社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2006 − SLP − 64 (24)
2006／12／21

音声ドキュメント検索評価のためのテストコレクションの試作

伊藤克亘 (法政大学)　相川清明 (東京工科大学)　秋葉友良 (豊橋技術科学大学)
伊藤慶明 (岩手県立大学)　河原達也 (京都大学)
南條浩輝 (龍谷大学)　西崎博光 (山梨大学)　安田宜仁 (NTT)　山下洋一 (立命館大学)

あらまし　情報処理学会音声言語情報処理研究会の音声ドキュメント処理ワーキンググループの活動の一環として行っている、音声ドキュメント検索評価用テストコレクションについて報告する。試作したテストコレクションは、対象を日本語話し言葉コーパス (CSJ) の講演音声データならびに書き起こしデータとし、テキストクエリとそれに対する正解データ、ならびに音声認識結果から構成される。答が 1 分前後の音声区間となり、5 区間程度の正解が得られるようなクエリを目標に作成して、正解判定をしたところ 100 件程度作成したクエリのうち、33 件のクエリが条件を満たしていた。

キーワード　音声ドキュメント, 情報検索, テストコレクション, 評価

A Prototype of a Test Collection for evaluation of Spoken Document Retrieval System

Katunobu Itou (Hosei Univ.), Kiyoaki Aikawa (Tokyo Univ. of Technology), Tomoyosi Akiba (Toyohashi Univ. of Technology), Yoshiaki Itoh (Iwate Prefectural Univ.), Tatsuya Kawahara (Kyoto Univ.), Hiroaki Nanjo (Ryukoku Univ.), Hiromitsu Nishizaki(Univ. of Yamanashi), Norihito Yasuda(NTT), and Yoichi Yamashita(Ritsumeikan Univ.)

Abstract　The Spoken Document Processing Working Group, which is organized in special interest group of spoken language processing, information processing society of Japan, are developing a test collection for evaluation of spoken document retrieval system. A prototype of the test collection consists of a set of textual queries, relevant segment lists, and the transcription by the automatic speech recognition system to retrieve from the Corpus of Spontaneous Japanese (CSJ). As a result to design queries according to the criteria that a query should have more than five relevant segments that consist from about one minute speech segment, thirty three queries were obtained among about 100 queries.

Key words　Spoken Document, Information Retrieval, Test Collection, Evaluation

1. はじめに

ことばには、二つの重要な機能がある [1]。一つは、相互作用的 (interactional) な機能である。これは、人間がことばを用いて、他人に気持ちを伝えたり、社会的関係を構築・維持したりするための機能である。ことばには、もう一つ重要な機能がある。それは、業務的 (transactional) な機能である。これは、ことばを用いて、知識や技術・情報を伝達する機能である。

業務的な機能は、人類がことばを使いはじめた初期の段階では、口伝えという形で、話しことばが担っていただろう。しかし、文字が発明されると業務的な機能は次第に書きことばが担うようになり、紙、印刷術といった発明によって、書きことばの業務的な機能は飛躍的にひろがってきた。

古代においては、業務的な機能を多く担っていたはずの話しことばは、近代以降においては、相互作用的な機能の面ばかりに使われているように思える。しかし、音声認識などの技術をより洗練させることで、話しことばそれ自体を業務的な機能のメディアとして利用できるようになるのではないだろうか。そのような発明によって、書きことばより話しことばの方が表現しやすい内容の共有、再利用を促進し、人間の新たな知的活動

を刺激することになるのではないだろうか。

書きことばの世界で、特に情報管理の面を強調した形態を「文書」「ドキュメント」とよぶ。この名称にちなんで、我々は、情報処理学会音声言語情報処理研究会に「音声ドキュメント処理ワーキンググループ」を発足させた。このワーキンググループでは、音声ドキュメント利用の第一歩である記号化 (文字化) は音声認識技術で一定の成果をおさめているという前提のもとで、大量のドキュメントを処理するために不可欠な技術である検索技術の研究を促進するために、まずは音声ドキュメント検索研究のための基盤整備に取り組むこととした。

最初の活動として、情報検索の研究には欠かせないテストコレクションの構築を目標としている。本稿では、音声ドキュメント検索研究のためのテストコレクションの試作について報告する。

