Vol.2010-SLP-82 No.10
2010/7/23

音声ドキュメント検索の現状と課題

秋　葉　友　良[†1]

情報通信網の発展とデータ記録コストの低減により、テキストデータに加えてマルチメディアコンテンツの増大が加速している。音声ドキュメント検索は、音声データに含まれる言語情報を利用した検索技術であり、今後マルチメディアコンテンツの情報爆発時代に必要不可欠な技術になると考えられる。本稿では、音声ドキュメント検索の現状と課題について論じる。まず、音声ドキュメント検索の問題設定と、現在評価が行われている 2 つのタスクについて述べるとともに、比較的研究が進んでいるテキスト検索と比べた独自性を述べる。また、音声ドキュメント検索に必要な要素技術を整理し、関連研究を紹介する。最後に、手法の性能評価に不可欠なテストコレクションの現状について述べる。

Trends and Challenges for Spoken Document Retrieval

Tomoyosi Akiba[†1]

The growth of the internet and the decrease of the storage costs are resulting in the rapid increase of multimedia contents today. Spoken Document Retrieval (SDR) is a promising technology for enhancing the utility of such data. In this paper, the trends and the challenges for SDR are discussed. Firstly, the problem definition of SDR and the two current tasks for SDR are given, comparing with its counterpart, information retrieval for text. Then, the several component technologies for SDR are shown with their related works. Finally, the currently available test collections for SDR are presented, which are indispensable for evaluating SDR technologies.

†1 豊橋技術科学大学
Toyohashi University of Technology

1. はじめに

音声・画像・ビデオの記録・編集機器の拡大、およびインターネットをはじめとする情報通信網の発展により、誰でも気軽にコンテンツを作成・公開することが可能となり、マルチメディアコンテンツの増大が加速している。これらのコンテンツには、ファイル名やタイトル以外にはメタデータが付与されていないことが多く、従来のテキストベースの検索技術だけでは、目的のコンテンツにたどり着くことは困難である。一方、話し言葉を含むコンテンツの場合には、大語彙連続音声認識技術を利用することで言語情報を抽出し、テキスト検索技術を利用した検索が可能である。このような音声言語情報を対象とした検索技術は「音声ドキュメント検索 (Spoken Document Retrieval; SDR)」と呼ばれ、マルチメディアコンテンツの情報爆発時代に必要不可欠な技術になると考えられる。

本稿では、音声ドキュメント検索の現状と課題を整理し、それらに対する関連研究を紹介する。まず 2 節にて、音声ドキュメント検索の問題設定と、現在評価が行われている 2 つのタスクについて述べるとともに、比較的研究が進んでいるテキスト検索と比べた独自性を述べる。次に 3 節では、音声ドキュメント検索に必要な要素技術を整理し、関連研究を紹介する。4 節では、手法の性能評価に不可欠なテストコレクションの現状について述べる。

2. 音声ドキュメント検索の問題設定と課題

2.1 音声ドキュメント検索の問題設定

音声ドキュメント検索とは、音声データと検索クエリが入力として与えられ、検索クエリに適合するような音声データの部分を特定する問題である。ここで、検索クエリはテキストで与えられると仮定する。音声クエリからの音声ドキュメント検索[1)] も興味深い問題ではあるが、本稿では扱わない。[⋆1]

音声ドキュメント検索は、音声認識、特にワード・スポッティング、と深く関係している。音声認識は、音声データが入力として与えられ、入力の全部または一部に対応するテキスト (書き起し) を求める問題である。音声ドキュメント検索がテキスト (検索クエリ) を入力として対応する音声データ区間を出力するのに対し、音声認識は音声データ区間を入力として対応するテキストを出力する。両者は、入力と出力が逆転してはいるが、基本的に同じ問題を解いていると考えられる。特に、問題を解くためのリソース (計算コスト、空間コスト、

⋆1 音声クエリからの検索 (検索対象は音声データとは限らない) をボイスサーチ、または単に音声検索、などと呼ぶ。

利用可能なデータ) が無限に使える状況では、音声認識と音声ドキュメント検索のための手法に本質的な違いはない。したがって、両者の差異は、利用可能なリソースの差異にあり、その制約の元で適切な手法が検討されることになる。

音声ドキュメント検索における利用可能なリソースの制約は以下の通りである。

- 対象の音声データのサイズが大きい (数十時間〜数千時間以上)。
- 対象の音声データは、検索処理に先だって入手できる (前処理できる)。
- 音声データの前処理に必要なコスト (時間・空間コスト) を低く押さえることが要求される。
- 検索クエリが入力されてから、効率良く (時間・空間コスト)、特に短時間 (1 秒以内〜数分) で出力を返すことが要求される。

