cross Adaptation For LVCSR System Combination,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 342–345, 2010.

- [9] Icksang Han, Chiyoun Park, Jeongmi Cho and Jeongsu Kim, “A Hybrid Approach to Robust Word Lattice Generation Via Acoustic-Based Word Detection,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 210–213, 2010.
- [10] Hung-yi Lee, Chia-ping Chen, Ching-feng Yeh, Lin-shan Lee, “Improved Spoken Term Detection by Discriminative Training of Acoustic Models based on User Relevance Feedback,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 1273–1276, 2010.
- [11] Aren Jansen, Kenneth Church, Hynek Hermansky, “Towards Spoken Term Discovery At Scale With Zero Resources,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 1676–1679, 2010.
- [12] Sha Meng, Wei-Qiang Zhang, Jia Liu, “Combining Chinese Spoken Term Detection Systems via Side-information Conditioned Linear Logistic Regression,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 685–688, 2010.
- [13] Carolina Parada, Abhinav Sethy, Mark Dredze, Frederick Jelinek, “A Spoken Term Detection Framework for Recovering Out-of-Vocabulary Words Using the Web,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 1269–1272, 2010.
- [14] Javier Tejedor, Doroteo T. Toledano, Miguel Bautista, Simon King, Dong Wang and José Colás, “Augmented set of features for confidence estimation in spoken term detection,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 701–704, 2010.
- [15] Taisuke Kaneko, Tomoyosi Akiba, “Metric Subspace Indexing for Fast Spoken Term Detection,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 689–692, 2010.
- [16] Daniel Schneider, Timo Mertens, Martha Larson, Joachim Köhler, “Contextual Verification for Open Vocabulary Spoken Term Detection,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 697–700, 2010.
- [17] Mirko Hannemann, Stefan Kombrink, Martin Karafiát, Lukáš Burget, “Similarity Scoring for Recognizing Repeated Out-of-Vocabulary Words,” in *Proceedings of the*

- 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 897–900, 2010.
- [18] Sebastian Tschöpel, Daniel Schneider, “A lightweight keyword and tag-cloud retrieval algorithm for automatic speech recognition transcripts,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 1277–1280, 2010.
  - [19] Yoshiaki Itoh, Hiromitsu Nishizaki, Xinhui Hu, Hiroaki Nanjo, Tomoyosi Akiba, Tatsuya Kawahara, Seiichi Nakagawa, Tomoko Matsui, Yoichi Yamashita and Kiyooki Aikawa, “Constructing Japanese Test Collections for Spoken Term Detection,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 677–680, 2010.
  - [20] J. G. Fiscus, “A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER),” in *Proc. of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU’97)*, pp. 347–354, 1997.
  - [21] T. Utsuro, Y. Kodama, T. Watanabe, H. Nishizaki, and S. Nakagawa, “An empirical study on multiple lvsr model combination by machine learning,” in *Proc. of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, pp. 13–16, 2004.
  - [22] K. Iwata, K. Shinoda, and S. Furui, “Robust spoken term detection using combination of phone-based and word-based recognition,” in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2008*, pp. 2195–2198, 2008.
  - [23] Jonathan Mamou, Yosi Mass, Bhuvana Ramabhadran and Benjamin Sznaider, “Combination of Multiple Speech Transcription Methods for Vocabulary Independent Search,” in *Proc. of the 2nd workshop on Searching Spontaneous Conversational Speech (SSCS) 2008*, pp. 20–27, 2008.
  - [24] Roy Wallace, Brendan Baker, Robbie Vogt and Sridha Sridharan, “The Effect of Language Models on Phonetic Decoding for Spoken Term Detection,” in *Proc. of the 3rd workshop on Searching Spontaneous Conversational Speech (SSCS) 2009*, pp. 31–36, 2009.
  - [25] 小野寺悠二, 伊藤慶明, 小嶋和徳, 石亀昌明, 田中和世, 李時旭, “複数のサブワード・言語モデルを用いた音声中の検索語検出の高精度化”, 第4回音声ドキュメント処理ワークショップ講演論文集, 2010.

- [26] Lidia Mangu, Eric Brill and Andreas Stolcke, “Finding consensus in speech recognition: word error minimization and other applications of confusion networks,” *Computer Speech and Language* 14(4), pp. 373–400, October 2000.
- [27] Yi-cheng Pan, Hung-lin Chang, Berlin Chen and Lin-chan Lee, “Subword-based Position Specific Posterior Lattices(S-PSPL) for Indexing Speech Information,” in *Proc. of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2007*, pp. 318–321, 2007.
- [28] 堀 貴明, リー ハセリントン, ティモシー ヘイゼン, ジェームズ グラス, “コンフュージョンネットワークを用いたオープン語彙発話検索法とその評価”, 電子情報通信学会技術研究報告 SP11-8, pp. 43–48, 2007.
- [29] S. Meng, J. Shao, R. P. Yu, J. Liu, and F. Seide, “Addressing the out-of-vocabulary problem for large-scale chinese spoken term detection,” in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2008*, pp. 2146–2149, 2008.
- [30] Jie Gao, Qingwei Zhao, Yonghong Yan and Jian Shao, “EFFICIENT SYSTEM COMBINATION FOR SYLLABLE-CONFUSION-NETWORK-BASED CHINESE SPOKEN TERM DETECTION,” in *Proceedings International Symposium on Chinese Spoken Language Processing (ISCSLP) 2008*, pp. 366–369, 2008.
- [31] 伊藤慶明, 岩田耕平, 石亀昌明, 田中和世, 李時旭, “語彙制限のない音声文書検索における複数サブワードの統合—検索語彙に依存した検索性能推定指標の導入”, 情報処理学会論文誌, Vol50, No.2, pp.524–533, 2009.
- [32] 神田直之, 住吉貴志,, 小窪浩明, 佐川浩彦, 大淵康成, “多段リスクアリングに基づく大規模音声中の任意検索語検出”, 電子情報通信学会論文誌 D, Vol50, No.2, pp.524–533, 2009. 電子情報通信学会論文誌 D, Vol.J95-D No.4, pp. 969–981, 2012.4.
- [33] 岩見圭祐, 山本一公, 中川聖一, “複数音声認識システムを併用した音節 n-gram 索引による検索性能の改善”, 第 6 回音声ドキュメント処理ワークショップ講演論文集, SDPWS2012-05, 2012.
- [34] 勝浦広大, 桂田浩一, 入部百合絵, 森本容介, 辻靖彦, 青木久美子, 新田恒雄, “放送大学の講義音声を対象とした高速キーワード検索の性能評価” 第 6 回音声ドキュメント処理ワークショップ講演論文集, SDPWS2012-05, 2012.
- [35] 斉藤裕之, 伊藤慶明, 小嶋和徳, 石亀昌明, 田中和世, 李時旭, “複数音節の事前検索結果を利用した音声中の検索語検出の高速化”, 第 6 回音声ドキュメント処理ワークショップ講演論文集, SDPWS2012-05, 2012.



- [36] 金子泰輔, 秋葉友良, “部分距離空間上の索引を用いた STD における距離順計算の厳密化と非直線検出への拡張”, 第 6 回音声ドキュメント処理ワークショップ講演論文集, SDPWS2012-05, 2012.
- [37] Tomoyosi Akiba, Hiromitsu Nishizaki, Kiyooki Aikawa, Xinhui Hu, Yoshiaki Itoh, Tatsuya Kawahara, Seiichi Nakagawa, Hiroaki Nanjo and Yoichi Yamashita, “Overview of the NTCIR-10 SpokenDoc-2 Task,” *Proceedings of the 10th NTCIR Conference*, pp. 573–587, 2013.6.
- [38] 太田晃平, 西崎博光, 関口芳廣, “マルチメディア情報を利用した電子ノート作成支援システム,” 情報処理学会第 75 回全国大会講演論文集, Vol.4, 4ZE-4, pp. 737–738, 2013.3.
- [39] 西崎博光, 胡新輝, 南條浩輝, 伊藤慶明, 秋葉友良, 河原達也, 中川聖一, 松井知子, 山下洋一, 相川清明, “Spoken Term Detection のためのテストコレクション構築とベースライン評価”, 情報処理学会研究報告 SLP-81-13, NL-196-13, 2010.
- [40] 北研二, 津田和彦, 獅子堀正幹, “情報検索アルゴリズム”, 共立出版, 1, 2002.
- [41] John S. Garofolo, Cedric G. P. Auzanne, Ellen M. Voorhees, “The TREC Spoken Document Retrieval Track: A Success Story,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTER-SPEECH) 2010*, pp. 210–213, 2010.
- [42] Tomoyosi Akiba, Kiyooki Aikawa, Yoshiaki Itoh, Tatsuya Kawahara, Hiroaki Nanjo, Hiromitsu Nishizaki, Norihito Yasuda, Yoichi Yamashita, and Katunobu Itou, “Construction of a Test Collection for Spoken Document Retrieval from Lecture Audio Data,” 情報処理学会論文誌, Vol.50. No.2, pp. 501–513, 2009.
- [43] Chang Woo Han, Shin Jae Kang, Chul Min Lee, and Nam Soo Kim, “Phone Mismatch Penalty Matrices for Two-Stage Keyword Spotting Via Multi-Pass Phone Recognizer,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 202–205, 2010.
- [44] Carolina Parada, Abhinav Sethy, Bhuvana Ramabhadran, “Query-by-Example Spoken Term Detection For OOV Terms,” in *Proc. of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2009)*, pp. 404–409, 2009.
- [45] NIST. (2006) The spoken term detection (STD) 2006 evaluation plan. [Online]. Available: <http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf>