2. 音声ドキュメント処理

書きことばを対象とした研究のうち、情報管理の側面に着目した処理を総称して「ドキュメント処理」とよぶことがある。ドキュメント処理の研究項目には、文書の交換、閲覧、加工などが含まれる。さらに、文書の構造化や文書の検索 (タグを用いるものと全文検索の両方が含まれる)、質問応答、要約などの技術も含まれる。

我々は、音声ドキュメント処理を音声メディアに対するこれらの処理を総称するものとして定義したい。その処理の対象となるメディアとしては、口頭での指示 (を録音したもの)、電話、ラジオなどがあげられるだろう。その他にも、講義・講演、テレビ、会議、インタビューなどのデータも対象となるだろう。しかし、後者のメディアは、ドキュメントとして利用されるときには音声だけで構成されていることはほとんどなく、テキストや画像・映像 (話し手の身振りや表情も含む) が併用されている。これは、音声が「ライブ」の場、つまり、聴衆と話し手が、(ある種の) 時間や空間を共有する場で用いられる手段であることと密接に関係していると考えられる。この点を考慮すると、音声ドキュメント処理は、マルチメディアドキュメント処理に進展しなければならない必然性を持っていると考えられる。

これまで音声ドキュメント処理として組織的に研究が行われているのは、ニュース (ラジオ、テレビ) の検索 [2], [3] くらいである。しかし、例えば、講演とニュースでは、ドキュメントとしての性質がかなり異なる。ニュースは通常 30 秒から 1 分程度で自己完結的に作られている。一方で、講演は、通常 10 分以上の長さを持ち、さらにサブトピックに分けられるなどの構造を持つ。講演の内部では、一定の内容を持つ 1 分程度のサブトピックを設定することも可能であるが、その区間は、必ずしも自己完結的ではない。このような特徴を持つ講演音声を検索対象とした研究は日本語以外においても、まだほとんどおこなわれていない。したがって、検索対象の単位をどうするかなど、評価の観点、つまり、テストコレクションの設計自体が試行錯誤せざるをえない段階であり、研究的要素を多く含んでいる。

3. 音声ドキュメント検索研究用テストコレクション

3.1 テストコレクション

テストコレクションとは、情報検索や自動要約などの分野で開発したシステムを、ある程度限定した設定のもとで定量的に評価するためのベンチマークのためのデータセットである。特に、正解かどうかを人間が判断して評価せざるをえないような観点の評価に利用される。また、性能評価に再現率を示さなければならない場合にも利用される。

これまで、情報検索の分野では、web 検索 [4]、新聞検索 [5]、論文抄録検索、多言語検索 (新聞) [6]、特許広報検索 [7]、ニュース音声検索 [2]、ニュース画像検索 [3]、著作権切れ映像検索などのテストコレクションが作られてきている。また、質問応答 [8] や自動要約 [9] のテストコレクションも作られている。

3.2 検索評価用テストコレクション

検索評価用のテストコレクションは、クエリ集合、検索対象の文書集合、正解文書集合から構成される。

クエリ集合は一般に数十から数百程度の数のクエリから構成される。クエリは図書館学の専門家が作成した例や、特許検索の専門家が作成した例がある。これらのアプローチは、対象文書集合が現実のタスクと同等規模である場合には有効である。しかし、対象文書集合に対応するタスクが実世界にはない場合などは、検索対象となる文書を決めてから、逆にそれに見合ったクエリを作成してテストコレクションを作ることもある。

対象文書集合は、より実際に即した評価を行うためには、実際の検索の対象となるような分野や多様性や量でなければならない。それらを考慮し、実際に電子化されたデータが手に入るという観点から、日本語の場合は、新聞 (数年分程度)、web 文書 (10GB から 1TB 程度)、特許広報のコレクションが作られている。新聞の場合、1 年間で 10 万記事程度であるため、対象文書の数は 10 万から 1000 万程度の量である。

正解文書集合は、それぞれのクエリに対して、対象文書集合の中から関連する文書のリストを作ったものである。複数の文書が正解となりうることと、関連の度合は一つではないという点に配慮しなければならない。

また、検索の評価の場合、システムが出力した検索結果が正解かどうかが判定できるだけでは不十分であり、検索漏れがないかどうかも重要な尺度となる。この点を評価するためには、全ての正解をリストアップする必要がある。しかし、前述のように、正解が複数あり、正解判定は関連があるかどうかという観点でおこなうため、厳密には、全ての文書を検査しない限り、検索漏れがあるかどうかはわからない。

