Vol.2009-SLP-78 No.4
2009/10/23

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-

伊藤慶明[†1]，西崎博光[†2]，胡新輝[†3]，南條浩輝[†4]，
秋葉友良[†5]，相川清明[†6]，河原達也[†7]，
中川聖一[†5]，松井知子[†8]，山下洋一[†9]

TREC において Spoken Document Retrieval（SDR:音声ドキュメント検索・音声文書検索）の Track が 1996 年～2000 年に設定され，2006 年には NIST を中心に音声中の検索語検出（Spoken Term Detection : STD）タスクが設定され[1][2]，以降海外では盛んに SDR，STD に関する研究が行われるようになった．情報処理学会音声言語情報処理研究会（SIG-SLP）で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を立ち上げ，これまでに SDR 評価用テストコレクションの構築を進めてきた[3]．これに続き 2008 年から音声中の検索語検出（STD）の評価用テストコレクションの構築を開始した．本稿ではこのテストコレクション構築に当っての方針，進捗状況，問題点，構築スケジュールを説明するとともに，現段階で構築したテストセットについて解説する．

Development of Test Collection for Spoken Term Detection – Interim Report -

Yoshiaki Itoh[†1], Hiromitsu Nishizaki[†2], Xinhui Hu[†3], Hiroaki Nanjo[†4], Tomoyosi Akiba[†5], Kiyoaki Aikawa[†6], Tatsuya Kawahara[†7], Seiichi Nakagawa[†5], Tomoko Matsui[†8] and Yoichi Yamashita[†9]

Spoken Document Retrieval (SDR) was dealt with in one of tracks of TREC from 1996 to 2000. NIST supplied a task for Spoken Term Detection (STD) in 2006. Many researches have been conducted as for SDR and STD after these projects. A working group for spoken document processing of SIG-SLP (Spoken Language Processing) in IPSJ also aimed to activate the researches for spoken document processing, and developed a test collection for SDR so far. The working group started to develop a test collection for STD last year. This paper reports the policy of the collection, the development progress, the problems, the schedule and the details of the test collection developed so far.

1. はじめに

近年，パソコンのマルチメディア環境，高速なインターネット，大容量の HDD-ビデオレコーダが普及し，撮り溜めた TV 放送，長期間に渡り録画した家庭用ビデオ，講義や教材用ビデオ，動画サイトでのビデオコンテンツなど，音声・動画を含んだマルチメディアコンテンツが増加・大容量化している．これに伴いこれらの大量のデータから見たい・聞きたい部分を検索したいという機能が求められるようになった．音声を含むデータに対しては，音声認識技術を適用してデータを検索する方式が有望であり，音声ドキュメント検索あるいは音声文書検索（Spoken Document Retrieval: SDR）として既に様々な研究が行われてきている．

音声ドキュメント検索においては，ビデオや講義音声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，クエリに関連する音声ドキュメントを特定することをアドホック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と呼ぶ．検索の基本的な枠組みでは，まず音声ドキュメントを単語ベースで音声認識しておき，その認識結果である単語列に対してテキスト検索の技術を用いて音声ドキュメントを特定する．性能を評価する際，音声認識では音声データの「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データの「質」だけでなく「長さ」，「正解箇所の数」にも影響される（例えば，1 時間の音声データから探す場合，10 時間の音声データから探す場合，正解が全く含まれていない場合，これらの検索性能の比較は困難）．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，クエリ，正解に基づいて評価が行われることが望ましい．

TREC (Text REtrieval Conference) においては，Spoken Document Retrieval の Track が 1996 年の TREC-6 から取り上げられ，TREC-7～9 を経て 2000 年まで行われた[4]．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を立ち上げ，既に SDR 評価用テストコレクションの構築を進めてきた[3].