- [46] K. Maekawa, “Corpus of Spontaneous Japanese: Its design and evaluation,” In *Proc. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, pp. 7–12, 2003.
- [47] Tomoyosi Akiba, Hiromitsu Nishizaki, Kiyoaki Aikawa, Tatsuya Kawahara and Tomoko Matsui, “Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop”, *Proceedings of NTCIR-9 Workshop Meeting*, pp. 223–235, 2011.12.
- [48] Keisuke Iwami and Seiichi Nakagawa, “High speed spoken term detection by combination of n-gram array of a syllable lattice and LVCSR result for NTCIR-SpokenDoc”, *Proceedings of NTCIR-9 Workshop Meeting*, pp. 242–248, December 6-9, 2011.12.
- [49] Taisuke Kaneko, Tomoko Takigami and Tomoyosi Akiba, “STD based on Hough Transform and SDR using STD results: Experiments at NTCIR-9 SpokenDoc”, *Proceedings of NTCIR-9 Workshop Meeting*, pp. 264–270, December 6-9, 2011.12.
- [50] Kouichi Katsurada, Koudai Katsuura, Yurie Iribe and Tsuneo Nitta, “Utilization of Suffix Array for Quick STD and Its Evaluation on the NTCIR-9 SpokenDoc Task”, *Proceedings of NTCIR-9 Workshop Meeting*, pp. 271–274, December 6-9, 2011.12.
- [51] Hiroyuki Saito, Takuya Nakano, Shirou Narumi, Toshiaki Chiba, Kazuma Kon’no and Yoshiaki Itoh, “An STD system for OOV query terms using various subword units”, *Proceedings of NTCIR-9 Workshop Meeting*, pp. 281–286, December 6-9, 2011.12.
- [52] Yoichi Yamashita, Toru Matsunaga and Kook Cho, “YLABRU at Spoken Term Detection Task in NTCIR-9”, *Proceedings of NTCIR-9 Workshop Meeting*, pp. 287–290, 2011.12.
- [53] Kouichi Katsurada, Koudai Katsuura, Kheang Seng, Yurie Iribe and Tsuneo Nitta, “Using Multiple Speech Recognition Results to Enhance STD with Suffix Array on the NTCIR-10 SpokenDoc-2 Task”, *Proceedings of the 10th NTCIR Conference*, pp. 588–591, 2013.6.
- [54] Kazuma Kon’no, Hiroyuki Saito, Shirou Narumi, Kenta Sugawara, Kesuke Kamata, Manabu Kon’no, Jinki Takahashi and Yoshiaki Itoh, “An STD System for OOV Query Terms Integrating Multiple STD Results of Various Subword units”, *Proceedings of the 10th NTCIR Conference*, pp. 592–596, 2013.6.
- [55] Satoru Tsuge, Ken Ichikawa, Norihide Kitaoka, Kazuya Takeda and Kenji Kita, “Spoken Content Retrieval Using Distance Combination and Spoken Term De-

- tection Using Hash Function for NTCIR10 SpokenDoc2 Task“, *Proceedings of the 10th NTCIR Conference*, pp. 597–603, 2013.6.
- [56] Tomoyosi Akiba, Tomoko Takigami, Teppei Ohno and Kenta Kase, “DTW-Distance-Ordered Spoken Term Detection and STD-based Spoken Content Retrieval: Experiments at NTCIR-10 SpokenDoc-2”, *Proceedings of the 10th NTCIR Conference*, pp. 618–625, 2013.6.
- [57] Iori Sakamoto, Kook Cho, Masanori Morise and Yoichi Yamashita, “YLABRU at Spoken Term Detection Task in NTCIR-10 SpokenDoc-2”, *Proceedings of the 10th NTCIR Conference*, pp. 638–642, 2013.6.
- [58] Nagisa Sakamoto and Seiichi Nakagawa, “Spoken Term Detection by N-gram Index with Exact Distance for NTCIR-SpokenDoc2”, *Proceedings of the 10th NTCIR Conference*, pp. 643–647, 2013.6.
- [59] Naoki Yamamoto and Atsuhiko Kai, “Spoken Term Detection Using Distance-Vector based Dissimilarity Measures and Its Evaluation on the NTCIR-10 SpokenDoc-2 Task”, *Proceedings of the 10th NTCIR Conference*, pp. 648–653, 2013.6.
- [60] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄, “IT Text 音声認識システム”, オーム出版, p. 198, 2001.
- [61] A. Lee and T. Kawahara, “Recent Development of Open-Source Speech Recognition Engine Julius,” in *Proc. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 2009*, 2009.
- [62] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸, “日本語形態素システム茶釜 使用説明書”, 奈良先端科学技術大学院大学松本研究室, 2000.
- [63] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto “Applying Conditional Random Fields to Japanese Morphological Analysis”, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pp.230-237, 2004.
- [64] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland, “*The HTK Book*,” Morgan Kaufmann, 1995.
- [65] Philip Clarkson and Ronald Rosenfeld, “Statical Language Modeling using the CMU-Cambridge Toolkit,” in *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*,

- [66] 国立国語研究所. The Corpus of Spontaneous Japanese [Online]. Available: [http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/public/index\\_j.html](http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/public/index_j.html)
- [67] 情報検索システム評価用テストコレクション構築プロジェクト. 第9回NTCIR ワークショップ, 情報アクセス技術の評価: 情報検索, 質問応答, 言語横断情報アクセス [Online]. Available: <http://research.nii.ac.jp/ntcir/ntcir-9/tasks.html>
- [68] B. Varadarajan, D. Yu, L. Deng, and A. Acero, “Maximizing global entropy reduction for active learning in speech recognition,” *Proc. ICASSP*, pp. 4721–4724, (2009).
- [69] Yoshiaki Itoh, Kohei Iwata, Masaaki Ishigame, Kazuyo Tanaka, Shi-wook Lee, “Spoken Term Detection Results Using Plural Subword Models by Estimating Detection Performance for Each Query” in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2012*, pp. 2117–2120, 2011.
- [70] 濱中悠三, 江森 正, 越仲孝文, 篠田浩一, 古井貞熙, “音声認識のための複数の認識器を利用した能動学習,” 情報処理学会研究報告 SLP-79-4, 2009.
- [71] 小暮悟, 西崎博光, 土屋雅稔, 中川聖一, “講義コンテンツの収集・分析および講義音声の認識手法に関する検討,” 第1回音声ドキュメント処理ワークショップ, 豊橋技術科学大学メディア科学リサーチセンター, pp. 1–8, 2007.
- [72] 根本雄介, 秋田祐哉, 河原達也, “講義音声認識のためのスライド情報を用いた言語モデル適応,” 第1回音声ドキュメント処理ワークショップ, 豊橋技術科学大学メディア科学リサーチセンター, pp. 89–94, 2007.
- [73] Alex Park, Timothy J. Hazen, and James R. Glass, “Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling,” *In Proc. of the ICASSP2005*, Vol.1, pp. 497–500, 2005.
- [74] 小暮悟, 西崎博光, 土屋雅稔, 富樫慎吾, 山本一公, 中川聖一, “日本語講義音声コンテンツコーパスの構築と講義音声認識手法の検討,” 第2回音声ドキュメント処理ワークショップ, 豊橋技術科学大学メディア科学リサーチセンター, pp. 7–14, 2008.
- [75] 梶浦泰智, 鈴木基之, 伊藤彰則, 牧野正三, “WWW を用いた言語モデルの教師なし反復適応法,” 第1回音声ドキュメント処理ワークショップ, 豊橋技術科学大学メディア科学リサーチセンター, pp. 109–114, 2007.
- [76] 踊堂憲道, 伊藤克亘, 鹿野清宏, 中村哲, “N-gram モデルのエントロピーに基づくパラメータ削減に関する検討,” 情報処理学会, 情報処理学会論文誌, Vol. 42, No. 2, pp. 327–333, 2001.

- [77] A.Stolcke, “Entropy-based pruning of backoff language models,” *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, pp. 270–274, 1998.
- [78] A.Sethy and P.G.Georgiou, “An iterative relative entropy minimization-based data selection approach for N-gram model adaptation,” *Trans. on AUDIO, SPEECH, AND LANGUAGE PROCESSING*, pp. 13–23, Vol. 17, No. 1, pp. 13–23, 2009.
- [79] 株式会社ボイザー, 製品紹介, “<http://www.voiser.co.jp/products.html>” (参照日 : 2013.5.17).
- [80] 株式会社アニモ, AnimoSearch, “<http://www.animo.co.jp/record/as/>” (参照日 : 2013.5.17).
- [81] 小林彰夫, 奥貴裕, 本間真一, 佐藤庄衛, 今井亨, “コンテンツ活用のための報道番組自動書き起こしシステム”, 電子情報通信学会論文誌, Vol. J93-D, No.10, pp. 2085–2095, 2010.
- [82] 伊藤慶明, 西崎博光, 中川聖一, 秋葉友良, 河原達也, 胡新輝, 南條浩輝, 松井知子, 山下洋一, 相川清明, “音声中の検索語検出のためのテストコレクション構築 -中間報告-”, 情報処理学会研究報告, Vol.2009-SLP-78, no.4, pp. 1–8, 2009.

# 学外発表

## 論文誌掲載 (査読付き)

1. Satoshi Natori, Yuto Furuya, Hiromitsu Nishizaki, Yoshihiro Sekiguchi, “Spoken Term Detection Using Phoneme Transition Network from Multiple Speech Recognizers’ Outputs,” *Journal of Information Processing*, Vol.21, No.2, pp. 176–185, 2013.4.