しかし、全ての文書を検査するのは、コストの面や作業時間の面で現実的でないことが多い。そこで、これまでのテストコレクションの正解判定では、プーリングという手法がとられることが多かった。プーリングに基づく手法では、まず、評価対象のシステムが各クエリに対して十分な数の結果を出力する。それらをクエリごとに全て集めてプールを作る。そのプールについて人手で関連性があるかどうかを判定する。つまり、プー

リングに基づく評価とは、評価すべきシステムや手法が有限の場合は、それらが出力した結果だけを考慮すれば、システム間の優劣は評価できるという考えに基づいている。

3.3 音声ドキュメント検索評価用テストコレクション

音声ドキュメント検索評価用コレクション構築では、テキスト検索評価用コレクションとは異なる点にも配慮しなければならない。まず、対象が音声ドキュメントとなる。しかし、音声データのみの提供では、そもそも正解判定作業自体が困難になることが予想される。また、音声認識を含めた実験環境を全て整えられる研究者しか利用できないものとなってしまい、テストコレクションとしての価値が下がってしまう。したがって、書き起こしも対象としては不可欠である。さらに、音声認識器の性能を考慮した検索の評価を行うために、音声認識器による標準的な書き起こしも必要であろう。認識結果の書き起こしも利用可能であれば、音声研究者以外にも、認識誤りのある文書に対する検索研究の機会を与えることになる。

以下、我々が今回構築している音声ドキュメント検索評価用テストコレクションの設計方針について述べる。

3.3.1 検索対象

広く利用可能な日本語音声データで、数百時間以上の規模を持ち、書き起こしも利用できるものとしては「日本語話し言葉コーパス (以下 CSJ と略す)」[10] しか存在していない。したがって、CSJ を対象とした。

今回のテストコレクションでは、CSJ のうち、「学会講演」と「模擬講演」を用いた。どちらも独話で自発発話である。両方の講演をあわせると 600 時間を越える。

表 1 利用した音声の種類と規模

音声の種類	話者数	講演数	データ量 (時間)
学会講演	838	1007	299.5
模擬講演	580	1699	324.1

学会講演は、自然科学系 (工学系も含む) 3 学会 621 講演、人文科学系 4 学会 187 講演である。講演の長さは、ほとんどが 12 分から 25 分であるが、中には、1 時間以上のものもある。模擬講演は、聴衆が 3 名から 5 名で、12 分程度のスピーチである。テーマは表 2 に示した 12 種類である。

表 2 模擬講演のテーマ分類

テーマ	講演数
(指定なし)	222
うれしかった出来事	137
悲しかった出来事	134
住んでいる町	134
よく知っていること・興味関心のあること	151
印象に残っていること	167
過去数年のニュース	152
無人島に持っていくもの 3 つ	101
何かのやり方・作り方	151
何かの歴史	100
いちばん大事なもの・人	100
21 世紀に残したいもの・残したくないもの	150

3.3.2 検索クエリ

検索対象文書 (CSJ) の性質、講演音声を対象とすることなどを考慮した上で、テストコレクションとしてどのような検索クエリが適切であるかを検討した。まずは 10 人程度の委員が各自 10 件程度のクエリを作成し、その結果を持ち寄って議論を行った。この手順を数回繰り返し行い、以下のような方針を固めていった。

- 内容を問う質問

音声文書特有の特徴を対象にした質問 (「言い間違いを笑って取り繕っている箇所をみつけたい」など) も提案されたが、検索に要する技術要素が分散してしまうことを避け、従来の内容検索技術で扱えないような質問は対象外とした。

- 講演の一部が答となるような質問

講演を検索の単位とすると、既存の文書検索テストコレクションに比べて対象文書コレクションのサイズが極端に小さくなり、検索の問題設定が簡単になり過ぎるという問題が生じる。例えば、年間約 10 万記事を有する新聞記事の数年分を対象とする既存の文書検索テストコレクションに対し、CSJ の講演数は約 2700 である。最低限の対象文書サイズを確保するため、講演よりも小さい単位 (例えばスライド 1 枚分の説明にあたる 1 分程度の長さのセグメントにすれば、全体で約 40000 の文書があるのと同等になる) を検索対象文書としなければならない。そのため、一般の文書検索における質問よりも検索対象の特定性が高い質問を作成する。例えば、固有名を問う質問応答的な質問が想定される。

