特集：『検索』のゆくえ

音声ドキュメント検索：マルチメディアデータを対象とした音声言語情報検索

秋葉　友良*

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諸相について論じる。まず，音声ドキュメント検索の問題設定と，現在技術目標として評価が行われている2つのタスク設定である音声中の検索語検出（STD）と音声ドキュメントの内容検索（SCR）について述べる。また，音声ドキュメント検索に必要な技術課題を整理し，研究動向を紹介する。最後に，音声ドキュメント検索手法の性能評価に不可欠なテストコレクションの整備状況と，現在評価が進行しつつあるNTCIR SpokenDocタスクについて紹介する。

キーワード：音声ドキュメント，マルチメディアコンテンツ，音声中の検索語検索，内容検索，音声認識，認識語彙外語，認識誤り，索引付け，テストコレクション

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問題設定を整理し，それらに対する技術課題を示す。まず2章にて，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章では，音声ドキュメント検索に関する関連研究を紹介しながら，必要な技術的課題を整理する。4章では，音声ドキュメント検索手法の性能評価のためのテストコレクションについて述べるとともに，現在進行中の評価プロジェクトNTCIR SpokenDocについて紹介する。

*あきば　ともよし　豊橋技術科学大学
〒441-8580　愛知県豊橋市天伯町雲雀ヶ丘1-1
Tel. 0532-44-6758
（原稿受領　2012.11.05）

2. 音声ドキュメント検索とは

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。本稿では，検索クエリはテキストで与えられると仮定するが，音声クエリからの音声ドキュメント検索も興味深い問題である。

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，根本的にはテキストと音声データ区間の対応関係を見つける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，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，音声ドキュメント検索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，の組み合わせである。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多くの研究で報告されている。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 2つのタスク：STDとSCR

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（図1）。

一つは，Spoken Term Detection（以下，STD），日本語では音声中の検索語検出，音声キーワード検索などと呼ばれる。STDは，単語あるいは数単語の列からなる用語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を技術評価タスクに設定したこと[1]，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。

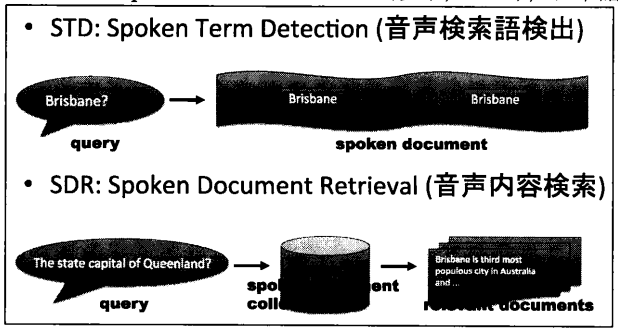

図1 STDとSDR

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（Spoken Content Retrieval；以降SCRと呼ぶ）である。このタスクを狭義でSpoken Document Retrieval(SDR)と呼ぶことも多い。SCR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，SCR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 テキスト検索との対応関係

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」を指し，これは音声ドキュメント検索でのSCR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」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」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，有向非循環グラフで表現された単語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技術課題

3.1 認識語彙外語の問題

現在，音声認識方式の中でも比較的よく利用される大語彙連続音声認識は，数万から数十万程度の単語辞書を仮定し，音声区間と最も合致する単語列候補を探索する手法である。したがって，単語辞書に現れない稀な単語については認識することができない。この問題を認識語彙外語（Out-Of-Vocabulary; OOV）の問題と呼ぶ。

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（使用頻度の低い）語が含まれ

ることが多く，認識語彙外語に成りやすい。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。例えば，日本語の音節は約 150 種類であり，全音節を語彙として連続音節認識を行えば，どのような音節列でも認識することができる。このような単語より小さい単位を総称して，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。

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，認識結果に対して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）[2]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[3]。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[4]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[5][6]も提案され効果が示されている。また，単語とサブワード[7][8]，あるいは複数のサブワード[9][10][11]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[6]。

SCR タスクにおいてサブワードを使う場合は，ベクトル空間モデルなどの検索モデルにおいて，単語の代わりにサブワードを単位とした文書間関連度を用いることができる[12]。しかし，意味を担う最小単位である単語と比べると，サブワードの表現能力は劣るため，検索性能も低下してしまう。この問題に対し，検索の前処理として検索クエリ中の単語に対して STD を用い，単語の検出結果を単位として文書関連度を求める SCR 手法も提案されている[13]。