## 国際会議発表 (査読付き)

1. Satoshi Natori, Hiromitsu Nishizaki, and Yoshihiro Sekiguchi, “Japanese Spoken Term Detection Using Syllable Transition Network Derived from Multiple Speech Recognizers’ Outputs,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, pp. 681-684, 2010.9.
2. Satoshi Natori, Hiromitsu Nishizaki and Yoshihiro Sekiguchi, “Network-formed Index from Multiple Speech Recognizers’ Outputs on Spoken Term Detection,” in *the proceedings of the 2nd Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2010) (student symposium)*, p.1, 2010.12.
3. Yuto Furuya, Satoshi Natori, Hiromitsu Nishizaki, and Yoshihiro Sekiguchi, “Introduction of False Detection Control Parameters in Spoken Term Detection,” *the Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2012)*, Digital (4 pages), 2012.12.
4. Satoshi Natori, Yuto Furuya, Hiromitsu Nishizaki, and Yoshihiro Sekiguchi, “Entropy-based False Detection Filtering in Spoken Term Detection Tasks,” *the Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2013)*, Digital (4 pages), 2013.10-11.
5. Chifuyu Yonekura, Yuto Furuya, Satoshi Natori, Hiromitsu Nishizaki and Yoshihiro Sekiguchi, “Evaluation of the Usefulness of Spoken Term Detection in an

Electronic Note-Taking Support System,” *the Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2013)*, Digital (4 pages), 2013.10-11.

## 国際会議発表 (査読なし)

1. Hiromitsu Nishizaki, Yuto Furuya, Satoshi Natori and Yoshihiro Sekiguchi, “Spoken Term Detection Using Multiple Speech Recognizers’ Outputs at NTCIR-9 SpokenDoc STD subtask,” *Proceedings of NTCIR-9 Workshop Meeting*, pp. 236–241, December 6-9, 2011.
2. Yuto Furuya, Daiki Nakagomi, Satoshi Natori, Hiromitsu Nishizaki and Yoshihiro Sekiguchi, “STD and SCR Techniques and Their Evaluations on the NTCIR-10 SpokenDoc-2 Task,” *Proceedings of the 10th NTCIR Conference*, pp. 626-633, 2013.6.

## 口頭発表 (査読なし)

1. 小林健司, 宗宮充宏, 名取賢, 西崎博光, 関口芳廣, “講義音声の自動評価のための各種特微量の調査”, 豊橋技術科学大学メディア科学リサーチセンター&情報処理学会音声言語処理研究会, 第2回音声ドキュメント処理ワークショップ, pp.143–148, 2008.2.
2. 名取賢, 西崎博光, 関口芳廣, “任意語彙発話音声検索のための複数の認識モデルを利用した音節遷移ネットワークの構築”, 日本音響学会 2009 年秋季研究発表会講演論文集, pp.205–206, 2009.9.
3. 名取賢, 西崎博光, 関口芳廣, “複数音声認識システムを用いた音声中の検索語検出の検討”, 情報処理学会, 情報処理学会技術報告, Vol.2009-SLP-79, No.19, 6pages, 2009.12.
4. 名取賢, 西崎博光, 関口芳廣, “複数音声認識システムに基づいた音声中の検索語検出手法の検討と CSJ テストコレクションでの評価”, 豊橋技術科学大学メディア科学研究会&情報処理学会 SLP 研究会音声ドキュメント処理ワーキンググループ, 第4回音声ドキュメント処理ワークショップ講演論文集, 2010.2.
5. 名取賢, 西崎博光, 関口芳廣, “音声中の検索語検出のための複数の音声認識結果を用いたネットワーク型インデキシング”, 日本音響学会 2010 年秋季研究発表会講演論文集, pp.61–64, 2010.9.

6. 名取賢, 西崎博光, 関口芳廣, “複数音声認識システムを利用した STD における誤検出を抑制した検出方法の検討”, 日本音響学会 2011 年春季研究発表会講演論文集, 3-5-19, 2011.3.
7. 藤原裕幸, 名取賢, 西崎博光, 関口芳廣, “話し言葉音声認識のための STD を利用した語彙推定手法の検討”, 日本音響学会 2011 年春季研究発表会講演論文集, 3-5-4, 2011.3.
8. 古屋裕斗, 名取賢, 西崎博光, 関口芳廣, “音声中の検索語検出における検出誤り抑制パラメータの検討”, 情報処理学会 SLP 研究会音声ドキュメント処理ワーキンググループ・豊橋技術科学大学メディア科学研究会, 第 6 回音声ドキュメント処理ワークショップ講演論文集, 8 pages, 2012.3.
9. 古屋裕斗, 名取賢, 西崎博光, 関口芳廣, “音声中の検索語検出のための音素遷移ネットワークのエントロピー分析”, 情報処理学会 SLP 研究会音声ドキュメント処理ワーキンググループ・豊橋技術科学大学メディア科学研究会, 第 7 回音声ドキュメント処理ワークショップ講演論文集, 6 pages, 2013.3.
10. 西崎博光, 古屋裕斗, 名取賢, 関口芳廣, “条件付き確率場を用いた音声中の検索語検出の検討”, 日本音響学会 2013 年秋季研究発表会講演論文集, 2-P-26, 2013.9.
11. 古屋裕斗, 名取賢, 西崎博光, 関口芳廣, “クエリのエントロピーを利用した STD 手法の検討”, 日本音響学会 2014 年春季研究発表会講演論文集, 3-4-6, 2014.3.
12. 澤田直輝, 古屋裕斗, 名取賢, 西崎博光, 関口芳廣, “STD システムへの音素間距離の導入方法の検討”, 日本音響学会 2014 年春季研究発表会講演論文集, 3-Q5-11, 2014.3.
13. 米倉千冬, 古屋裕斗, 澤田直輝, 名取賢, 西崎博光, 関口芳廣, “音声ドキュメントからの頻出発話語句の発見”, 第 8 回音声ドキュメント処理ワークショップ講演論文集, 2014.3.



# 付 録 A      日本語 STD 用テストコレクション                  ションのコア講演用未知語テ                  ストセットの50検索語

本論文で用いた日本語 STD 用テストコレクションのコア講演用未知語テストセットの 50 検索語を表 A.1, 表 A.2 に示す.

表 A.1: コア講演用未知語テストセットの 50 クエリ (1)

モーラ	検索語	tf	df
13	石川島造船所	1	1
12	コンテキストディペンデント	5	1
11	クリントイーストウッド	2	1
10	ボスニア・ヘルツェゴビナ	1	1
	ユニバーサルスタジオ	3	2
	ホテルニューハンプシャー	2	1
9	春桜亭円紫	1	1
	談洲楼焉馬	1	1
	竹取物語	5	1
	高島平駅	2	1
	タンチョウの飛来地	2	1
	チトー大統領	2	1
8	スティーブンキング	1	1
	名犬ラッシー	2	1
	駒沢公園	8	1
	まほろば連邦	5	1
	南大泉	5	1
	伊曾保物語	2	1
	営団赤塚	1	1
	キラウエア火山	5	1

表 A.2: コア講演用未知語テストセットの 50 クエリ (2)

モーラ	検索語	tf	df
7	ユーゴスラビア	7	1
	代々木上原	2	2
	釧路湿原	3	2
	コザクラインコ	4	1
	奄美大島	1	1
	オスマントルコ	6	1
	奥穂高岳	1	1
6	光が丘	10	3
	ノーベル賞	2	1
	西日暮里	7	1
	常盤平	7	1
	拝島駅	12	1
	本駒込	3	1
	メーンランド	2	1
	バンクーバー	4	2
5	アルバニア	9	1
	三河島	3	1
	美堀町	4	1
	屈斜路湖	3	1
	スリーピー	7	1
	ワイコロア	6	1
	九品仏	6	1
	NATO 軍	3	1
4	那覇港	2	1
	ネパール	27	1
	安保理	5	1
	ヒマラヤ	4	2
	知床	14	1
	八潮市	7	1
	ケベック	7	1

# 付 録 B    NTCIR-9 SpokenDoc タスク formal-run テストセットの 50 クエリ

NTCIR-9 SpokenDoc タスク formal-run テストセットの 50 クエリを表 B.1, 表 B.2 に示す.

表 B.1: formal-run テストセットの 50 クエリ (1)

モーラ	検索語	tf	df
14	サンクリストバル・デ・ラスカサス	7	1
13	冷泉家時雨亭文庫	4	1
12	津軽海峡冬景色	2	1
10	最大エントロピー	10	3
	和英語林集成	6	1
9	言語処理学会	7	2
	形態素解析	9	3
	竹取物語	5	1
	東京競馬場	3	2
8	バービー人形	4	1
	ドムドムバーガー	4	1
	キラウエア火山	5	1
	駒澤大学	3	1
	工場実習	10	1
	メルケプストラム	11	4
	オーウェンミーニー	8	1
	知床半島	4	1
	ウィザードオブオズ	8	2

表 B.2: formal-run テストセットの 50 クエリ (2)

モーラ	検索語	tf	df
7	不信任案	4	1
	海外派兵	4	1
	釧路湿原	3	2
	中野新橋	10	1
	ネパール旅行	6	1
	ピッチパターン	19	2
	ステロイド剤	4	1
	高島平	13	1
	ユーゴスラビア	7	1
6	バンクーバー	4	2
	ボルネオ島	2	1
	N T C I R	6	1
	フラタニティ	13	1
	光が丘	11	3
	京王線	7	3
	キディーランド	6	1
	水ぼうそう	14	1
	残効量	11	2
5	ゴーカート	3	2
	八王子	6	4
	花屋さん	5	3
	ハワイ島	12	1
	カトマンズ	4	2
	コリー犬	4	1
	九品仏	6	1
	土佐日記	9	2
	東京都	23	17
4	福生市	5	1
	髪型	10	3
	目黒区	11	3
	サイパン	5	1
	山梨	5	2

# 付 録 C    NTCIR-10 SpokenDoc-2 タ スク large-size タスク large-size テストセットの 100 クエリ

NTCIR-10 SpokenDoc-2 タスク large-size タスク large-size テストセットの 100 クエリを表 C.1, 表 C.2, 表 C.3, 表 C.4 に示す.