したがって、音声ドキュメント処理を計算機処理の観点から見た場合、大量の音声データを、後の高速な検索処理に備えて、いかに効率良く前処理するかが主な課題となる。

この問題に対する現在の典型的な解法は、(1) 音声データに対する音声認識、(2) 認識結果に対する索引付け、(3) テキスト検索手法の適用、の組合わせである。まず (1) で、音声データに対して音声認識を使ってテキストに変換 (量子化) しておくことで、後の検索時の効率化とそれに必要な記憶容量 (空間コスト) を低減する。さらに (2) で、より高速な検索に備えたデータ構造を、低コスト (特に、空間コスト) で構築しておく。最後に (3) で、前処理で構築したデータ構造を利用して短時間で結果を出力する。

本稿では、これらのうち (2) と (3) の処理について論じることとし、(1) の音声認識の問題については立ち入らないこととする。しかし、実際は (1) の音声認識は、音声ドキュメント検索にとって無視できない処理であることは明らかである。実際、音声認識の性能は後段の検索結果と強い相関があることが報告されている。また、逐次増加する傾向のある大量の対象音声データを前処理するには、高速な音声認識が必要になる。今後、認識対象音声データに適応するためのリソースが十分に用意できない場面や、認識処理自体を高速に行うことが要求される応用場面が考えられ、音声ドキュメント検索の問題に特化した音声認識技術は重要な研究課題になると思われる。

2.2 音声ドキュメント検索のタスク

現在、音声ドキュメントを対象とした検索は、2 種類のタスクが設定され、研究・評価が行われている。

一つは、Spoken Term Detection(以下、STD)、日本語では音声検索語検出、音声キーワード検索などと呼ばれる。後述する評価型会議の TREC SDR[2] Track では、Known Iterm Retrieval と呼ばれた。STD は、単語あるいは数単語の列をクエリとして与え、音声ドキュメント中からクエリがそのまま現れる位置を特定するタスクである。2006 年に NIST が STD をタスクに設定したこと[3]、およびタスクの評価基準が明確であることから、近年研究が活発化している。STD は、次の SDR タスクの前段となるタスクであると考えることができる。

もう一つのタスクは、テキスト検索における内容検索に相当する音声内容検索である。このタスクを狭義で Spoken Document Retrieval(SDR) と呼ぶことも多い (以降、こちらのタスクを SDR と呼ぶ)。また、TREC SDR Track では、Ad-hoc Retrieval[⋆1] と呼ばれた。SDR は、検索者の知りたい内容を表現した文やキーワードリストなどの比較的長いクエリを与え、その内容を含む文書を見つけるタスクである。正解は、クエリと文書から人手で判定される。必ずしも検索クエリ中の表現 (語) が含まれているとは限らない。

STD は、検索者が検索の対象 (用語) を既に知っている状況 (ナビゲーショナルな質問) を想定したタスクである。一方、SDR は、人の曖昧な情報要求 (インフォメーショナルな質問) から関連情報を見つけるタスクである。

2.3 テキストを対象とした検索との対応

音声ドキュメント検索の第一近似は、音声認識を用いて音声データをテキストに自動書き起ししておき、これに対して既存のテキスト検索手法を適用することである。しかし、このナイーブな手法は、音声認識で生じる認識誤りを扱うことができない。既存の文書検索手法は、検索対象のテキストに誤りが含まれていることを仮定していないからである。特に、音声認識の認識語彙外語 (OOV) は自動書き起し結果に現れることがないため、検索することができない。したがって、音声ドキュメントを対象とした検索では、フロントエンドで導入されるノイズを、検索手法でどのように扱うかが課題になる。

一般にテキストに対する検索と言えば「内容検索」[4] を指し、これは音声ドキュメント検索での SDR タスクに対応する。テキストを対象とした内容検索の手法は、誤りのない線状のテキストを対象としてきた。音声認識結果を対象とするためには、誤りを含むテキストを対象とするとともに、ラティスなどで表された複数候補を扱う手法が必要となる。

一方、テキストを対象とした、STD に対応するタスクは「文字列照合」[5] と呼ばれる。特に、検索語とのずれを許した文字列照合は「近似文字列照合」[6] と呼ばれる。テキストを対象とした近似文字列照合では、文字単位の一致・不一致をベースとした離散的な編集距離を

⋆1 情報検索分野で、検索クエリがその場で与えられる検索タスクを指す。

指標に検索語と近い文字列出現個所を見つける。音声認識結果を対象とする場合、音響的な類似度や認識尤度を考慮に入れたより連続的な距離尺度を考慮に入れた手法が必要となる。また、線状のテキストに対して、ラティスなどの複数候補表現を検索対象とするように拡張が必要である。

3. 音声ドキュメント検索の関連研究

3.1 認識・検索の単位

検索クエリは検索結果を絞るように選択されるため、固有名詞などの特定性の高い語が含まれることが多く、認識語彙外語 (OOV) に成りやすい。したがって、OOV 対策は音声ドキュメント検索の主要な課題の一つである。