3.2 認識誤りへの対応

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，コンフュージョン・ネットワークなどが知られている。ここでラティスとは，辺に単語やサブワードを対応させた有向非循環グラフで，音声認識結果の複数認識候補を効率よく表現するために利用されるデータ構造である。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，連続する単語の検索（フレーズ検索）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，すなわち出現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，ラティス上に辺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辺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。

以上の観点から，認識結果のラティスを圧縮して表現しておき，それを対象に検索を行う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。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。音声ドキュメント検索で用いられている複数認識候補の代表的な非可逆圧縮表現として，コンフュージョン・ネットワーク[14][2]，Position Specific Posterior Lattice(PSPL)[15]，Time-based Merging for Indexing (TMI)[16]，Time-Anchored Lattice Expansion (TALE)[17]，などが提案されている。

3.3 索引と照合

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を手がかりに検索を行

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。一方，音声ドキュメントを対象とする場合は，OOV や認識誤りの問題があるため，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，すなわち近似文字列照合を行う必要がある。

音声ドキュメント検索のための近似文字列照合手法としては，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。これまで音声ドキュメント検索に適用されたオフライン近似文字列照合の索引付け手法には，音節 tri-gram 索引による inverted file を用いる手法[18]，音素サフィックスアレイによる索引に対し音素単位の DP マッチングを用いる手法[19]，距離空間上の索引を用いる手法[20]，などが報告されている。

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[21]，サブワードの音響モデル間の距離（KL-divergence[22]や Bhattacharyya 距離[18)20)]），音素弁別特長間のハミング距離[19]などが利用されている。

3.4 検索モデル

3.4.1 文書の順序付け

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[23]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[24]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。

3.4.2 文書拡張・質問拡張

SCR 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。

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，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。この視点からの研究では，言語モデルに基づく検索モデルと Probabilistic LSI(PLSI)を組み合わせた確率的文書拡張法[25]，認識単語から正解単語への翻訳モデルを用いる手法[26]，などが提案されている。

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[23]。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[27]。

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[9]，検出単語に依存した信頼度[28]なども提案されている。

4. 音声ドキュメント検索の評価

4.1 テストコレクション

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 SCR 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。

TREC での音声ドキュメント検索タスク SDR Track の初年度（1996 年，TREC-6）[29]では，Known Item Task と呼ばれる検索クエリに含まれているキーワードを含む音声ドキュメントを検索するタスクがあった．これは今日での STD に類似したタスクと言える。2006 年には，米国規格協会（NIST）が STD を新たなタスクに設定し[1]，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。一方の SCR 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[30]。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。その後，欧州の評価型ワークショップ CLEF では，2003 年 2004 年に前述の TREC SDR Track のデータを用いた Spoken Document Retrieval[31]タスクが，2005 年から 2007 年にはインタビュー音声を対象とした Speech Retrieval タスク[32]

が実施された。

一方，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，SCR および STD のテストコレクション構築が行われた[33)34)]。検索対象の音声ドキュメントは，国立国語研究所で編纂された日本語話し言葉コーパス（CSJ）[35)]の学会講演と模擬講演の約 623 時間である。

4.2 NTCIR SpokenDoc タスク

NTCIR（NII Testbets and Communities for Information access Research）[36)]は，国立情報学研究所が主催する情報アクセス技術の評価型ワークショップである。1999 年から開始され，1 年半を 1 サイクルとして 2011 年末までに 9 回のワークショップを実施，2013 年 6 月には第 10 回目の NTCIR-10 ワークショップが開催される予定である。

これまで NTCIR で様々な情報検索タスクが採択され評価が行われる中，2011 年の NTCIR-9 において音声ドキュメント検索タスク SpokenDoc が採択され実施された[37)]。

NTCIR 初の音声ドキュメント検索タスクである SpokenDoc-1 では，先に述べた音声ドキュメント処理ワーキンググループで開発した STD および SCR のテストコレクションを基盤に，CSJ を検索対象文書コレクションとした新たな検索課題を用意して，STD および SCR の 2 つのタスクが実施された。

音声ドキュメント検索の性能は音声認識の精度に依存するため，タスクオーガナイザはタスク参加グループ共通で利用できる音声認識結果を用意した。これにより，参加グループの検索結果を共通の土台の上で比較することが可能になる。また，音声認識システムを保有しない参加グループや，音声認識よりも検索手法に興味を持つ参加グループに対し，参加を容易にする環境を提供できる。参加グループの持つ様々な技術や手法に対応するため，音声認識手法には連続単語認識と連続音節認識の 2 種類を用い，認識結果の 10-best リスト，単語（音節）ラティス形式，コンフュージョンネットワーク形式，などの様々な認識結果，出力形式を提供した。