[†1] 岩手県立大学 Iwate Prefectural University, Japan
[†2] 山梨大学 University of Yamanashi, Japan
[†3] 情報通信研究機構 NICT
[†4] 龍谷大学 Ryukoku University
[†5] 豊橋技術科学大学 Toyohashi University of Technology
[†6] 東京工科大学 Tokyo University of Technology
[†7] 京都大学 Kyoto University
[†8] 統計数理研究所 The Institute of Statistical Mathematics
[†9] 立命館大学 Ritsumeikan University

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できたとしても，その結果は一覧性・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聞いてみないと確かめられない．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中から特定（音声中の検索語検出：Spoken Term Detection: STD）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[5][6]，未知語の検索機能は不可欠である． NIST が 2006 年に STD を新たなテーマとして設定して以降，未知語の検出を重視した STD の研究が盛んに行われるようになり，2009 年の ICASSP でも音声ドキュメント検索（SDR，STD）のセッションが組まれていた．このような状況を踏まえて先のワーキンググループが日本語アドホック音声ドキュメント検索用テストコレクションに続き，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，現在公開に向けてテストコレクションの構築作業を進めている．

本稿では，まず NIST が 2006 年設定したテストコレクションを概説し，日本語テストコレクション構築に当っての方針，進捗状況，問題点，構築スケジュールを説明するとともに，現段階で構築したテストセットについて解説する．また本報告を機に関連する研究者から意見を伺い，最終的なテストコレクションに反映させたいと考える．

2. NIST における Spoken Term Detection（STD: 音声中の検索語検出）用テストコレクション[1][2]

本章では 2006 年に NIST が設定した STD タスクにおける，音声ドキュメントデータ，クエリ，正解，評価方式について簡単に説明する．

2.1 音声ドキュメントデータ

音声ドキュメントデータは表 1 の通りで，アラビア語（近代標準語とレバノン系），中国語（標準語），英語（米語）の 3 種の言語について，放送ニュース音声(Broadcast News)，電話での会話音声(Telephone Conversation)，会議音声(Roundtable Meetings)の 3 種の音声タイプについて，1～3 時間のデータが用意された．

表 1. 評価用音声ドキュメント評価用データ（言語と音声タイプとドキュメント時間）とクエリ数とその出現回数

		Arabic	Chinese	English
Broadcast News		1 hour	1 hour	3 hours
Telephone Conversation		1 hour	1 hour	3 hours
Roundtable Meetings		No	No	2 hours
Number of	Reference	1100	1120	1100
	Occurrences	5240	3684	14421

2.2 検索語（Query terms）

1～4単語から成るクエリ1100個が用意された．3，4グラムは100個，バイグラムは約400個，ユニグラムは約500個が選定された．3単語から成るクエリの場合，トライグラムを高頻度順に並べ，トライグラム中の各単語のユニグラム確率が高いものを除外し上位80個選定する．10個はコーパス外の最新のトライグラムを選定し計90個とした（4グラムは10個）．バイグラムは3,4グラムをバイグラムに展開したもの201個，トライグラムと同様な方法で189個，コーパス外のトライグラムを展開したもの20個が選定された．以下にいくつかの例を示す．単語表記上，同一なら発音が異なっても同一と見なしている（Ex. “wind” (moving air) と“wind” (twist)）．一方，“grasshoppers” は “grasshopper”，“cat” が部分一致する “catalog” は同一とは見なさない．

“grasshopper”, “New York”, “in terms of”, “overly protective”, “Albert Einstein”, “Giacomo Puccini”.

NISTのSTDタスクにおいてはコンテスト形式での評価が行われた（参加者はForbidden Data以外は学習用に利用可能）．このためクエリの単語が既知語か未知語かは明確には定められていない．表2に英語についての各音声のタイプにおけるクエリの種類数と出現数を，図1にクエリの種類数について音声タイプ間での重複などを示す．

表2．各音声タイプ別クエリの種類数・出現数(英語)

		#Term	#Occ.
Broadcast News		898	4893
Telephone Conversation		411	5856
Roundtable Meetings		241	3672
Number of	Reference	1100	
	Occurrences		14421

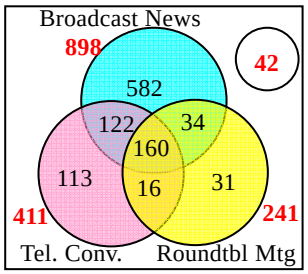

図1. 音声タイプ間のクエリの重複

2.3 正解と評価方式

ある検索語に対して，システムが出力した区間の中心時刻が，実際にその検索語が話されている位置との差が 0.5 秒以内であれば正解と見なされる．

評価は検出精度と速度で行われる．速度については以下の項目が評価の対象となる．

- Indexing time
- Indexing memory consumption (high water mark)
- Index size (on disk)