OOV の問題を直接避ける方法は、単語より小さな認識単位を設定して認識語彙を閉じることである。そのような単位を総称して、サブワード (subword) と呼ぶ。サブワードを単位とした音声認識を行い、サブワードを単位としたテキスト検索を行えば、少なくとも OOV の問題は解決する。問題は、どのような単位をサブワードとして使うかである。サブワードの選択は、検索対象の言語に強く依存する。例えば、語形変化の激しい言語の場合には、認識精度を上げるためにも検索性能を上げるためにも、サブワードの導入は必須である[7]。

サブワードとして、書記の単位を用いるか、発音の単位を用いるかの 2 つの選択が考えられる。書記単位を使う場合は、検索クエリと音声ドキュメントの表現が一致するため、直接検索が可能である。書記単位の候補としては、単語、形態素 (morpheme)、書記素 (grapheme)、文字 (特に中国語における漢字) などが挙げられる。また、テキストを自動処理により分割した単位 (morph)[7] を使うことも可能である。これらの単位を使う場合、検索の観点からは検索クエリとドキュメント中での単位が一致していることが要求されるため、曖昧なく定義できる単位であることが望ましい。例えば、日本語の場合、単語の境界に明確な区切りが無いため、自動処理 (形態素解析) によって単語区切りを見つけることになるが、検索性能はその精度に左右されることになる。

発音の単位の選択肢としては、音節 (sylable)、音素 (phoneme) などの単位と共に、その音響コンテキストを含めるかどうかの選択 (例えば、triphone など) が考えられる。また、より短い単位である半音素や音素片 (SPS) を使う方法も提案されており、時間的に精緻な単位を使うことで検索性能が向上することが報告されている[8]。発音単位をサブワードとした認識結果を検索する場合は、テキスト (書記単位列) として表現された検索クエリを発音単位に変換する必要があり、発音の多様性の大きい言語 (英語など) ではこの変換性能が重要になる。一方、音声認識とは別に用意した知識源から学習した、発音への変換モデル[9] を陽に導入できるという利点もある。

サブワードの導入は OOV 対策として有効な手段ではある。しかし一般に、認識語彙内語 (IV) の場合は、単語を単位とした方が認識率は高く、したがって検索性能も向上する。検索語が IV か OOV かは認識辞書から判定できるので、IV の場合は単語認識結果、OOV の場合はサブワード認識結果、というように両者を併用する手法[10),11)] も提案され効果が示されている。また、単語とサブワード[12),13)]、あるいは複数のサブワード[14)–16)] を使った検索結果を組合せることで信頼性を向上させる方法も提案されている。

また、認識の単位と検索の単位は必ずしも一致しなくてもよい。単語認識の結果をサブワード列に展開することで OOV の検索性能を向上させる方法も考えられる[11)]。

3.2 複数代替候補の表現

検索対象音声ドキュメントの認識誤りの影響を軽減するために、音声認識結果の複数代替候補を利用することが考えられる。音声認識結果の複数代替候補の表現方法としては、N-best リスト、単語 (あるいはサブワード) ラティス、Confusion Network[17] などが知られている。以降では、単語を単位とした表現を仮定するが、単語の代わりにサブワードを利用することも可能である。

ラティスなどで表現した複数候補表現を検索対象とする場合の問題点は、まず候補が増えることによる空間コストの増大にある。検索時に利用する索引のための記憶容量は、なるべく小さく押さえることが望まれる。また、フレーズ検索 (単語列の検索) に利用する単語の隣接情報を効率良く利用できることが要求される。線状のテキストでは、単語の位置情報だけで、すなわち位置を表す番号が連続しているかどうかを調べることで、単語間の隣接関係が判定できる。一方、ラティス上にアークとして表現された単語の場合、単語の隣接はラティス上でのアークの隣接関係を調べる必要があり、計算コストが高い。特にサブワード単位を索引の単位にする場合、検索クエリは必ずサブワード列となるため、フレーズ検索を効率良く実行する必要がある。Saraclar ら[11)] は、ラティスから直接、単語の隣接関係を保持した索引を構築する手法を提案している。しかし、索引のサイズがテキストと比べ大幅に増大するという問題点がある。

以上の観点から、認識結果のラティスの圧縮手法が提案されている。それらの手法の多くは、元のラティスを非可逆に圧縮する手法である。非可逆圧縮により、元のラティスでは現れていない候補 (パス) も新たなパスとして表現されることで、候補数が増大し検索の再現率を向上させる効果も得られる。