STD タスクでは，コア講演と呼ばれる 44 時間のサブセットを対象としたタスクと，全講演 623 時間を対象としたタスクの 2 種類のサブタスクが設定された。STD タスクには 7 グループが参加した。全グループがコア講演を対象としたタスクに参加し，合計 12 の run が提出された。全講演を対象としたタスクには 2 チームが参加し，合計 4 つの run が提出された。検索処理の効率の観点から見ると，5 つの run は索引を使った検出の高速化を実装しており，44 時間の音声から 1 ミリ秒前後で検索語の検出が行えている。一方，性能の観点から見た場合，高い検出性能を示した run は，共通して複数の音声認識結果を使用する手法を用いていた。例えば，最も高い性能を示した run は，合計 10 種類の認識結果を利用して高精度かつ高再現率の検出を実現している。その他，各参加グループが採用した手法の詳細は NTCIR-9 の Proceedings[38)]を参照されたい。

SCR タスクでは，次の 2 種類のサブタスクを設定した。

- 講演検索タスク：検索クエリに適合する講演を見つけるタスク
- パッセージ検索タスク：正解音声区間そのものを見つけるタスク

講演検索タスクは，テキスト検索における文書検索に相当する。しかし，音声ドキュメント検索では講演のような大きな単位が検索されたとしても，検索結果を確認するためには音声の再生が必要となり，テキストのように全体をざっと一覧することができない。したがって，よりピンポイントに検索の適合箇所（音声区間）を見つける技術が必要となる。これがパッセージ検索タスクを設定した理由である。SCR タスクには 5 グループが参加し 21 の run が提出された。このうち，講演検索タスクには 4 チームが参加し 11 の run が提出された。パッセージ検索タスクには 3 チームが参加し，10 の run が提出された。評価結果の詳細は NTCIR-9 の Proceedings[38)]を参照されたい。特に，パッセージ検索タスクでは，全体的に参加 run の評価指標の値は低く，本タスクの難しさが示されており，課題が残されていることがわかっている。

SpokenDoc-1 の評価結果を踏まえて，2013 年の NTCIR-10 に向けて 2 回目の音声ドキュメント検索評価タスク SpokenDoc-2[39)]が進行中である。SpokenDoc-2 では，新しい検索対象音声ドキュメントや，低い認識率のもとでの音声認識結果を提供し，各タスクのより詳細な評価を行う予定である。

註・参考文献

（web 参照日は全て，2012 年 10 月 30 日です）

1) National Institute of Standards and Technology. Spoken Term Detection Evaluation Portal. http://www.nist.gov/speech/tests/std/
2) Turunen, V. T.; Kurimo, M. Indexing Confusion Networks for Morph-based Spoken Document Retrieval. In Proceedings of SIGIR, 2007, p.631-638.
3) 岩田耕平，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．情報処理学会論文誌，2007，Vol.48，No.5，p.1990-2000.
4) Bisani, M; Ney, H. Joint-sequence models for grapheme-to-phoneme conversion. Speech Communication, 2008, Vol.50, No.5, p.434-451.
5) 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，2003，Vol.J86-D-II，No.10，p.1369-1381.
6) Saraclar, M.; Sproat, R. Lattice-Based Search for Spoken Utterance Retrieval. In Proceedings of Human Language Technology Conference, 2004.
7) Yu, P.; Seide, F. A Hybrid Word / Phoneme-based Approach for Improved Vocabulary-Independent Search in Spontaneous Speech. In Proceedings of International Conference on Spoken Language Processing, 2004.
8) Iwata, K. et al. Robust Spoken Term Detection Using Combination of Phone-Based and Word-Based Recognition. In Proceedings of International Conference on Speech Communication and Technology, 2008, p.2195-2198.
9) Tejedor, J. et al. A Posterior Probability-Based System Hybridisation and Combination for Spoken Term