・Search time for each term
・Searching memory consumption (high water mark)
・Computing platform benchmarks

STD においては性能が検索対象の音声ドキュメント長，正解の出現頻度，検索語の長さ等，様々な要素により影響される．今回，誤って検出する確率として性能を評価することで出現頻度等の問題を回避すべく，DET (Detection Error Tradeoff:ミスの確率と FA の確率のカーブ)と ATWV (Actual Term Weighted Value) ，MTWV (Maximum Term Weighted Value)等の新たな指標が導入された．

3. 日本語版：音声中の検索語検出テストコレクション概要

本章では，SIG-SLP の音声ドキュメント処理ワーキンググループが現在構築を進めている日本語 Spoken Term Detection テストコレクションの構築について説明する．

3.1 基本的な方針

既に SDR，STD 研究を行っている研究者だけでなく，新たに関連する研究を始める研究者も利用できるよう，「SDR/STD 研究・開発を行う多様な利用者を想定し，複数かつ単純な検索語と正解セットの提供」を目指すこととした．

提供・公開するものは，音声ドキュメントの認識結果と使用した音声認識辞書，複数の検索語セット，正解情報，一致検索等のベースライン検索性能を想定している．音声認識システムの辞書に登録されていない単語(未知語)についても考慮したいと考えている．

3.2 現在の構築状況とテストコレクションのサンプル

3.2.1 音声ドキュメント

音声ドキュメントは日本語話し言葉コーパス（CSJ）を検索対象ドキュメントと想定している．CSJ 自体は利用者が別途入手することを前提にしている．CSJ は実際の講演音声と模擬講演から構成されており，全 2702 講演，604 時間の音声コーパスである．そのうち 177 講演には詳細なラベルが付与され，コア講演と呼ばれる．語彙数約 25,000 語の音声認識システム Julius を用いてこの音声ドキュメントを認識し，N ベストの認識結果を得た．したがって，音声ドキュメントについてはこれらの認識結果と辞書の提供が可能である．

3.2.2 検索語セット

検索語のセットは以下の４つのセット案を現段階で策定した．

(1) 既知検索語セット：2702 講演用 100 検索語
(2) 既知検索語セット：コア講演用 50 検索語
(3) （低頻度）未知検索語セット：コア講演用 50 検索語，全講演用 50 検索語
(4) 簡易性能評価用 50 検索語

以下，それぞれの検索語セットを設定するに当っての考え方を紹介する．

(1) 既知検索語セット：2702 講演用 100 検索語

全音声ドキュメント 2702 講演を対象として，先の音声認識システムの語彙に登録されている検索語を 100 個用意した．検索語は 1 語以上からなる内容語とした．実際の検索場面を想定し，その検索語が検索において意味がある言葉であるよう TF・IDF 値も検索語を選定する際に考慮した．検索語長に性能が左右されるため，現在 4〜10 モーラの検索語を 12 個ずつ程度用意した．以下にモーラ数毎にサンプルを示す．付録 A-1 に全リストを示す．

12 広島風お好み焼き	11 音声対話システム	10 重要文抽出
9 形態素解析	8 基本周波数	7 イントネーション
6 商店街	5 お婆ちゃん	4 地下鉄

(2) 既知検索語セット：コア講演用 50 検索語

CSJ のコア 177 講演を対象とした．(1)に比べ小規模な検索対象とした．検索語は全て (1)に含まれる．現在 4〜12 モーラの検索語を 8 個ずつ程度用意した（9 モーラ以上は適切なものが少ないため 8 個未満)．付録 A-1 にリストを示す．

(3) （低頻度）未知検索語セット：コア講演用 50 検索語，全講演用 50 検索語

前述した音声認識システムの辞書はカットオフ 4 で作られたため出現回数が 3 回以下の単語（列）は（コア講演以外）辞書に登録されていない（学習データについては課題で述べる)．即ち未知語となる．このため低頻度の検索語だけを未知検索語として用意した．コア講演用，全 2702 講演用，それぞれについて検索語 50 個とした．付録 A-2 にリストを示す．