- 検索対象の長さは、発話を単位として可変長とする

あらかじめ講演を一定の単位で分割して検索対象文書集合とする案も検討した。しかし、対象文書としてどのような単位・サイズが適当かを質問作成に先立って決めることは困難であったこと、検索対象の範囲は質問毎に異なることが予想されたこと、発話単位で可変長に正解判定しておけばテストコレクションの利用時の自由度が高まること (たとえば、単位固定の検索を評価することも可能)、などの理由から可変長とした。

- 1 分程度の発話区間が検索対象となる

前項の通り答の長さは不定ではあるが、答の粒度をある程度揃えるために基準を設けた。

- 10 件程度の適合区間が存在する

性能評価に用いることを考え、過度に特定性が高く適合区間が CSJ 全体で一カ所となるような質問や、一般的すぎてあらゆる講演に答が見つかる質問は避け、ある程度質問の質をそろえる。

- 特定分野にかたよること無く、CSJ 全体を網羅するような検索クエリセット

作成者の専門分野から、音声や言語に関する質問が多くなる傾向にあるので、それ以外の分野の講演を対象とすることを心掛ける。

3.3.3 適合性判定

次に、作成した検索クエリについて、CSJ から正解部分を特定する作業を行った。この作業は、(a) 文書全体ではなく文書の一部に対する適合性判定を行うこと、(b) 適合性判定と同時

に適合する区間 (連続した発話) も特定すること、の 2 点において、TREC や NTCIR などのテストコレクション構築プロジェクトで行われてきた文書単位の適合性判定と大きく異なる。

(a) に関して、文書を単位に検索を行う文書検索に対し文書の一部を検索対象とするタスクは、情報検索の分野では「パッセージ検索」と呼ばれている。しかし、パッセージ検索のテストコレクション構築、特に検索対象を可変長としたコレクション構築は、過去にあまり例がない。TREC においては、新規性検索タスク (Novelty Track, 2002 年〜2004 年) [11] が最初の試みで、文単位の適合性判定が行われた。NTCIR では、特許検索タスク [7] において、段落単位の適合性判定が行われている。

(b) に関して、検索クエリに対する答の部分だけを抽出するタスクとして「オープンドメイン質問応答」(以下、質問応答) がある。質問応答は 1999 年の TREC [12] からテストコレクション構築が行われてきたが、その対象は名称や数量を問う事実型質問 (factoid question) が中心であった。近年では、定義を問う質問 (定義型) や、理由を問う質問 (WHY 型) などの非事実型質問 (non-factoid question) を対象とした質問応答も研究が行われており、2006 年の NTCIR-6(QAC4) では任意の質問を対象とした質問応答のテストコレクション構築が試行されている。本稿で対象とする適合性判定は、答の部分そのものではなく発話単位で回答するという違いはあるものの、このような任意の質問に対する質問応答に近い。

以上のように、過去の事例から見ても今回の適合性判定タスクは多くの点で新しい試みであった。質問作成者が自ら作成した質問に対して正解判定を試み、その結果を持ち寄って議論を行い、以下のような方針で適合性判定を行うことにした。

- 一つの適合性判定対象は、連続した発話区間とする
- 判定は、適合 (Relevant)、部分適合 (Partially Relevant)、不適合 (Irrelevant) のいずれかとする

ただし、適合、部分適合に関する判定基準は、現段階では決めておらず、各判定作業者の判断に任せている。

- 根拠の存在は必須とする

例えば、「機械学習の手法が知りたい」という質問に対し、「決定木」だけが現れる発話は、それが機械学習手法かどうか文書を見ただけでは判定できないので、適合ではない。

- 根拠は、適合性判定対象の発話区間中に現れていなくても、同一講演の他の個所から読み取っても良い

互いに独立した文書を対象とした文書検索と異なり、講演の一部を適合性判定する場合、判定対象は他の検索対象 (同一講演のある部分) と依存関係を持ち得る。特に、質問に対する答そのものが現れる個所とその根拠に関する個所は講演中に離れて出現する場合がある。その際、適合区間に根拠を含めようとすると、(間に無関係の区間を挟んで) 長い区間を適合区間としなければならない。これを避けるため、根拠区間は適合区間とは別に指定することを許した。また、講演全体から漠然と根拠の存在が読み取れるが根拠個所は特定できない場合も、適合とした。