Confusion Network (CN)[17] は、ラティスの時間情報と音響的類似度を元に、同一候補あるいは対立する候補をクラスタリングし、ソーセージ型のラティスを構築する手法である。本来は音声認識率の向上のために利用されてきたが、STD への適用例は多い[7),18)]。ソーセージ型ラティスの利点は、線状のテキストと同様に、単語の隣接関係を位置情報だけで判定できることにある。しかし、Confusion Network の場合は、アークに単語が無いイプシロン遷移が含まれるので、単純に位置番号の連続だけでは隣接関係は判定できず、検索時に高価な処理が必要となる。

Position Specific Posterior Lattice (PSPL)[19),20)] は、ラティスの先頭からの位置情報だけを使って圧縮する手法である。すなわち、各単語についてラティス先頭からのパスの長さ (位置) を求め、その位置にその単語が現れたとして、ソーセージ型ラティスにクラスタリングする。その作り方から、イプシロン遷移は現れないので、テキストの場合と同様の単語隣接関係の判定が可能である。しかし、先頭からのパスの数に応じて一つの単語が複数の位置にコピーされることから、長い文ではラティスサイズが大きくなり、実用上は短い発話単位を見つけて適用することが要求される。

Time-based Merging for Indexing (TMI)[21] は、ラティスの時間情報を使って圧縮を行う手法である。ラティスのアークをマージする手法と、ノードをマージする手法が提案されており、それぞれ TMI-arc、TMI-node と呼ぶ。TMI では、必ずしもソーセージ型のラティスにはならないが、各単語は開始時刻と終了時刻の情報を使ってクラスタリングされるので、時間情報だけで隣接関係の判定が可能になる。

Time-Anchored Lattice Expansion (TALE)[22] は、TMI をベースにソーセージ型のラティスを構築する手法である。ラティスの位置情報を使って圧縮を行うが、先頭からの位置を用いる PSPL と異なり、各単語について前後 N 単語の隣接情報を利用する。イプシロン遷移は現れないので、テキストの場合と同様の単語隣接関係の判定が可能である。

3.3 索引付けと照合

3.3.1 Exact Matching

テキストを対象とした検索の場合は、検索クエリと検索対象文書の間での文字列の完全一致 (exact matching) を手がかり検索を行うのが基本である。その際、効率的に検索を行うために、検索対象文書に対して索引付けが行われる。代表的な索引付け手法として、転置ファイル (inverted file) がある。転置ファイルは、検索対象文書に現れる単語の辞書を作り、各単語エントリに出現文書および文書中の位置を記録したものである。転置ファイルは辞書から洩れた語は検索できない。一方、文書中の任意の部分文字列に対する索引付け手法として、サフィックスアレイ[23] が知られている。

音声ドキュメントを対象とした検索の場合も、完全一致の検索を行うことが多い。ノイズを含む文書の場合でも、3.2 節で述べた複数認識候補および非可逆圧縮による候補の増加により、認識誤りにある程度対応できるからである。完全一致の検索であれば、効率化手法として転置ファイルが利用できる。サブワードなどの短い単位を利用する場合は、検索候補数の増加による検索効率の悪化を避けるため、サブワード n-gram を索引にすることができるが、n-gram の長さが増えると索引の辞書サイズが増加するため、時間と空間効率のバランスをとる必要がある。また、3.2 節で述べたように、フレーズ検索を効率的に行うためにはラティスの表現手法を考慮する必要がある。一方、サフィックスアレイによる索引は、検索対象文書中の共通部分列をツリー状に圧縮する手法であるため、共通部分をまとめることが難しいラティスなどの複数候補表現には適用が難しい。

3.3.2 Soft Matching

より積極的にノイズに対応するためには、検索クエリと検索対象文書の間のずれを許容した一致判定を行う必要がある。これらの手法を音声ドキュメント検索では soft matching あるいは fuzzy search などと呼ぶ。一方、テキストを対象とした検索において、検索クエリと検索対象文書の間で誤りを許した一致を求める問題は近似文字列照合 (approximate string matching) と呼ばれる。以下では、近似文字列照合の手法を紹介し、音声ドキュメント処理での研究との関連について述べる。

近似文字列照合は、テキストがその場で与えられることを仮定してテキストを前処理することなしに照合を行うオンライン手法[6] と、あらかじめテキストが与えられていることを仮定してテキストを前処理して索引付けを行うオフライン手法[24] の 2 つに分類される。音声ドキュメント検索における soft matching 手法は、オンライン手法である連続 DP マッチングが使われることが多かった。しかし今後、検索対象の音声ドキュメントのサイズが大規模 (数百〜数千時間) になると、対象文書を一通り調べる必要のあるオンライン手法は現実的ではなく、索引を使ったオフライン手法が必須になる。

オフライン近似文字列照合の索引付け手法は、(1)n-gram 索引、または n-sample 索引を用いる手法、(2) サフィックスアレイによる索引を用いる手法、(3) 距離空間上の索引を用いる手法、に分類される[24]。