(4) 簡易性能評価用 50 検索語

コア講演中の 49 講演約 13 時間を音声ドキュメントとして，既に[6][7]で検索語が未知語として検索性能の評価を行っている．この実験の際に用いた検索語，音声ドキュメントセットと性能を提供することにより，簡便に自分のシステムを評価する枠組みを提供する．付録 A-3 にリストを示す．

3.2.3 サンプル検索性能

上記(1)および(2)の既知語検索について簡単な評価実験を行った．3.2.1 節で述べたように，音声ドキュメントを音声認識し，認識結果のテキストデータを用意した．クエリはすべて既知語であるため，Unix の”grep”コマンドを利用することで，認識テキストデータ中にクエリが存在するか否かを単純に調査した．その結果は以下の通りである．

(1) Recall:68.0 (12.5〜96.0)　Precision:97.0 (76.2〜100.0)　F 値:80.0 (22.2〜98.0)
(2) Recall:58.6 (9.09〜90.9)　Precision:97.5 (42.9〜100.0)　F 値:73.0 (16.7〜95.2)

付録 A-1 に個々の検索語，および正解数，検索性能を示す．上記のように Precision がほぼ 100%なので（FA=0)，実質，ATWV=Recall となり，(1)に対しては，0.6，(2)に対しては 0.5 程度となる．

3.2.4 考察と課題

（1）検索対象の音声ドキュメントについて

・既知語セットの場合，検索語自体へ誤認識するケースが少なく precision は 100%になるものが多い．今回のセットには，湧きだしエラーが多い検索語はほとんど編入していないが，湧きだし誤りが多い検索語（ただし再現率は 100%に近い）を編入させることもできる

・単純な検索においても性能が高いと思われる

・検索対象ドキュメントについてコアと全講演を比較した場合，講演数の少ないコアの方が性能が低い．これは音響モデル・言語モデルの学習データにコア以外の講演を用いたためで，コアの方はオープンな評価となっている．再度，コア以外の講演をクロスバリデーション等でオープンな条件で認識を行う必要がある．しかし，学習データによって既知語・未知語が異なるため，この条件が必要であれば，検索語選定から全てをやり直さなければならない

・上述した学習後，辞書作成時のカットオフを 4 としたため，コア講演以外では出現回数が 3 以下のものは未知語となる．したがって出現回数が 4 回以上あるものは未知語にはならない（コア講演以外）

・未知語を含む検索語を全 2702 講演から探すのは難しすぎるという指摘もあり，1 講演から数箇所の正解を探すタスクで良いとの考え方もある．NIST の評価では 2,3 時間の検索対象ドキュメントであり，国際会議で報告する場合，15 分の検索対象では評価されないことが危惧される．そこで， 以下のように検索対象講演の時間数を変化させた性能評価方法を検討中である

  ・未知語検索語を含む講演　　→　ターゲット講演
  ・未知語検索語を含まない講演→非ターゲット講演
  ・１つの未知語検索語に対しターゲット講演性能だけでなく，非ターゲット講演を追加した時の性能を提示(１時間，10 時間等）

上記で付加する非ターゲット講演のセットを具体的に指定する．

（2）正解セットについて

検索語が決まれば正解セットが決まる．正解の定義案は以下 2 通りが考えられる．

案 1：正解区間の中心と検出区間の中心との差が 0.5 秒以内（NIST 基準）

案 2：検出区間が正解フレーズに含まれる

案 1 の NIST 基準とする場合，CSJ にはフレーズ単位の境界時刻情報はあるが，単語境界の時刻情報はないため，正解区間の境界時刻情報を新たに付与する必要がある．正解判定はそのフレーズを聞くことになるため，案 2 のように正解フレーズを特定できれば良いとの考え方もある．案 1 とした場合，音声認識システムを用い強制アライメントにより正解区間の開始終了時刻を算出することを検討している（その結果を人手で再確認が必要か，精度をみて判断したいと考える)．

（3）検索性能評価について

①ベースライン性能

ベースラインとなる検索性能を提示する必要がある．3.2.3 で示した性能のみで良いか検討中である．未知語については，音声認識した結果得られる音素系列からの検索結果は提示できる．以下 2 つのベースライン検索性能の要望も聞いている．