表 C.1: large-size テストセットの 100 クエリ (1)

モーラ	検索語	tf	df
12	音声合成システム	25	13
	サポートベクターマシーン	7	2
	チャイルドトランスミッション	9	2
	横浜国立大学	7	6
11	ウェアラブルコンピューター	5	2
	スペースダイバーシティ	10	2
	第二次世界大戦	27	23
	内閣不信任案	14	3
	ベクトル空間モデル	19	7
	マルチビームフォーミング	37	1
10	宇宙戦艦ヤマト	11	1
	逆フーリエ変換	6	6
	最大エントロピー	19	5
	サザンオールスターズ	32	4
	周波数ワーピング	17	1
	セクシャルハラスメント	27	2
	フジモリ大統領	9	1
	北海道大学	14	9
	マスカルポーネチーズ	13	2

表 C.2: large-size テストセットの 100 クエリ (2)

モーラ	検索語	tf	df
9	石原裕次郎	17	4
	英会話学校	13	8
	オーサリングツール	9	2
	原子力発電	44	10
	港北ニュータウン	11	5
	D P マッチング	40	16
	阪神タイガース	7	4
	ビーフストロガノフ	11	1
	ベルサイユ宮殿	8	5
	ポートフォリオ評価	30	2
	類聚名義抄	15	1
8	暗証番号	7	4
	ウィザードオブオズ	15	7
	オーウェンミーニー	8	1
	キラウエア火山	7	3
	銀山温泉	15	2
	楕形フィルター	15	3
	甲州街道	21	12
	駒沢公園	11	2
	タロット占い	10	1
	花束贈呈	10	5
	プレイステーション	9	4
	プロ野球選手	7	5
	分類語彙表	26	4
	有毛細胞	14	2

表 C.3: large-size テストセットの 100 クエリ (3)

モーラ	検索語	tf	df
7	杏仁豆腐	16	3
	新婚旅行	18	9
	信用金庫	11	5
	総理大臣	35	18
	高島平	21	2
	東京タワー	16	10
	中野新橋	11	2
	名古屋大学	25	20
	夏目漱石	16	11
	ホワイトリカー	20	1
	村上春樹	18	4
	室町時代	11	9
	ユーゴスラビア	9	3
	ユースホステル	40	6
	ラジオ体操	18	9
	レンタルビデオ	13	9
6	L E D	28	3
	N T C I R	16	6
	岡山県	10	6
	グリム童話	17	2
	京王線	31	17
	サーフボード	22	2
	散歩コース	10	8
	正倉院	28	1
	新選組	12	2
	占星術	27	3
	パラトグラム	39	3
	バンクーバー	13	5
	半導体	21	2
	光が丘	13	5
	フラタニティー	13	1
	ペットボトル	36	19
	防空壕	14	7
	水ぼうそう	14	1

表 C.4: large-size テストセットの 100 クエリ (4)

モーラ	検索語	tf	df
5	愛知県	21	14
	阿波踊り	11	6
	火山灰	16	2
	カメハメハ	18	3
	金メダル	34	12
	ジュウシマツ	22	2
	隅田川	26	8
	ドーピング	21	3
	土佐日記	9	2
	ハワイ島	21	4
	ホトトギス	26	1
	マスメディア	26	16
	メンチカツ	8	3
	ライオンズ	20	3
	ワンピース	10	8
4	髪型	17	8
	サイパン	14	6
	日暮里	30	2
	ベーグル	26	2
	目黒区	14	5
	山梨	45	24
3	土岐市	12	1



# 付 録 D    NTCIR-10 SpokenDoc-2 タ スク moderate-size タスク moderate-size テストセット の 100 クエリ

NTCIR-10 SpokenDoc-2 タスク moderate-size タスク moderate-size テストセットの  
100 クエリを表 D.1, 表 D.2, 表 D.3, 表 D.4 に示す.

表 D.1: moderate-size テストセットの 100 クエリ (1)

モーラ	検索語	tf	df
18	WWE R 最小化	9	1
12	サポートバクターマシーン	12	5
	質問応答システム	12	2
11	S T D の性能	9	2
	音声ドキュメント処理	25	17
	P L S A モデル	12	1
10	M M I システム	19	1
	機械翻訳モデル	5	3
	ジェフェリー情報量	4	1
	W F S T	6	1
	短時間スペクトル	10	1
	発話区間検出	10	2
	S P S モデル	8	1
9	おはようございます	4	4
	多項式カーネル	4	4
	デモンストレーション	3	2
	転置インデックス	9	2
	背景と目的	12	10
	パッセージ検索	16	3
	マイクロフォンアレイ	16	2

表 D.2: moderate-size テストセットの 100 クエリ (2)

モーラ	検索語	tf	df
8	アーティキュレーション	7	1
	M R R	8	4
	カラオケ方式	9	1
	擬似三音節	12	1
	キタちゃんキタロボ	7	1
	Q A システム	9	1
	高次モーメント	8	1
	五体不満足	8	1
	C J L C	12	5
	情報工学	5	3
	スピードワープロ	7	1
	センシングデータ	4	1
	非可逆圧縮	4	1
	弁別特徴	12	5
	村山富市	4	1
7	M P 3	13	3
	木構造辞書	11	1
	携帯電話	7	3
	講義スライド	13	2
	時論公論	5	1
	セミクローズド	6	1
	単語トレリス	9	2
	名古屋大学	5	3
	バタチャリヤ距離	10	4
	パワーポイント	12	9
	プログラミング	8	5
	P o d c a s t l e	10	2

表 D.3: moderate-size テストセットの 100 クエリ (3)

モーラ	検索語	tf	df
6	I B M	7	3
	A d a b o o s t	5	1
	ウェーブレット	4	1
	A P I	7	4
	S L P	9	6
	L D A	21	2
	エントロピー	24	6
	産総研	3	2
	GMM	20	7
	バイノーラル	4	2
	バッファサイズ	7	1
	ハンズフリー	6	1
	ヒストグラム	14	6
	プライバシー	3	1
	プロトタイプ	11	1
	マトリックス	5	3
	ウィキペディア	6	6
	エンドレス	4	1
	句読点	14	5
5	シーケンス	12	4
	ソーティング	4	2
	大丈夫	8	6
	チューニング	5	2
	聴診器	6	1
	NAMマイク	17	1
	非流暢	11	2
	不完全	5	2
	プロポーズ	11	2
	v o t i n g	24	1
	緑色	10	7
	モダリティー	25	1
	ワイヤレス	15	5

表 D.4: moderate-size テストセットの 100 クエリ (4)

モーラ	検索語	tf	df
4	折れ線	8	2
	キャプション	8	3
	きらきら	3	1
	三振	5	1
	色相	4	1
	シラバス	8	1
	S P O J U S	29	9
	声量	9	2
	中国	5	1
	デフォルト	3	3
	投球	7	2
	東京	6	6
	東北	5	4
	爆発	10	6
	ラッシー	6	1
	ロボット	12	3
3	アニメ	14	2
	茶釜	9	6
	N I S T	5	3
	ブログ	12	1
	劣化	9	4

# 付 録 E    NTCIR-10 SpokenDoc-2 タ スク iSTD タスク用テストセッ トの 100 クエリ

NTCIR-10 SpokenDoc-2 タスク iSTD タスク用テストセットの 100 クエリを表 E.1, 表 E.2, 表 E.3, 表 E.4 に示す.

表 E.1: iSTD 用テストセットの 100 クエリ (1)

モーラ	検索語	tf	df
14	長岡技術科学大学	0	0
12	アカデミックハラスメント	0	0
	WWW	0	0
	ネットワークスペシャリスト	0	0
	山梨学院大学	0	0
11	ウェアラブルコンピューター	0	0
	グロッサリーショッピング	0	0
	日経平均株価	0	0
10	逆フーリエ変換	0	0
	サザンオールスターズ	0	0
	フォルマント周波数	0	0
	ホイールアライメント	0	0
	ユニバーサルスタジオ	0	0
9	英会話学校	0	0
	オバマ大統領	0	0
	グローバリゼーション	0	0
	原子力発電	0	0
	チューリングマシン	0	0
	ピアノ協奏曲	0	0
	ポートフォリオ評価	0	0
	よろしくメカドック	0	0

表 E.2: iSTD 用テストセットの 100 クエリ (2)

モーラ	検索語	tf	df
8	ウィザードオブオズ	0	0
	英語リスニング	0	0
	江南スタイル	0	0
	キッズステーション	0	0
	セメント協会	0	0
	WHO	0	0
	ピボット溶接	0	0
	V T L N	0	0
	プレイステーション	0	0
	プロ野球選手	0	0
	ペナルティーゴール	0	0
	読売新聞	0	0
	ライン川下り	0	0
	ロイター通信	0	0
7	コンサルティング	0	0
	サンタクロース	0	0
	スマート家電	0	0
	タイムテーブル	0	0
	トランザクション	0	0
	夏目漱石	0	0
	ネゴシエーション	0	0
	パリコレクション	0	0
	V T R	0	0
	マルチトラック	0	0
	分かりかねます	0	0

表 E.3: iSTD 用テストセットの 100 クエリ (3)

モーラ	検索語	tf	df
6	厚かましい	0	0
	A T A	0	0
	L T E	0	0
	かわいらしい	0	0
	サブカルチャー	0	0
	C S 研	0	0
	G P U	0	0
	ジブリアニメ	0	0
	ばかばかしい	0	0
	爆弾テロ	0	0
	ホームページ	0	0
	ほったらかし	0	0
	みっともない	0	0
	もっての外	0	0
	U S B	0	0
	喜ばしい	0	0
	量子化誤差	0	0
5	案の定	0	0
	好ましい	0	0
	サバイバル	0	0
	ショットガン	0	0
	セキュリティー	0	0
	ないがしろ	0	0
	名古屋城	0	0
	夏休み	0	0
	ハイジャック	0	0
	ハイジャンプ	0	0
	ばかでかい	0	0
	マスメディア	0	0
	丸の内	0	0

表 E.4: iSTD 用テストセットの 100 クエリ (4)