(1) の n-gram 索引を用いる手法は、検索対象のテキストを文字 n-gram を単位に転置ファイルで索引付けしておく。検索時には、検索クエリを複数の n-gram 区間に分割し、各 n-gram で完全一致したテキスト位置を見つけ、その前後区間をオンライン近似文字列手法

で照合する。ここで、誤りとして許容する距離を d とすると、検索クエリを $d+1$ 個の区間に分割した場合、少なくとも 1 つの区間は誤り無しで完全一致することを手がかりとしている。音声ドキュメントの検索の場合は、距離が連続的になることに注意が必要であるが、基本的には同じ原理で適用が可能である[25]。しかし、検索クエリが短い場合など、クエリの分割の際に誤りを含まない区間が作れない場合は、検索洩れが生じてしまう。岩見ら[26]は、音節を脱落させた索引作成と検索時の検索クエリへの音節脱落操作により、検索洩れを押さえる手法を提案している。

(2) のサフィックスアレイを索引に使う手法は、検索対象文書の共通部分文字列をツリー状に圧縮したデータ構造 (サフィックスツリー) に対し、DP マッチングで近似文字列照合を行う手法である。文書の複数個所への探索が共有されることで、高速な照合が可能になる。音声ドキュメントの検索に適用する場合は、共通部分をまとめることが難しいラティスなどの複数候補表現には適用が難しい。しかし、複数候補表現を用いる代わりに、1-best 候補に対する soft matching だけで音声ドキュメントのノイズに対応した検索を行うことも考えられる。Katsurada ら[27]は、本手法を音声ドキュメント検索に適用し、大規模音声ドキュメントに対する高速な検索システムを構築している。

(3) の距離空間上の索引を用いる手法は、テキストに限らず距離が定義された空間上のオブジェクト一般に適用できる索引付け手法である。検索対象となるオブジェクト集合に対して、検索クエリオブジェクトを与え、最も距離の小さいオブジェクトを検索結果として求める。その際、検索クエリとすべての対象オブジェクトとの間の距離計算を、索引付けによって効率化する。典型的な手法は、検索対象オブジェクトから少数をピボットとして抽出し、ピボットと他のオブジェクト間の距離を予め計算しておくというものである。検索時には、検索クエリとピボットとの距離だけを計算し、距離の三角不等式関係により近似的に距離計算を行う。本手法を音声ドキュメントの検索に適用した例は少ないが、テキストと異なり連続的な距離を扱う必要のある音声ドキュメント検索との相性は良いと考えられる。金子ら[28]は、検索クエリを分割した部分距離空間上に索引を作る STD 手法を提案している。この手法は、検索時に距離のしきい値を必要とせず、尤らしい順番に検索結果を出力するという特長がある。

音声ドキュメントに対して近似文字列照合を行う場合、サブワード間の距離の決め方も重要である。認識尤度[29]、サブワードの音響モデル間の距離 (KL-divergence[30] や Bhattacharyya 距離[31])、音素弁別特長[32] 間のハミング距離などが利用されている。

3.4 検索モデル

3.4.1 文書のランキング

SDR タスクでは、検索クエリの単語が含まれるなどとして候補となった文書集合に対し、順序付けを行い出力する必要がある。これには、テキストを対象とした検索で広く利用されているベクトル空間モデル[4] が利用できる。音声ドキュメントを対象とする場合、単語の重み付けに利用する TF(Term Frequency) や IDF(Inverse Document Frequency) は、ラティスなどの複数候補表現から計算する必要がある。ラティスから求めた TF の期待値を使う手法[33),34] などが提案されている。

また、情報検索分野で新しく提案された言語モデルに基づく検索モデル[35] は、確率的な枠組みを基盤としているため、ラティスなど不確性を表す音声ドキュメントとの相性が良いと考えられ、近年音声ドキュメント検索での利用が進んでいる[36),37]。

3.4.2 質問拡張・文書拡張

SDR タスクは、検索クエリと内容が一致した文書を見つけるタスクであるので、検索結果には必ずしも検索クエリ中の表現 (語) が含まれているとは限らない。検索クエリと文書の間の表現のギャップを埋める手法を、処理対象に応じて質問拡張または文書拡張と呼ぶ。

Latent Semantic Indexing (LSI)[38] は、質問拡張・文書拡張の一手法である。LSI では、検索クエリと文書の単語集合 (bag of words) を、それらの概念を表す潜在意味空間上へマッピングし比較を行う。テキストを対象とした検索では様々な LSI の拡張手法が提案されているが、これらは音声ドキュメントを対象とした検索でも利用可能である。Hu ら[39] は、SDR タスクに対して、検索クエリと文書の表現の違いを扱うために LSI の拡張手法である NMF を適用している。また、ノイズのある音声ドキュメントを検索対象とする場合、文書拡張はそもそも文書中の語の出現が不確実であることを考慮するのが望ましい。Chen[36] は、言語モデルに基づく検索モデルと Probabilistic LSI(PLSI)[40] を組み合わせた確率的文書拡張法を SDR に適用している。