・認識結果の音素系列に対し，Edit distance を用いた連続 DP による性能
・音素音響モデル／音素言語モデルを用いた連続音素認識の結果から得られる音素系列からの検索性能

表 3. SDR・STD 用テストコレクションの比較

名　称	検索時の想定状況	クエリ	検索対象音声ドキュメント	評価方法
TREC-6	既知語検索	47 トピック	ニュースを中心に 50 時間 (1451 ドキュメント)	1 位候補の正解率
TREC-7	アドホック検索	23 トピック 平均 14.7 語	ニュースを中心に 87 時間 (2866 ドキュメント)	MAP(平均適合率)
TREC-8	大規模アドホック音声ドキュメント検索	49 トピック	ニュースを中心に 557 時間 (21,754 ドキュメント，ドキュメント境界なし)	MAP
NIST STD	音声中の検索語検出	約 1000 個 1〜4 単語/個	ニュース 3H，電話会話 3H，円卓会議 2H (英語,アラビア語,中国語)	DET ATWV
JPN SDR	大規模アドホック音声ドキュメント検索	39 トピック	2702 ドキュメント(=講演, 604 時間) 正解は関連あるパッセージ	11 点 MAP F 値等
JPN STD	音声中の検索語検出	50 個× 5 セット	2702 ドキュメント(=講演, 604 時間), 177 コア講演, 1 講演(約 15 分)	MAP F 値, DET ATWV

②評価指標

検索指標としては以下の 3 つを提示したいと考える．

・Precision-Recall および F 値
・DET:ミスの確率と FA の確率のカーブ
・ATWV：検索語の出現頻度を考慮した正解する確率

3.2.5 構築スケジュール

現在，一通りの作業を終えたところであるが，今回の発表を機に SDR/STD 研究者からの意見・要望を伺い，最終的なテストコレクションに反映したいと考える．具体的には 2009 年 12 月までに要望等を集約し，2010 年 3 月末までに最終的なテストコレクションの完成を目指す．

3.3 SDR・STD 用テストコレクション比較

上記テストコレクションとこれまで提供されてきた SDR・STD 用テストコレクションの比較を表 3 にまとめる．

4. まとめ

情報処理学会音声言語情報処理研究会における音声ドキュメント処理ワーキンググループが SDR 評価用テストコレクションの構築に続き STD の評価用テストコレクションの構築を進めている．本稿ではこのテストコレクションとその構築状況について解説した．2009 年 3 月までに最終的なテストコレクションに反映させる予定である．本報告を機に関連する研究者から意見を伺い，SDR/STD 研究の活性化，発展に資することのできる STD テストコレクションを構築していきたいと考えている．

参考文献

1) 2006 Spoken Term Detection Evaluation: http://www.itl.nist.gov/iad/mig/tests/std/2006/index.html.
2) NIST, “The spoken term detection (STD) 2006 evaluation plan,” 2006. http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf
3) Tomoyosi Akiba, Kiyoaki 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,” IPSJ Journal Vol. 50 No. 2 1234–1245 (2009).
4) J. Garofolo, et al.. “The TREC spoken document retrieval track: A success story,” Ninth Text Retrieval Conference (TREC-9), NIST, 2000.
5) B. Logan et. al., “Confusion-based query expansion for OOV words in spoken document retrieval,” Proc. ICSLP, 2002.
6) 岩田他，“語彙フリー音声文書検索手法における新しいサブワードモデルとサブワード音響距離の有効性の検証,” 情処論文誌，vol. 48, 5, pp. 1990-2000, 2007.
7) 伊藤他，“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ー検索語彙に依存した検索性能推定指標の導入ー,” 情処会論文誌, Vol. 50, 2, pp. 524−533, 2009.