モーラ	検索語	tf	df
4	安心	0	0
	E T	0	0
	うきうき	0	0
	嘘つき	0	0
	駅前	0	0
	エジソン	0	0
	大阪	0	0
	押し上げ	0	0
	鹿児島	0	0
	ぎゃあぎゃあ	0	0
	くちゃくちゃ	0	0
	月並み	0	0
	でたらめ	0	0
	のほほん	0	0
	ぶっちゃけ	0	0
	ぺしゃんこ	0	0
	マグナム	0	0
	まろやか	0	0
	めろめろ	0	0
	横浜	0	0
	わくわく	0	0
3	宛て名	0	0
	B I G	0	0
2	J A L	0	0



## 付 録F      コンフュージョンマトリクス スコア

誤検出を抑制するパラメータとして利用した，コンフュージョンマトリクスのスコアのうち，ある音素が正解している確率を表 F.1 に示す．また，ある音素が挿入している確率を表 F.2 に，ある音素が脱落している確率を表 F.3 に示す．

表 F.1: ある音素が正解している確率

音素	正解している確率
a	0.905918
i	0.840033
u	0.709136
e	0.795875
o	0.806821
k	0.841231
g	0.609706
ky	0.638441
gy	0.470101
kw	0.000000
gw	0.000000
s	0.833229
z	0.726940
sh	0.823324
j	0.781061
t	0.776874
d	0.698583
ch	0.750851
q	0.708220
ts	0.689849
ty	0.000000
dy	0.018868
n	0.797946
ny	0.480685
h	0.743694
b	0.748198
p	0.700119
hy	0.682074
by	0.438062
py	0.631466
f	0.679543
fy	0.000000
m	0.842834
my	0.253531
y	0.579507
r	0.823989
ry	0.541125
w	0.640532
N	0.828727
sp	0.000000

表 F.2: ある音素が挿入している確率

音素	挿入している確率
a	0.109641
i	0.107181
u	0.117153
e	0.084395
o	0.114957
k	0.020079
g	0.014515
ky	0.001121
gy	0.000843
kw	0.000000
gw	0.000000
s	0.010562
z	0.005489
sh	0.007146
j	0.004580
t	0.016585
d	0.014598
ch	0.004674
q	0.090356
ts	0.006875
ty	0.000000
dy	0.000002
n	0.019421
ny	0.000469
h	0.026087
b	0.008356
p	0.008492
hy	0.001798
by	0.000177
py	0.000333
f	0.004425
fy	0.000000
m	0.012535
my	0.000233
y	0.015319
r	0.034284
ry	0.001568
w	0.021794
N	0.113957
sp	0.000000

表 F.3: ある音素が脱落している確率

音素	脱落している確率
a	0.034345
i	0.049959
u	0.112373
e	0.065707
o	0.068803
k	0.037307
g	0.065660
ky	0.018061
gy	0.027304
kw	0.000000
gw	0.000000
s	0.026681
z	0.027115
sh	0.028188
j	0.028686
t	0.044199
d	0.063314
ch	0.029187
q	0.175062
ts	0.037061
ty	0.022727
dy	0.088050
n	0.042350
ny	0.021828
h	0.079495
b	0.040183
p	0.061708
hy	0.058776
by	0.025353
py	0.019704
f	0.075562
fy	0.000000
m	0.038254
my	0.035311
y	0.131483
r	0.048815
ry	0.041812
w	0.165074
N	0.071398
sp	0.000000

# 付 録 G      コンフュージョンマトリックススコアの検索性能

本研究での、検索語の検出アルゴリズムは DP を用いた単純な方法である．第 4 章ならびに第 5 章ではこの DP を用いた検索語の検出手法について述べた．これらの用語検索エンジンに用いる DP の各遷移コストは編集距離に基づいており、一致の場合は 0、誤りの場合は置換・挿入・脱落に関わらず全て 1 とした．また、ネットワーク型インデックスには NULL 遷移が存在しており、この NULL 遷移に対するコストとして 0.1 を設定した．

本研究では、この編集距離に基づく DP によるインデックスと検索語間の距離計算を用いることで、高い検索性能を示すことができた．

しかし、更なる検索性能の向上を図るためには異なる距離計算尺度を検討する必要がある．

本付録では、この距離計算尺度をコンフュージョンマトリックススコアに置き換えた検索語の検出方法について述べる．

## G.1    コンフュージョンマトリックススコアの導入方法

コンフュージョンマトリックススコア (CM スコア) に基づくインデックスと検索語間の距離の計算は、式 (G.2) から式 (G.4) に示すように算出され、式 (G.1) に示すように適用される．

$$D(i, j) = \min \begin{cases} D(i, j-1) + Cm_{Del}(j) \\ D(i-1, j) + Cm_{Ins}(i) \\ D(i-1, j-1) + Cm_{Cor}(i, j) \end{cases} \quad (G.1)$$

$$Cm_{Del}(j) = 1.0 - P(\phi, Query(j)) \quad (G.2)$$

$$Cm_{Ins}(i) = \min \begin{cases} 1.0 - P(p, \phi) : \forall p \in PTN(i) \\ 0.1 : NULL \in PTN(i) \end{cases} \quad (G.3)$$

$$Cm_{Cor}(i, j) = \begin{cases} 1.0 - P(p, Query(j)) \\ \quad : \exists p \in PTN(i), \\ \quad \quad p = Query(j) \\ 0.0 : Query(j) \notin PTN(i) \end{cases} \quad (G.4)$$

表 G.1: コンフュージョンマトリックススコアベースの距離計算を行う PTN の構成内容

音声認識システムの種類	N-Best	仮説数
WBC/*, WBH/*, CB/*, BM/*, Non/*	1	10

表 G.2: 距離計算尺度による検索性能の比較

距離計算尺度	F-measure	MAP	MRP
EditDist.	0.64	0.81	0.75
CM Score	0.50	0.78	0.71

$D(i, j)$  は DP 格子上的  $(i, j)$  の位置に至るまでの距離である。

$Query(j)$  は検索語の  $j$  番目の音素を表し,  $PTN(i)$  は PTN の  $i$  番目の Node が持つ Arc の集合を表す. また,  $p$  は PTN の  $i$  番目の Node が持つ, ある Arc の音素を表す.

$P(i, j)$  は CM の確率を表し,  $\phi$  は空文字を表す. つまり  $P(i, j)$  において  $i = j$  のとき正解率を表し,  $P(\phi, j)$  のとき  $j$  が脱落する確率,  $P(i, \phi)$  のとき  $i$  が挿入する確率を表す.

## G.2 評価実験

検索性能の比較のためのインデックスは, 10 種類の音声認識システムの 1-Best 出力を音素単位でネットワーク型インデックスとして構築した PTN である. この PTN は表 G.1 に示す内容で構築されている.

この評価実験で用いたテストセットは, 日本語 STD 用テストコレクションの未知語テストセットである. また, 用いた評価尺度は, Recall-Precision カーブと F-measure, MAP, MRP である.

表 G.2 に, 編集距離に基づく距離計算 (EditDist.) とコンフュージョンマトリックススコアに基づく距離計算 (CM Score) の検索性能を示す. また, 図 G.1 に Recall-Precision カーブを示す.

実験結果より, 距離計算尺度に編集距離を用いることが, コンフュージョンマトリックススコアを用いる場合より高い検索性能が得られることが示された.

しかし, MAP や MRP に関してはあまり違いがないことから, コンフュージョンマトリックススコアのインデックスと検索語間の距離計算式への適用方法を変更することによって, 検索性能が改善される可能性がある.

今回の実験では CM スコアを単純に導入している. 特に, どの音素がどの音素に誤認識され易いかというスコアを用いていない. この置換誤りのコンフュージョンマトリックススコアを導入することによって, 検索性能が改善される可能性がある.

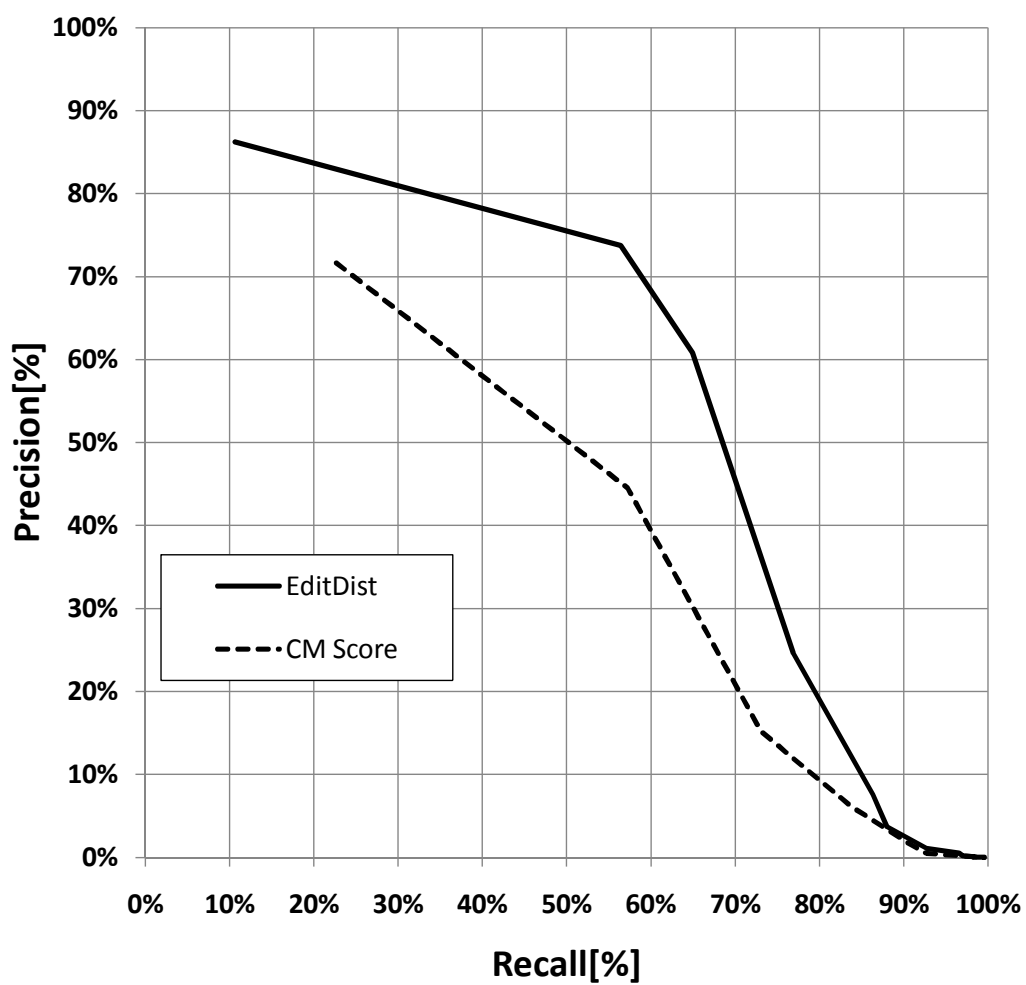


図 G.1: 距離計算尺度による検索性能の比較

## 付 録 H 単一の音声認識システムの検索性能

本研究で用いた 12 種類の音声認識システムのうち、LM に CSB を用いた音声認識システム以外の 10 種類の音声認識システムごと音声中の検索語検出性能を示す。