表 3 質問作成&適合性判定の概要

	質問あたり判定区間数	質問あたり異なり講演数	判定区間あたり平均発話数
適合	8.42	6.09	11.10
適合&部分適合	9.84	7.36	11.14

3.4 試作結果

3.4.1 クエリ作成と適合性判定

適合性判定作業を行うにあたり、CSJ の書き起しテキストを対象とした文書検索システムを、汎用検索エンジン GETA [13] を用いて構築し、作業のためのツールとして利用した。作業を円滑に行えるように、複数の検索条件設定を作業者の判断で自由に切り替えられるように実装した。まず、検索対象の単位として、講演単位、15 発話単位、30 発話単位 (15 発話のオーバーラップあり) の 3 種類を用意し、各クエリに適切な単位を選択できるようにした。また、索引の単位として、形態素、文字バイグラム、それらの組合せ、の 3 種類を用意した。また、検索結果の順位づけには、TF-IDF、SMART、AND、など、GETA に用意されている尺度から選択して利用した。作業者は、これらの種々の組合せて検索を行い、適合部分を見つけようと試みる。例えば、15 発話単位、文字バイグラム、AND 尺度の組合せで検索すれば、検索キーとして入力した文字列を文字通りに含む文書が引ける可能性が高い。

本原稿作成時点において、約 10 人の委員によって一人あたり平均 10 件程度の検索質問の作成を行い、各自で作成した質問に対して適合性判定作業を行った。作成した質問のうち、5 区間以上の「適合」区間を持つ質問は 33 件であった。33 件の統計量および事例を、それぞれ表 3、表 4 に示す。

3.4.2 問題点

テストコレクション構築を通して、いくつかの問題点が明らかになった。

第一に、講演を対象とした検索タスクは、従来の検索タスクに比べ難易度が高く、適合性判定も難しい。これは、次のような理由に因る。

- 学会講演は同じ研究分野の研究者を対象に行われるので、聴講者との知識の共有を前提に基本的な説明を省略する傾向がある。そのため、重要なキーワードが講演中に現れないことも多い。
- 講演は、音声発話だけでなくスライドを併用して行われる。そのため、重要な用語を発声せずにスライドを指示するだけで進める場合も少なくない。例えば、研究の目的となるようなキーワードが表題の部分だけで述べられていることも多い。(しかし、CSJ の書き起しでは講演題目が削除されている。) また、「見れば分かる」実験結果の数字はスライドに掲載されるだけで、それに関するコメント (結果の差を述べるなど) だけが発話されることも多い。特に、全国大会など時間の短い (12 分程度の) 講演では、発表時間を補うためスライド資料を効率的に使うので、この傾向が大きい。