付録 A-1　既知検索語セット：全講演用（左），コア用（右）と検索性能

	For all spoken documents						For core			
Word	tf	df	tfidf	grep hit/df	rcl	prc sn	tf	df	tfidf	grep hit/df
12 モーラ	**4 個**						**1 個**			
国立 国語 研究 所	35	19	174	26/16	74	100	7	5	25	4/4
統計 数理 研究 所	10	8	58	4/4	40	100				
大 語彙 音声 認識	6	5	37	4/3	67	100				
広島 風 お 好み 焼き	14	3	87	9/2	78	64				
11 モーラ	**12 個**						**1 個**			
音声 対話 システム	89	34	389	77/31	85	100				
学習 指導 要領	59	18	296	49/17	85	100				
教師 なし 話者 適応	30	6	183	23/6	77	100				
コンビニエンス ストアー	30	21	146	27/19	90	100				
ウエーブレット 変換	25	3	170	24/3	96	100				
機械 翻訳 システム	24	7	143	16/6	67	100				
東京 ディズニー ランド	21	7	125	15/7	71	100				
内閣 不 信任 案	14	3	95	9/3	71	100				
人工 知能 学会	10	5	63	6/4	60	100				
安全 保障 理事 会	6	1	47	4/1	67	100	6	1	31	4/1
トレード オフ の 関係	9	8	52	8/7	89	100				
環境 音 の 識別	8	1	63	3/1	38	100				
10 モーラ	**12 個**						**2 個**			
TF IDF	79	18	396	66/18	81	94				
ニューラル ネットワーク	69	15	358	48/13	70	100				
重要 文 抽出	45	12	244	29/12	64	100	7	5	29	5/3
阪神 大 震災	33	15	171	29/15	85	97				
サザン オール スターズ	32	4	208	14/3	44	100				
シドニー オリンピック	29	16	149	25/15	86	100				
最大 エントロピー	19	5	120	13/5	68	100				
天体 望遠 鏡	14	3	95	6/2	43	100				

	For all spoken documents						For core			
Word	tf	df	tfidf	grep hit/df	rcl	prc sn	tf	df	tfidf	grep hit/df
ワン パス トライグラム	13	1	103	3/1	23	100	13	1	67	3/1
宇宙 戦艦 ヤマト	11	1	87	8/1	73	100				
羊 たち の 沈黙	10	3	68	5/3	50	100				
ドル トン の 原子 説	7	1	55	2/1	29	100				
9モーラ	**12 個**						**8 個**			
形態 素 解析	159	76	568	147/72	92	99	9	3	37	7/3
主 成分 分析	90	31	402	60/28	67	100				
原子 力 発電	44	10	246	35/10	80	100				
プロスペクト 理論	34	4	222	30/4	88	100				
音素 認識 率	32	7	191	28/1	94	100	16	1	83	12/1
ワーキング ホリデー	27	10	165	22/6	85	100				
ツー パス デコーダー	18	6	110	10/4	56	100	14	2	63	7/1
ヤ コ ビ 適応 法	16	1	126	12/1	81	100				
ヒト ゲノム 計画	14	3	95	13/3	93	100				
シラブル の 構造	9	1	71	7/1	89	100	9	1	47	7/1
キー ワード 抽出	14	5	88	13/5	86	92	5	1	25	4/1
エベレスト 街道	6	1	47	2/1	33	100	6	1	31	2/1
8モーラ	**12 個**						**8 個**			
基本 周波 数	287	61	1088	259/61	91	100	21	1	65	17/9
情報 検索	146	51	580	133/54	88	96	18	7	61	16/7
パープレキシティー	121	25	567	104/25	86	99	26	5	93	22/5
絶対 音感	86	3	585	59/3	71	100	38	1	197	19/1
就職 活動	72	29	326	51/25	69	98	5	4	19	3/3
インターラクション	61	27	281	53/30	85	93	18	3	73	11/4
英語 の 勉強	33	18	165	28/18	88	100				
平家 物語	22	4	143	19/4	86	100				
遮断 周波 数	19	6	116	12/6	63	100				
沖縄 の 文学	19	1	150	14/1	79	100				
ウィザード オブ オズ	15	7	89	6/4	40	100	8	2	36	3/2
中央 林間	12	4	78	5/2	42	100	9	2	40	5/2
7モーラ	**12 個**						**8 個**			
イントネーション	199	50	114	190/54	96	98	32	5	114	27/5
NHK	172	95	576	139/84	78	97	20	8	62	14/7
ATR	147	81	516	132/68	65	88	13	7	42	10/6
機械 翻訳	92	25	431	73/24	80	99	15	2	67	12/2
京都 大学	53	35	230	26/11	41	85				
交通 の 便	32	26	149	24/19	73	100	7	4	26	5/3
混合 重み	30	10	168	21/11	63	95	12	2	54	5/2
東南 アジア	44	28	201	11/10	23	91				
有声 休止	26	2	187	15/2	58	100	23	1	119	14/1
遅延 和 アレー	23	4	150	15/4	65	100				
ラジオ 体操	18	9	103	12/5	67	100				
お しゅうとめ さん	15	6	92	5/3	33	100	9	2	40	2/1
6モーラ	**12 個**						**8 個**			
商店 街	207	75	742	137/61	66	99	21	9	63	10/6
大学 院	149	104	485	89/54	60	96	9	8	28	2/2
アナウンサー	132	45	541	89/42	68	96	27	5	96	14/5
研究 室	98	79	346	45/39	45	98	6	6	20	4/4
留学 生	91	32	404	78/32	81	96	16	3	65	11/3
ペット ボトル	36	19	178	30/20	81	97	9	1	47	6/1
世界 遺産	26	14	137	13/11	50	93				
弥生 時代	22	3	150	8/3	36	100				
ADA ブー スト	19	3	129	10/2	37	100				
産 婦人 科	15	10	84	7/6	47	100				
貝殻 虫	13	2	94	5/1	38	100	12	1	62	5/1
調音 地図	12	1	95	4/1	33	100	12	1	62	4/1
5モーラ	**12 個**						**8 個**			
お 婆 ちゃん	244	100	804	162/70	66	97	27	6	91	11/4
アルバイト	221	102	724	236/126	94	88	11	6	37	10/6