インデックスの形態としては、サブワードベースインデックスとして PHO(1-Best) と PHO(10-Best)、ネットワーク型インデックスとして PCN の検索性能を示す。

この評価実験で用いたテストセットは、日本語 STD 用テストコレクションの未知語テストセットである。なお、検索性能として示す評価尺度は、Recall-Precision カーブ、F-measure、MAP である。

表 H.1 に、単一の音声認識システムの検索性能を示す。図 H.1 から図 H.10 に Recall-Precision カーブを示す。

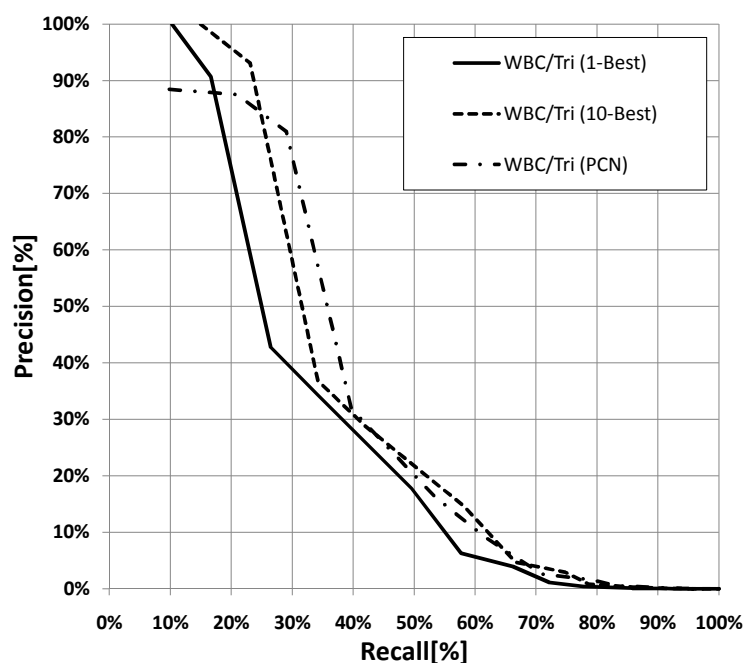


図 H.1: WBC/Tri の検索性能



表 H.1: 単一の音声認識システムの検索性能の比較

インデックス	F-measure	MAP
WBC/Tri(1-Best)	0.34	0.51
WBC/Tri(10-Best)	0.37	0.57
WBC/Tri(PCN)	0.43	0.59
WBH/Tri(1-Best)	0.43	0.57
WBH/Tri(10-Best)	0.48	0.60
WBC/Tri(PCN)	0.54	0.64
CB/Tri(1-Best)	0.49	0.62
CB/Tri(10-Best)	0.53	0.70
CB/Tri(PCN)	0.57	0.69
BM/Tri(1-Best)	0.51	0.62
BM/Tri(10-Best)	0.54	0.69
BM/Tri(PCN)	0.56	0.72
Non/Tri(1-Best)	0.49	0.65
Non/Tri(10-Best)	0.49	0.67
Non/Tri(PCN)	0.47	0.68
WBC/Syl(1-Best)	0.18	0.25
WBC/Syl(10-Best)	0.20	0.32
WBC/Syl(PCN)	0.18	0.33
WBH/Syl(1-Best)	0.26	0.31
WBH/Syl(10-Best)	0.28	0.39
WBC/Syl(PCN)	0.28	0.40
CB/Syl(1-Best)	0.32	0.41
CB/Syl(10-Best)	0.33	0.48
CB/Syl(PCN)	0.33	0.54
BM/Syl(1-Best)	0.32	0.39
BM/Syl(10-Best)	0.37	0.45
BM/Syl(PCN)	0.37	0.47
Non/Syl(1-Best)	0.28	0.41
Non/Syl(10-Best)	0.30	0.45
Non/Syl(PCN)	0.27	0.47