第二に、作成した質問は、CSJ の中でも日本音響学会の講演が答になるようなクエリが多いという傾向があった。これは、

表 4　クエリと適合性判定結果の例

発話区間	判定	根拠
質問: 有名もしくは個人的評価の高い温泉地について知りたい．どこの地域・都道府県にあるのかが分かれば尚良い．		
(F (? う)) 母が松山の出身なので/(F え) 親戚が愛媛に多いことから/(F えと) 道後温泉に行ったりであるとか後車の免許を取ってからは	部分適合	(有名という記述はない)
阿蘇には (F ま) 温泉がたくさんありますので年に一回二回と/(F えーと) 色々遊びに行ってその時に (F あーの) 大分県の別府に寄ったり/後は由布院今若い女性に非常に人気があるっていうことなんですけれども/由布院に行ったり/で (F まー) 阿蘇の温泉を楽しんだり	適合	
久住の山の中に/法華院温泉というのはこれは要注意なんですが法華院温泉というのがあります	適合	(F えーっと) まずはその法華院温泉が/私の/温泉への目を開かせた温泉でした
質問: 講演音声の特徴について知りたい		
それから講演音声は読み上げ音声のモデルよりも (F えーと) 対話音声のモデルに近い発話スタイルに/なっていると/いうことも (F まー) 言えると思います	適合	
(F えー) その一方で (F ま) 講演音声というものの特徴を考えていきますと (F えー)(F ま) 話し言葉の冗長的な表現というものを多く含みまして	適合	
講演である為に丁寧な/<FV> 口調で話されておりますので丁寧語が/各所入っております	部分適合	
質問: 尊敬されている人やものについて知りたい		
多分父親本当に尊敬してる人は父親だけだと思うんですけど	適合	
合掌造りの里とか曲がり屋とか	適合	.../(F あー) 私達が見て/とても尊敬 (D (? し))/に値すると思います
(F ま)(A ケー; K) 理事長	適合	(F あ) 尊敬する二人のトップを横軸に/話をしてみたい/と/思います/(F あ) まず
質問: DP マッチングを用いた研究を探したい		
でこれは二つのモジュールからなってまして第一段階で統合モジュール/これにより (A ディーピー; D P) マッチングを行ないまして各システムが出す単語列というものの対応を取ります	適合	
(F と) 本日の発表で/< 雑音 >/(F えー)/主眼を置いているのはこの/(A ディーピー; D P) マッチングを (D おく) 連続 (A ディーピー; D P) を行なう際の距離尺度なんですがここを/色々と/< 雑音 >/変えてみようと考えています	適合	キーワードを/(F えー)(A ディーピー; D P) マッチング連続 (A ディーピー; D P) を行なった結果 (D き)/(F えー) 得られたパスというものはこのように (D も)
システムの方で音声区間抽出/(F え)(A ディーピー; D P) マッチングを行ない/整合経路の表示を行ないます/これが/その (A ディーピー; D P) マッチングをした時の結果の例/です	適合	
質問: "悪いマナーの例にはどのようなものが挙げられるか？"		
銀行の (F えー) 応対マナーについて考えていこうという時期がありまして/それに合わせて/(F え)(F ま) 良い応対と悪い応対というのを実際に自分で	不適合	
(F まー) 例えば車内の/携帯電話の使用っていうな (F まー)/散々/アナウンスしてるにもかかわらずやっぱりマナーは悪いと	適合	
先生は必ず (F あの) 携帯電話について/一言注意を入れるんですけれども/やはり/大体一授業九十分授業に一回ぐらいは鳴ってしまいますし/で後/たまに/自分の周りにいる学生全然知らない学生の/(F あのー) 私語がうるさくて聞きたい自分は聞きたい授業なんだけれども	適合	確かに講義中の学生のマナー/っていうのは/凄く悪いのではないかと思います
質問: "情報検索性能を評価するにはどのような方法があるか知りたい。"		
他方が (F あーの) 犠牲になるというような関係に基本的になりますでしたがって評価尺度の再現率と精度っていうのも普通は	適合	ですから (F (? ん))(F えーっと) いい検索システムというのは <FV> 両方の尺度ができるだけ高いと/いうことになります
通常の情報検索システムの/出力 (D2 と)/でよく使われる (F え) 平均精度/で (F えー) ランキングを評価する方法そして	適合	検索結果を評価する基準ですが/でこれに関しても二通り
でその日英検索の十一点平均適合率を/取ると	不適合	(根拠がない)

検索クエリ作成者の専門が音声分野にかたよっているためである。特定の分野にかたよること無く、答が CSJ 全体に分散したクエリ集合となるような工夫が必要である。

3.5 音声認識

実際に音声ドキュメントが大量にアーカイブされるようになったとき、それらを認識するために利用できる音響モデルや言語モデルはどのようなものになるだろうか? 少なくとも、それらを学習データに含むようなモデルでないことだけは確かだろう。

しかし、ある程度の認識率を達成するには、似たような学習データである必要がある。CSJ を対象とする場合には、講演データで学習することが望ましい。しかし、現状では、前述のように、そのような言語資源は CSJ しかない。また、検索対象とすることを考えると、CSJ の一部を学習データとして用いることも難しい。また、クロスバリデーションのようなことをするのも一つの条件で音声認識を行うだけで何ヶ月もかかってしまうため費用対効果の面で問題がある。

そこで、今回は、次の条件で実験を行った。言語モデルはコアデータを除いた全講演で一つだけ作成した。音響モデルは音声認識評価用テストセットを除く全講演で作成したものを用いた [14]。言語モデル・音響モデルとも、多くの講演が学習データに含まれてしまう条件 (closed) となる。音響モデルに関しては、比較のために、学会講演と模擬講演でそれぞれモデルを作成し、学会講演モデルで模擬講演を認識、また、模擬講演モデルで学会講演を認識する、といった open の実験を行った。言語モデルに関しては、コアに含まれる講演の性能を open なデータとして観察することで比較実験とする。どちらも closed の場合にも open なものは含まれる。