質問拡張・文書拡張の手法として、検索クエリや文書に直接関連語を付け加える方法も考えられる。検索対象の音声ドキュメントを拡張する手法として、Web を使った文書拡張法が提案されている[41),42]。これらの手法では、検索対象文書を検索クエリとして Web 検索を行い、検索結果に含まれる単語で文書拡張を行う。

一方、質問拡張・文書拡張は、認識誤りや OOV の問題から生じる検索クエリと文書の間の表現のギャップを埋める手法としても有効である。この観点から見ると、3.2 節で述べたラティスによる複数候補表現は、文書拡張の一種であると考えられる。また、秋葉ら[43] は、

このギャップを直接埋め合わせるために、認識単語から正解単語への翻訳モデルを用いる手法を提案している。

3.5 多段階の検出

音声ドキュメント検索では、検索クエリが入力されてから短時間で検索結果を出力する必要があるため、音声認識のようにオンラインで全音声データとの詳細なマッチングを取ることはできない。しかし、見込みのある限られた区間だけマッチングを行い、検出の精度を上げることは可能である。このアイデアに基づき、処理の軽い高速な 1 段目の検索の後に、見込みのある区間に対して 2 段階目の照合を行って、実際に検索結果として出力するかどうかを決定する、2 段階の検出手法が提案されている[34]。また、中間段階で比較的高速な照合段階を挟む、3 段階の検出手法も提案されている[44]。

多段階検出における最後の段階を Decision Maker と呼び、検出が十分に信頼できるかどうかを Confidence Measure(信頼度) を使って判定する。信頼度には、ラティスベースの事後確率を使うのが一般的だが、多層パーセプトロンから直接計算した事後確率[14]、検出単語に依存した信頼度[45] なども提案されている。

4. 音声ドキュメント検索の評価

情報検索の分野で、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データセットをテストコレクションと言う。テキストを対象とした情報検索の分野では、TREC や NTCIR などの評価型ワークショップでの活動を中心に、多くのテストコレクションが積極的に構築されてきた。音声ドキュメント検索においても、性能評価のためには、テストコレクションが必要である。特に SDR タスクでは、正解を人手で判定する必要があるため、テストコレクションが不可欠である。

STD タスクでは、TREC SDR Track[2] の初年度 (1996 年, TREC-6) において、Known Iterm Retrieval のテストコレクションが構築された。また、2006 年には、米国規格協会 (NIST) が STD を新たなタスクに設定し[3]、共通の評価基盤が設定され、これを契機に STD 研究が活性化した。対象データは 3 時間程度の、ニュース音声、電話での会話、会議音声である。日本においては、情報処理学会音声言語処理研究会 (SIG-SLP) の音声ドキュメント処理ワーキンググループにおいて、STD のテストコレクション構築が進められており、2010 年には 2 回目の中間報告が行われた[46]。対象データは、日本語話し言葉コーパス (CSJ)[47] の学会講演と模擬講演で、セットによって約 44 時間または約 623 時間の長さである。

SDR タスクでは、やはり TREC SDR Track が契機となった。1997 年の TREC-7 から 1999 年の TREC-9 において、ニュース音声を対象としたテストコレクションが構築された。最終的には、557 時間、約 2 万文書を対象としたテストコレクションが構築された。TREC SDR Track は成功裏にタスクを終了したため、その後組織的に大規模なテストコレクションが作られることは少なかった。一方、日本では、音声ドキュメント処理ワーキンググループにおいて、日本語を対象とした SDR テストコレクションが構築され公開が行われている[48]。対象データは、日本語話し言葉コーパス (CSJ) の学会講演と模擬講演の約 623 時間で、39 の検索質問が設定されている。

5. ま と め

本稿では、音声ドキュメント検索に関する技術課題と関連研究について述べた。本稿で述べた音声ドキュメント検索の 2 つのタスクのうち、現在比較的活発に研究が行われているのは、検索者が検索の対象 (用語) を既に知っている状況を想定した STD タスクである。一方、人が検索を行なう実際の場面では、知りたい事項に対して漠然としたイメージしか持っていないこともあり、その場合は具体的な単語を想起できない。また、十分に記述内容を推敲するテキストの場合と異なり、音声ドキュメントは自発性の高い発話音声から構成される。そのため、検索者が想定するようなキーワードは必ずしも発話中に現れないと考えられる。このような状況を扱う SDR タスクは、評価のためのテストコレクション構築のコストが高く、研究が困難であった。現在、テストコレクションの整備も進みつつあることから、今後の SDR タスクの展開に期待したい。