Vol.2009-SLP-78 No.4
2009/10/23

	For all spoken documents						For core			
Word	tf	df	tfidf	grep hit/df	rcl	prc sn	tf	df	tfidf	grep hit/df
ダイエット	166	35	722	161/52	89	89				
クラシック	85	39	360	77/40	81	90				
ラーメン 屋	60	25	281	39/22	67	98	10	3	41	6/2
コンクール	39	14	205	36/17	85	92	15	2	67	11/2
鼻 濁音	39	6	238	27/6	69	100	21	2	94	12/2
レントゲン	34	15	177	27/14	79	96				
ドラえもん	30	13	160	25/15	83	93				
予測 誤差	28	6	95	21/6	79	100	16	2	72	13/2
豊島 園	24	5	151	17/6	64	94	21	2	94	13/2
連想 語	16	1	126	2/1	13	100	16	1	83	2/1
4モーラ	**12**						**8**	**個**		
地下 鉄	133	70	486	96/66	69	96	15	9	45	11/8
純音	119	21	578	112/43	77	80	15	3	61	8/5
世田谷	115	39	487	86/39	76	97	22	7	71	9/6
富士 山	92	38	392	68/34	73	99	13	6	44	9/6
鎌倉	65	26	302	66/31	94	90	13	4	49	11/4
青山	57	16	292	35/13	59	97	35	1	181	16/1
大田 区	43	11	237	36/13	73	89				
練馬 区	36	10	202	27/10	65	96	23	4	87	14/4
コサイン	30	18	150	24/19	67	80				
ベルギー	21	8	122	7/6	29	86	11	2	49	1/1
アイヌ 語	20	1	158	17/2	80	94				
ヤシ の 木	18	12	98	21/15	89	76				