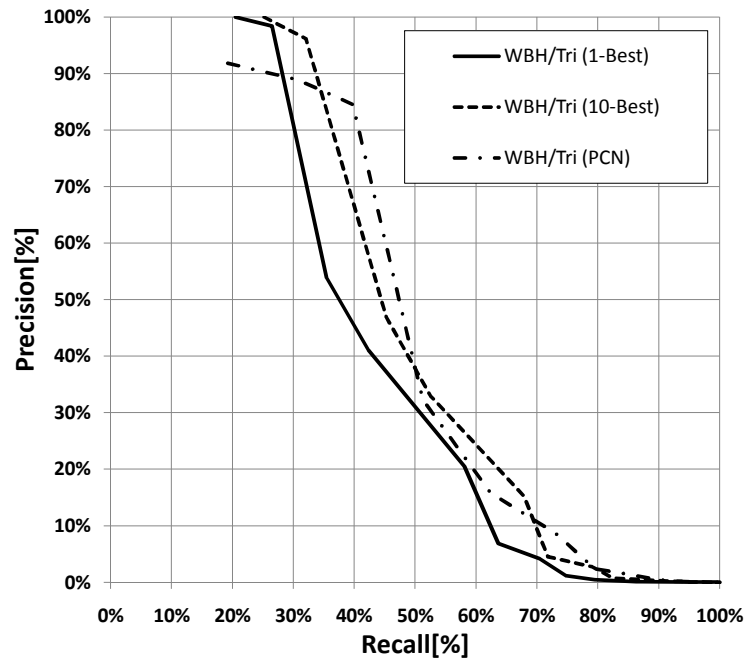


図 H.2: WBH/Tri の検索性能

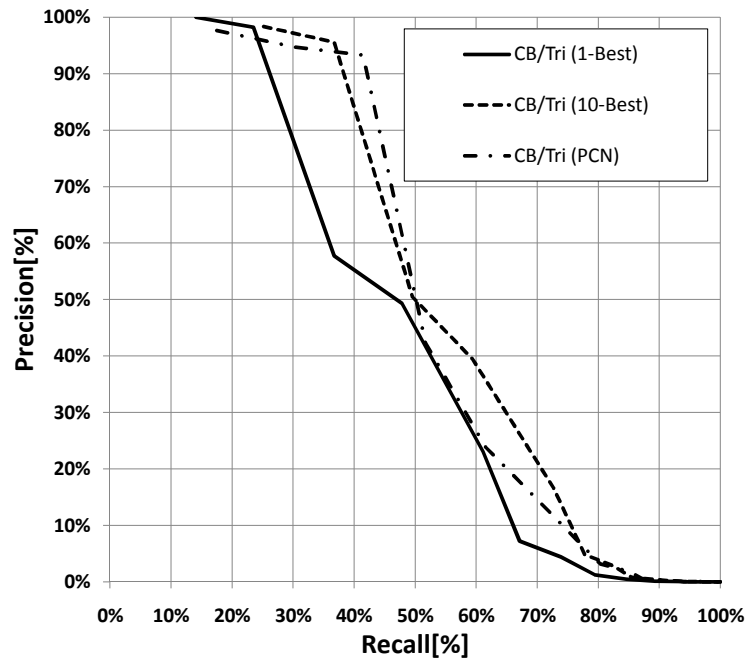


図 H.3: CB/Tri の検索性能

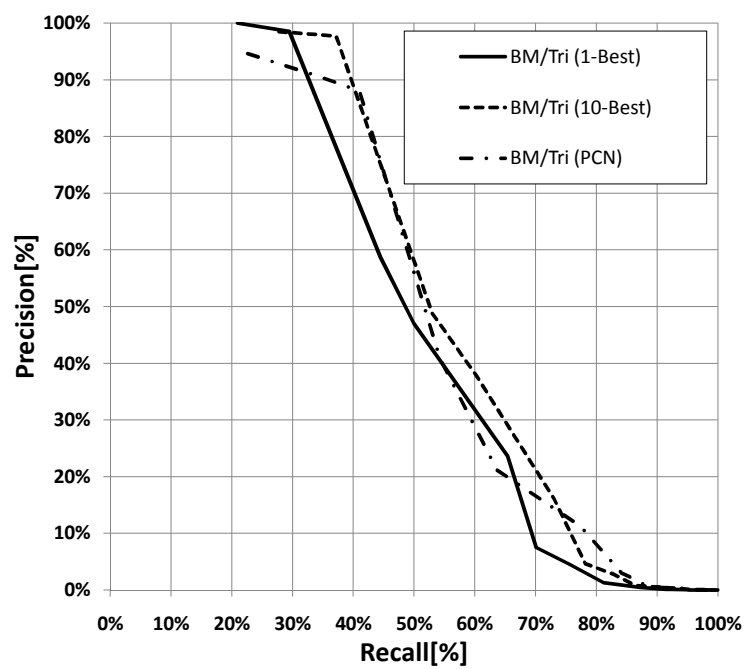


図 H.4: BM/Tri の検索性能

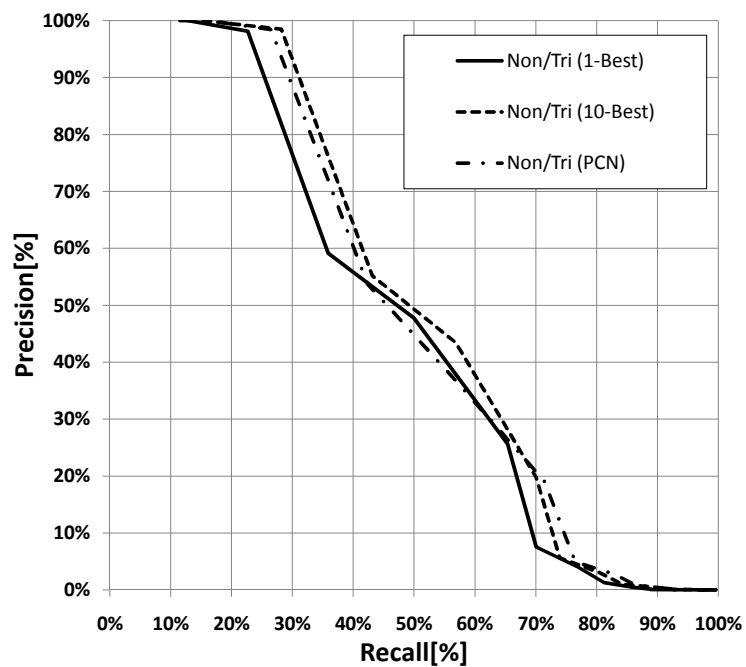


図 H.5: Non/Tri の検索性能

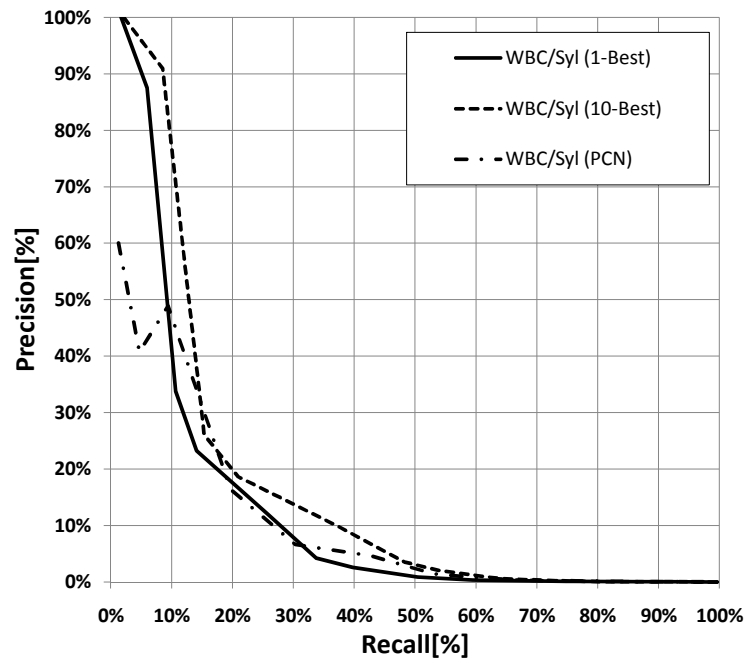


図 H.6: WBC/Syl の検索性能

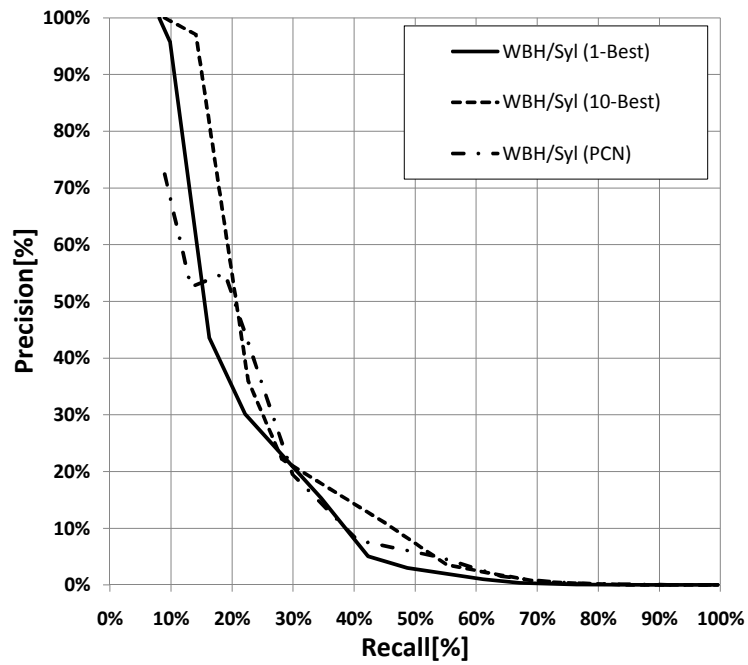


図 H.7: WBH/Syl の検索性能

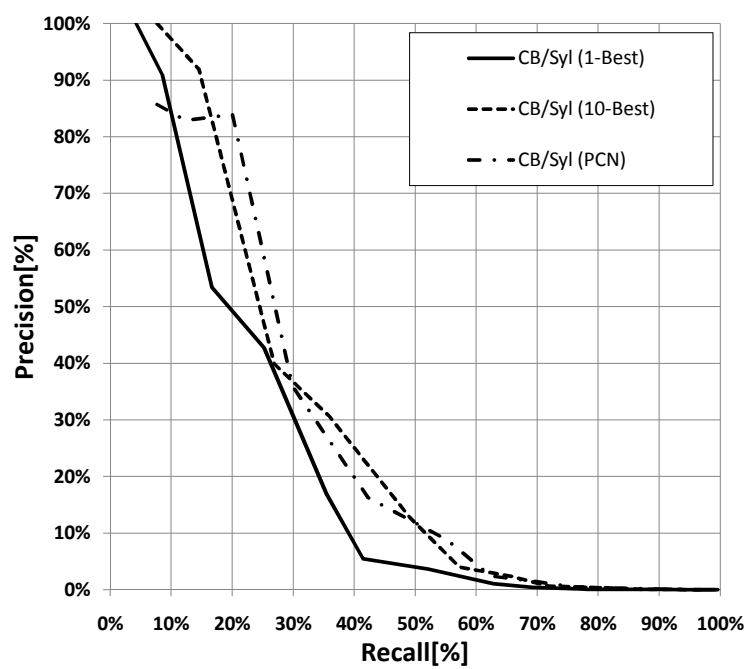


図 H.8: CB/Syl の検索性能

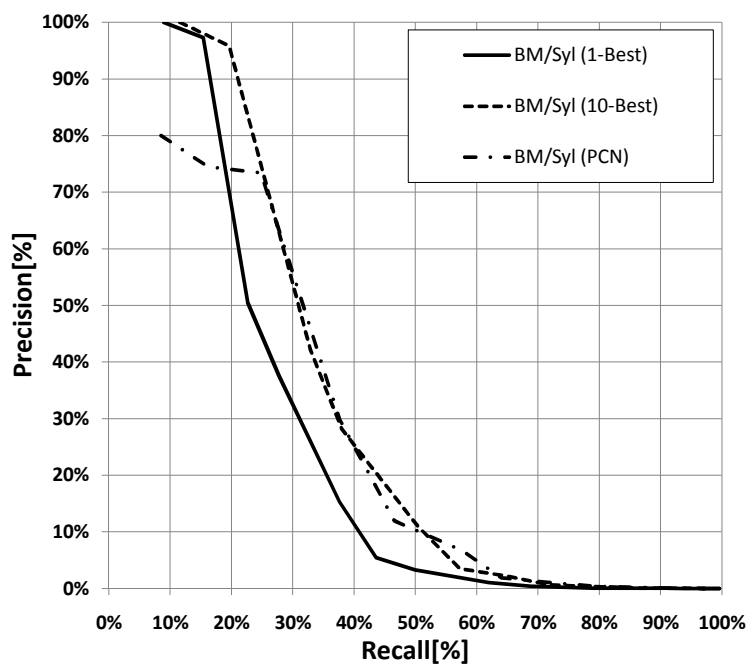


図 H.9: BM/Syl の検索性能

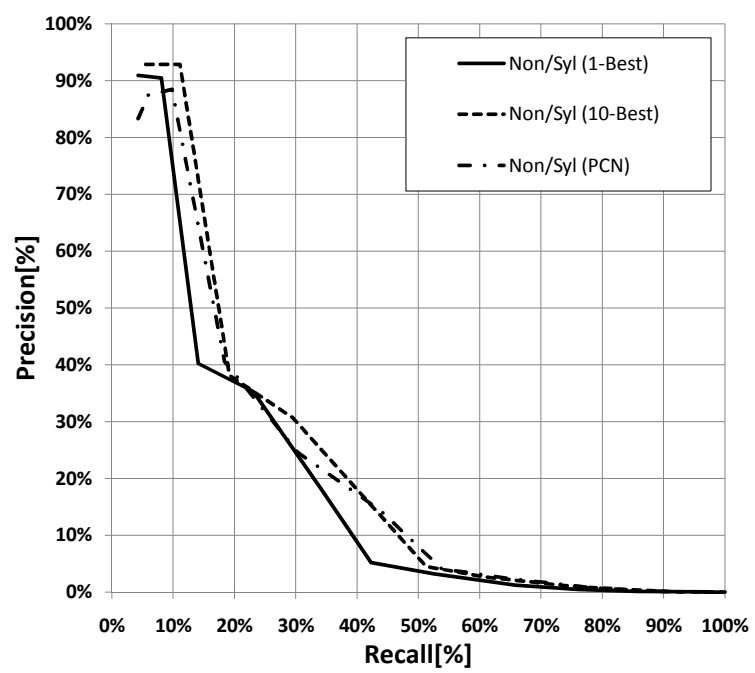


図 H.10: Non/Syl の検索性能

# 付 録I 既知検索語の検索性能

本研究では，検索語が未知語である場合に焦点を当て，検索語の検出性能を改善させる手法について述べた．

本付録では，検索語が既知語である場合において，提案する STD 手法が有効であるかを調査した．

## I.1 検索性能の比較実験条件

検索性能の比較に用いたインデックスは，表 I.1 に示すものとなる．

表 I.1 中の “\*” は全ての音響モデルを表す．Word-base は音声認識結果を形態素単位のまま利用するインデックスであり，この形態素単位の認識結果に対して完全に一致するものを検出したものである．PTN (with Voting) は，PTN に対して誤検出抑制パラメータの “Voting” を適用したものである．