参 考 文 献

1) Chia, T.K., Sim, K.C., Li, H. and Ng, H.T.: A Lattice-Based Approach to Query-by-Example Spoken Document Retrieval, *Proceedings of Annual International ACM SIGIR Conference on Research and development in information retrieval*, pp.363–370 (2008).
2) Garofolo, J.S., Auzanne, C. G.P. and Voorhees, E.M.: The TREC Spoken Document Retrieval Track: A Success Story, *Proceedings of TREC-9*, pp.107–129 (1999).
3) National Institute of Standards and Technology: Spoken Term Detection Evaluation Portal. `http://www.nist.gov/speech/tests/std/`.
4) Baeza-Yates, R. and Ribeiro-Nato, B.: *Modern Information Retrieval*, Addison-Wesley (1999).
5) Navarro, G. and Raffinot, M.: *Flexible Pattern Matching in String*, Cambridge University Press (2002).

6) Navarro, G.: A Guided Tour to Approximate String Matching, *ACM Computing Surverys*, Vol.33, No.1, pp.31–88 (2001).
7) Turunen, V.T. and Kurimo, M.: Indexing Confusion Networks for Morph-based Spoken Document Retrieval, *Proceedings of Annual International ACM SIGIR Conference on Research and development in information retrieval*, pp.631–638 (2007).
8) 岩田耕平，伊藤慶明，小嶋和徳，石亀昌明，田中和世，李　時旭：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，情報処理学会論文誌，Vol.48, No.5, pp.1990–2000 (2007).
9) Bisani, M. and Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion, *Speech Communication*, Vol.50, No.5, pp.434–451 (2008).
10) 西崎博光，中川聖一：音声認識誤りと未知語に頑健な音声文書検索手法，電子情報通信学会論文誌，Vol.J86-D-II, No.10, pp.1369–1381 (2003).
11) Saraclar, M. and Sproat, R.: Lattice-Based Search for Spoken Utterance Retrieval, *Proceedings of Human Language Technology Conference* (2004).
12) Yu, P. and Seide, F.: A Hybrid Word / Phoneme-based Approach for Improved Vocabulary-Independent Search in Spontaneous Speech, *Proceedings of International Conference on Spoken Language Processing* (2004).
13) Iwata, K., Shinoda, K. and Furui, S.: Robust Spoken Term Detection Using Combination of Phone-Based and Word-Based Recognition, *Proceedings of International Conference on Speech Communication and Technology*, pp.2195–2198 (2008).
14) Tejedor, J., Wang, D., King, S., Frankel, J. and Colas, J.: A Posterior Probability-Based System Hybridisation and Combination for Spoken Term Detection, *Proceedings of International Conference on Speech Communication and Technology*, pp.2131–2134 (2009).
15) 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ー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，Vol.50, No.2, pp.524–533 (2009).
16) 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，情報処理学会研究報告，Vol.2009-SLP-79, No.19 (2009).
17) Mangu, L., Brill, E. and Stolcke, A.: Finding Consensus in Speech Recognition: Word Error minimization and Other Applications of Confusion Networks, *Computer, Speech and Language*, Vol.14, No.4, pp.373–400 (2000).
18) Hori, T., Hetherington, L., Hazen, T.J. and Glass, J.R.: Open-Vocabulary Spoken Utterance Retrieval using Confusion Networks, *Proceedings of International Conference on Acoustic, Speech, and Signal Processing*, pp.73–76 (2007).
19) Chelba, C. and Acero, A.: Position Specific Posterior Lattices for Indexing Speech, *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pp.443–450 (2005).
20) Pan, Y., Chang, H., Chen, B. and Lee, L.: Subword-based Position Specific Posterior Lattices (S-PSPL) for Indexing Speech Information, *Proceedings of International Conference on Speech Communication and Technology*, pp.318–321 (2007).
21) Zhou, Z.-Y., Yu, P., Chelba, C. and Seide, F.: Towards Spoken-Document Retrieval for the Internet: Lattice Indexing For Large-Scale Web-Search Architectures, *Proceedings of Human Language Technology Conference*, pp.415–422 (2006).
22) Yu, P., Shi, Y. and Seide, F.: Approximate Word-Lattice Indexing with Text Indexers: Time-Anchored Lattice Expansion, *Proceedings of International Conference on Acoustic, Speech, and Signal Processing*, pp.5248–5251 (2008).
23) Manber, U. and Myers, G.: Suffix arrays: a new method for on-line string searches, *SIAM Journal on Computing*, Vol.22, No.5, pp.935–948 (1993).
24) Navarro, G., Baeza-Yates, R., Sutinen, E. and Tarhio, J.: Indexing Methods for Approximate String Matching, *IEEE Data Engineering Bulletin*, Vol.24, No.4, pp. 12–27 (2000).
25) Mamou, J., Mass, Y., Ramabhadran, B. and Sznajder, B.: Combination of Multile Speech Transcription Methods for Vocbulary Independent Search, *Proceedings of Annual International ACM SIGIR Conference on Research and development in information retrieval* (2008).
26) 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，日本音響学会春季研究発表会研究論文集，pp.203–206 (2010).
27) Katsurada, K., Teshima, S. and Nitta, T.: Fast Keyword Detection Using Suffix Array, *Proceedings of International Conference on Speech Communication and Technology* (2009).
28) 金子泰輔，秋葉友良：ハフ変換に基づく音声ドキュメントの高速検索語検出法，日本音響学会春季研究発表会研究論文集，pp.113–116 (2010).
29) Wallace, R., Vogt, R. and Sridharan, S.: A Phonetic Search Approach to the 2006 NIST Spoken Term Detection Evaluation, *Proceedings of International Conference on Speech Communication and Technology*, pp.2385–2388 (2007).
30) Liu, P., Soong, F.K. and Zhou, J.-L.: Divergence-based Similarity Measure for Spoken Document Retrieval, *Proceedings of International Conference on Acoustic, Speech, and Signal Processing*, Vol.IV, pp.89–92 (2007).
31) 山本一公，中川聖一：発話スタイルによる話速・音韻間距離・ゆう度の違いと音声認識性能の関係，電子情報通信学会論文誌，Vol.J83-D-II(11), pp.2438–2447 (2000).
32) Fukuda, T. and Nitta, T.: Orthogonalized Distinctive Phonetic Feature Extraction for Noise-Robust Automatic Speech Recognition, *IEICE transactions on information and systems*, Vol.E87-D(5), pp.1110–1118 (2004).
33) Allauzen, C., Mohri, M. and Saraclar, M.: General Indexation of Weighted Automata – Application to Spoken Utterance Retrieval, *Proceedings of the Workshop*