認識結果を表 5 に示す。音響モデルが open となることで、0.03 から 0.04 程度、さらに言語モデルが open となることで、0.05 から 0.07 程度認識率が悪化することがわかる。

学会講演の認識率の分布を図 1 に示す。平均が異なる以外の傾向はほぼ同じであり 0.65 から 0.95 くらいの範囲に分布することがわかる。これらの点を考慮し、最初のコレクションとしては少しでも認識率のよいものということで closed な結果を用いることとした。

表 5 CSJ の認識結果

	学会	コア (70 件)	模擬	コア (107 件)
closed	0.817	0.765	0.774	0.699
open	0.787	0.738	0.733	0.660

4. まとめと今後の課題

音声ドキュメント検索評価のためのテストコレクション第 0 版を構築した。今後は、認識結果に対する評価を行いテストコレクションとしての妥当性を評価する予定である。また、評価用途としての妥当性だけでなく、音声検索の応用分野に対する妥当性の検討も行っていく。構築したテストコレクションは、近日中に第 1 版として公開予定である。

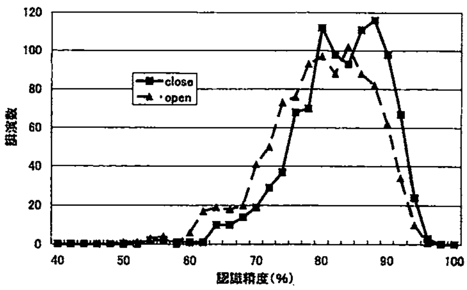

図 1 認識精度の分布 (学会講演)

文 献

[1] G. Brown and G. Yule: "Discourse Analysis", Cambridge University Press (1983).

[2] J. S. Garofolo, E. M. Voorhees, V. M. Stanford and K. S. Jones: "TREC-6 1997 spoken document retrieval track overview and results", Proceedings of the 6th Text Retrieval Conference, pp. 83–91 (1997).

[3] P. Over, T. Ianeva, W. Kraaij and A. Smeaton: "Trecvid 2005 - an overview", Proceedings of TRECVID 2005NIST, USA (2005).

[4] K. Oyama, M. Takaku, H. Ishikawa, A. Aizawa and H. Yamana: "Overview of the NTCIR-5 WEB navigational retrieval subtask 2", Proceedings of the Fifth NTCIR Workshop Meeting, pp. 423–442 (2005).

[5] T. Kitani, Y. Ogawa, T. Ishikawa, H. Kimoto, I. Keshi, J. Toyoura, T. Fukushima, K. Matsui, Y. Ueda, T. Sakai, T. Tokunaga, H. Tsuruoka, H. Nakawatase and T. Agata: "Lessons from BMIR-J2: A test collection for Japanese IR systems", pp. 345–346. Poster abstract.

[6] K. Kishida, K. hua Chen, S. Lee, K. Kuriyama, N. Kando, H.-H. Chen and S. H. Myaeng: "Overview of CLIR task at the fifth NTCIR workshop", Proceedings of the Fifth NTCIR Workshop Meeting, pp. 1–38 (2005).

[7] A. Fujii, M. Iwayama and N. Kando: "Overview of patent retrieval task at NTCIR-5", Proceedings of the Fifth NTCIR Workshop Meeting, pp. 269–277 (2005).

[8] T. Kato, J. Fukumoto and F. Masui: "An overview of NTCIR-5 QAC3", Proceedings of the Fifth NTCIR Workshop Meeting, pp. 361–372 (2005).

[9] T. Hirao, M. Okumura, T. Fukusima and H. Nanba: "Text summarization challenge 3 – text summarization evaluation at NTCIR workshop 4", Proceedings of the Fourth NTCIR Workshop (2004).

[10] 前川：“『日本語話し言葉コーパス』の概要”, 日本語科学, 15, pp. 111–133 (2004).

[11] I. Soboroff and D. Harman: "Novelty detection: The TREC experience", Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pp. 105–112 (2005).

[12] E. Voorhees and D. Tice: "The TREC-8 question answering track evaluation", Proceedings of the 8th Text Retrieval Conference, Gaithersburg, Maryland, pp. 83–106 (1999).

[13] "汎用連想検索エンジン GETA", http://geta.ex.nii.ac.jp.

[14] T. Kawahara, H. Nanjo, T. Shinozaki and S. Furui: "Benchmark test for speech recognition using the corpus of spontaneous japanese", ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, pp. 135–138 (2003).