この評価実験で用いたテストセットは，日本語 STD 用テストコレクションの既知語テストセットである．また，用いた評価尺度は，Recall-Precision カーブと F-measure, MAP である．

表 I.1: 既知検索語の検索性能の比較実験に用いたインデックスの種類

インデックス	インデックス の種類	音声認識システムの種類
Word-base	Word-base	WBC/Tri
WBC/Tri(1-Best)	PHO(1-Best)	WBC/Tri
WBC/Tri(10-Best)	PHO(10-Best)	WBC/Tri
WBC/Tri(PCN)	PCN	WBC/Tri
10PHOs(1-Best)	nPHOs(1-Best)	WBC/*, WBH/*, CB/*, BM/*, Non/*
PTN (only EditDist)	PTN(1-Best)	WBC/*, WBH/*, CB/*, BM/*, Non/*
PTN (with Voting)	PTN(1-Best)	WBC/*, WBH/*, CB/*, BM/*, Non/*

表 I.2: 既知検索語の検索性能の比較

インデックス	F-measure	MAP
Grep (simple)	0.69	N/A
WBC/Tri(1-Best)	0.72	0.68
WBC/Tri(10-Best)	0.73	0.71
WBC/Tri(PCN)	0.73	0.73
10PHOs(1-Best)	0.79	0.75
PTN (only EditDist)	0.77	0.78
PTN (with Voting)	0.77	0.81

## I.2 検索性能の比較結果

表 I.2 に、既知検索語の検索性能を示す。また、図 I.1 と図 I.2 に Recall-Precision カーブを示す。

## I.3 考察

実験結果より、単一の音声認識システムの出力を利用する場合と比較し、複数の音声認識システムの出力を利用することによって、検索性能が改善されることが示された。

Recall-Precision カーブでは、単一の音声認識システムの出力を用いた場合ではインデックスの形態によって検索性能が大きく変化することはなかった。また、複数の音声認識システムの出力を用いた場合においても、同様の結果が得られた。

しかし、MAP による比較結果では、ネットワーク型のインデックスを構築することによって検索性能が改善されている。

以上より、提案手法は検索語が未知語か既知語に限らず、音声中の検索語検出性能を改善させることに有効であることが示された。



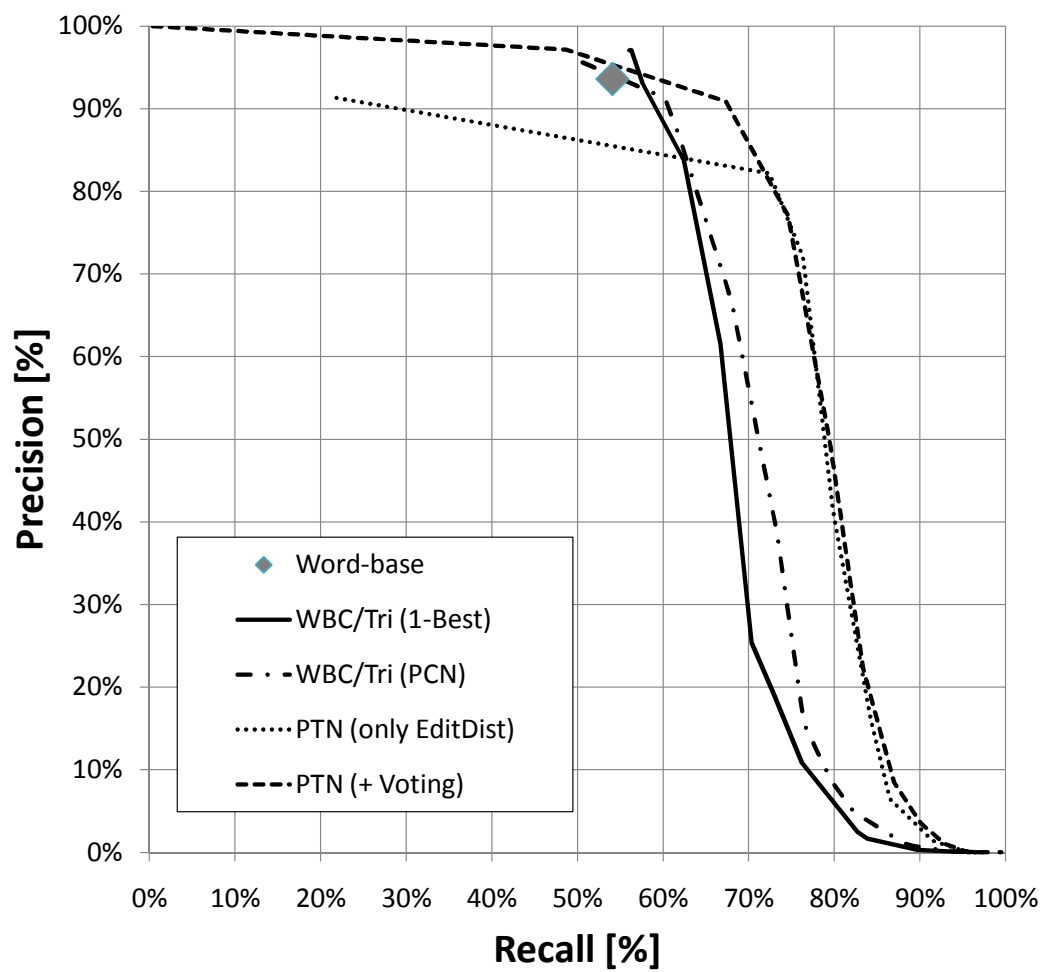


図 I.1: 単一の音声認識システムと提案手法の比較

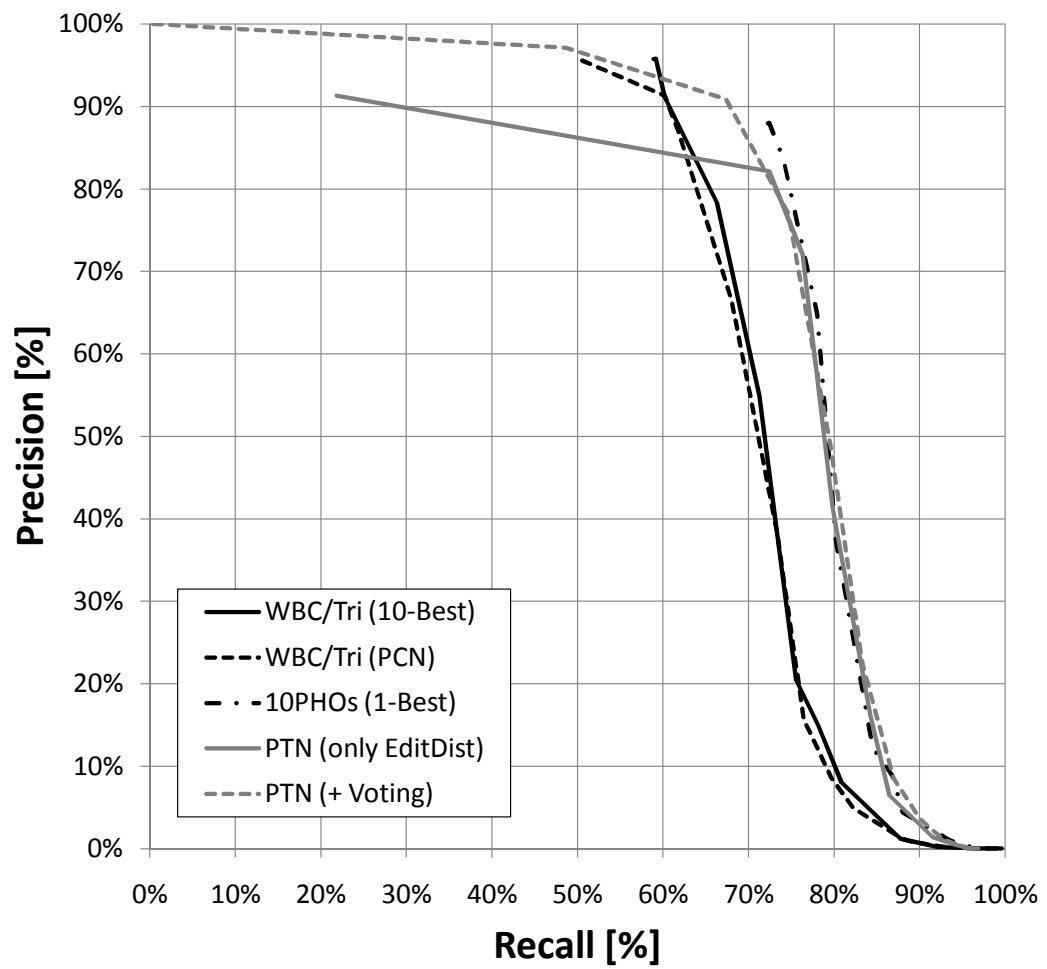


図 I.2: 10 個の音声認識結果を用いた場合の検索性能の比較