on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004 (2004).
34) Yu, P. and Seide, F.: Fast Two-Stage Vocabulary-Independent Search in Spontaneous Speech, *Proceedings of International Conference on Acoustic, Speech, and Signal Processing*, Vol.I, pp.481–484 (2005).
35) Croft, W.B. and Lafferty, J.(eds.): *Language Modeling for Information Retrieval*, Kluwer Academic Publishers (2003).
36) Chen, B.: Latent Topic Modeling of Word Co-Occurrence Information for Spoken Document Retrieval, *Proceedings of International Conference on Acoustic, Speech, and Signal Processing*, pp.3961–3964 (2009).
37) Honda, K. and Akiba, T.: Language Modeling Approach for Retrieving Passages in Lecture Audio Data, *Proceedings of International Conference on Language Resources and Evaluation*, pp.1525–1530 (2010).
38) Deerwester, S.C., Dumais, S.T., Furnas, G., Landauer, T.K. and Harshman, R.A.: Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, Vol.41(6), pp.391–407 (1990).
39) Hu, X., Kashioka, H., Isotani, R. and Nakamura, S.: Japanese Spontaneous Spoken Document Retrieval Using NMF-based Topic Models, *Proceedings of the Fifthe Asia Information Retrieval Symposium*, pp.149–156 (2009).
40) Hofmann, T.: Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning*, Vol.42, No.1, pp.177–196 (2001).
41) 杉本樹世貴，西崎博光，関口芳廣：音声ドキュメント検索における Web ページを用いたドキュメント拡張の効果，情報処理学会研究報告，Vol.2009-SLP-76, No.11 (2009).
42) 宇野　有，伊藤彰則，伊藤　仁，牧野正三：音声ドキュメント検索のための WWW を用いたインデクス改善，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，No.9 (2010).
43) 秋葉友良，横田悠右：認識候補から正解テキストへの翻訳に基づく講演音声ドキュメントのアドホック検索，情報処理学会論文誌，Vol.50, No.2, pp.514–523 (2009).
44) 神田直之，住吉貴志，戸上真人，大淵康成：任意語彙音声発話検索のための多段階リスコアリング手法の性能評価，第２回音声ドキュメント処理ワークショップ講演論文集，pp.73–78 (2008).
45) Wang, D., King, S., Frankel, J. and Bell, P.: Term-Dependent Confidence for Out-of-Vocabulary Term Detection, *Proceedings of International Conference on Speech Communication and Technology*, pp.2139–2142 (2009).
46) 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，Vol.2010-SLP-81, No.13 (2010).
47) Maekawa, K., Koiso, H., Furui, S. and Isahara, H.: Spontaneous Speech Corpus of Japanese, *Proceedings of International Conference on Language Resources and Evaluation*, pp.947–952 (2000).
48) Akiba, T., Aikawa, K., Itoh, Y., Kawahara, T., Nanjo, H., Nishizaki, H., Yasuda, N., Yamashita, Y. and Itou, K.: Construction of a Test Collection for Spoken Document Retrieval from Lecture Audio Data, *Journal of Information Society of Japan*, Vol.50, No.2, pp.501–513 (2009).