付録 A-2　未知検索語セット

コア講演用 50 検索語

mora	tf	df	phones	mora	tf	Df	phones
15	2	1	ホンコンマネタリーオーソリティー	9	1	1	NTTドコモ
15	1	1	コンシューマーキャラクタリスティック				
13	1	1	管領扇谷定正	8	2	1	サブスティチューション
12	2	1	ショアーズアットワイコロア	8	2	1	弱肉強食
12	1	1	灯台内村鑑三	8	2	1	ベネルクス三国
12	3	1	ムジークフェラインスザール	8	2	1	チャーミングセール
12	3	1	エクスポージャープロブレム	8	1	1	暗黒星団
11	2	1	ウエーテッドミューチュアル	8	1	1	アニマルマインド
11	1	1	仮面ライダースナック				
11	1	1	明日香親王一行	7	1	1	黎明書房
10	2	1	公序良俗違反	7	1	1	前代未聞
10	1	1	国旗国歌法案	7	1	1	坂本竜馬
10	1	1	昭島市美堀町	7	1	1	東茨城
10	1	1	冷暖房空調				
10	1	1	三四郎草枕	6	3	1	冷酷無比
9	2	1	レンペルジブ符号	6	3	1	藤子不二雄
9	3	1	ドロンボー一味	5	3	1	富士吉田
9	2	1	芳香族ニトロ	5	2	1	伊予三島
9	1	1	和漢混交文	5	1	1	千駄堀
9	1	1	普天間基地移設	5	1	1	青葉台
9	1	1	専制君主国	5	1	1	虎の門
9	1	1	仁義礼智	5	1	1	京田辺
9	1	1	別冊宝島				
9	1	1	田原総一朗	4	3	1	沙知代
9	1	1	西郷輝彦	4	2	1	多胡羊歯
9	1	1	鹿島アントラーズ	4	3	1	福引き
9	1	1	国電JR	4	3	1	売れ筋

全講演用 50 検索語

mora	tf	df	Query terms	mora	tf	df	Query terms
13	1	1	石川島造船所	7	1	1	一獲千金
12	1	1	セマティックプログレッション				
10	1	1	スーパーコンピューター	6	8	1	アドバイザー
10	1	1	少林寺拳法	6	7	6	LPC
				6	7	1	西日暮里
9	1	1	春桜亭円紫	6	3	1	本駒込
9	1	1	談洲楼焉馬	6	2	1	下北沢
9	1	1	ウエーティングロード	6	2	1	メーンランド
9	1	1	ボディーペインティング	6	1	1	九州佐多
				6	1	1	田舎育ち
8	1	1	スティーブンキング	6	1	1	安室奈美恵
8	1	1	訓読訓点	6	1	1	言い誤り
8	1	1	税制優遇				
8	1	1	春夏秋冬	5	9	1	アルバニア
8	1	1	イトーヨーカドー	5	5	2	アイランド
8	1	1	インターナショナル	5	2	1	岩清水
8	1	1	営団赤塚	5	3	1	三河島
8	1	1	順風満帆	5	4	1	美堀町
				5	1	1	江東区
7	7	1	ユーゴスラビア	5	2	1	那覇港
7	2	2	代々木上原				
7	2	1	ニューハンプシャー	4	27	1	定家
7	2	1	マツモトキヨシ	4	28	2	ネパール
7	1	1	悪戦苦闘	4	3	1	屈斜路
7	1	1	奄美大島	4	3	1	幡が谷
7	1	1	隔世遺伝	4	4	1	国入り
7	1	1	東武伊勢崎	4	5	1	安保理
7	1	1	奥穂高岳	4	6	1	集成

付録 A-3. 簡易性能評価用 50 検索語

mora	tf	Query terms	Mora	tf	Query terms
17	5	コンテキストインディペンデントモデル	6	15	年齢層
15	5	コンテキストディペンデントモデル	6	4	言語スコア
15	11	ワンパストライグラムデコーダー	6	9	声道長
11	3	セグメント統計量	6	29	話者適応
10	15	話者照合システム	6	65	言語モデル
9	8	周波数伸縮	6	7	話者間距離
9	13	ツーパスデコーダー	5	14	ホルマント
8	43	学習データー	5	11	評価話者
8	4	白色雑音	5	17	高齢者
8	29	合成音声	5	37	デコーダー
8	6	伸縮係数	5	5	カバレージ
7	9	リスコアリング	5	14	木構造
7	5	単語終端	4	3	枝刈り
7	5	ピッチ情報	4	36	閾値
7	24	不特定話者	4	32	空間
7	40	クラスタリング	4	45	有声
7	89	エイチエムエム	4	9	湧き出し
7	12	混合重み	4	59	音節
7	9	スポーツニュース	4	18	スポーツ
7	9	一般ニュース	4	166	学習
6	53	パラメーター	4	12	判別
6	21	ベースライン	4	8	マスク値
6	5	最推定	3	71	尤度
6	27	特徴量	3	12	子供
6	108	認識率	3	48	ニュース