Detection. In Proceedings of International Conference on Speech Communication and Technology, 2009, p.2131-2134.
10) 伊藤慶明，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－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，2009，Vol.50，No.2，p.524-533.
11) 名取賢，他．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，2009，Vol.2009-SLP-79，No.19.
12) Ng, K.; Zue, V.W. Subword-based Approaches for Spoken Document Retrieval. Speech Communication, 2000, Vol.32, No.3, p.157-186.
13) 瀧上智子；秋葉友良．音声検索語検出結果を用いた音声ドキュメントの内容検索．日本音響学会秋季研究発表会研究報告，2011，p.187-188.
14) Mangu, L. et al. Finding Consensus in Speech Recognition: Word Error minimization and Other Applications of Confusion Networks, Computer, Speech and Language, 2000, Vol.14, No.4, p.373-400.
15) Chelba, C.; Acero, A. Position Specific Posterior Lattices for Indexing Speech, In Proceedings of Annual Meeting of the Association for Computational Linguistics, 2005, p.443-450.
16) Zhou, Z.-Y. et al. Towards Spoken-Document Retrieval for the Internet: Lattice Indexing For Large-Scale Web-Search Architectures, In Proceedings of Human Language Technology Conference, 2006, p.415-422.
17) Yu, P. et al. Approximate Word-Lattice Indexing with Text Indexers: Time-Anchored Lattice Expansion, In Proceedings of International Conference on Acoustic, Speech, and Signal Processing, 2008, p.5248-5251.
18) Iwami, K. et al. Efficient Out-Of-Vocabulary Term Detection by N-gram Array Indices with Distance from a Syllable lattice, In Proceedings of International Conference on Acoustic, Speech, and Signal Processing, 2011, p.5664-5667.
19) Katsurada, K. et al. Fast Keyword Detection Using Suffix Array, In Proceedings of International Conference on Speech Communication and Technology, 2009, p.2147-2150.
20) 金子泰輔；秋葉友良．部分距離空間上の索引付けに基づく音声中の高速検索語検出法，電子情報通信学会論文誌，2012，Vol.J95-D，No.3，p.608-617.
21) Wallace R. et al. A Phonetic Search Approach to the 2006 NIST Spoken Term Detection Evaluation, In Proceedings of International Conference on Speech Communication and Technology, 2007, p.2385-2388.
22) Liu, P. et al. Divergence-based Similarity Measure for Spoken Document Retrieval, In Proceedings of International Conference on Acoustic, Speech, and Signal Processing, 2007, p.89-92.
23) Yu, P.; Seide, F. Fast Two-Stage Vocabulary-Independent Search in Spontaneous Speech, In Proceedings of International Conference on Acoustic, Speech, and Signal Processing, 2005, p.481-484.
24) Croft, W. B.; Lafferty, J. Language Modeling for Information Retrieval, Kluwer Academic Publishers, 2003.
25) Chen, B. Latent Topic Modeling of Word Co-Occurrence Information for Spoken Document Retrieval, Latent Topic Modeling of Word Co-Occurrence Information for Spoken Document Retrieval, 2009, p.3961-3964.
26) 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，2009，Vol.50，No.2，p.514-523.
27) 神田直之；他．多段階リスコ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，2012，Vol.J95-D，No.4，p.969-981.
28) Wang, D. et al. Term-Dependent Confidence for Out-of-Vocabulary Term Detection, In Proceedings of International Conference on Speech Communication and Technology, 2009, p.2139-2142.
29) Voorhees, E. et al. The TREC-6 Spoken Document Retrieval Track. In TREC-6 notebook, 1997, p.167-170.
30) Garofolo, J. S. et al. The TREC Spoken Document Retrieval Track: A Success Story, In Proceedings of Text Retrieval Conference (TREC) 8, 2000, p.107-129.
31) Federico, M. et al. CLEF 2004 Cross-Language Spoken Document Retrieval Track. Multilingual Information Access for Text, Speech and Images, 2005, LNCS, Vol.3491, p.816-820.
32) Pecina, P. et al. Overview of the CLEF-2007 Cross-Language Speech Retrieval Track. Advances in Multilingual and Multimodal Information Retrieval, 2008, LNCS, Vol. 5152, p. 674-686.
33) Akiba, T. et al. Construction of a Test Collection for Spoken Document Retrieval from Lecture Audio Data. Journal of Information Society of Japan, 2009, Vol.50, No.2, p.501-513.
34) Itoh, Y. et al. Constructing Japanese Test Collections for Spoken Term Detection. In Proceedings of International Conference on Speech Communication and Technology, 2010, p.667-680.
35) Maekawa, K. et al. Spontaneous Speech Corpus of Japanese. In Proceedings of International Conference on Language Resources and Evaluation, 2000, p.947-952.
36) NTCIR Project HOME
http://research.nii.ac.jp/ntcir/index-en.html
37) Akiba, T. et al. Designing an Evaluation Framework for Spoken Term Detection and Spoken Document Retrieval at the NTCIR-9 SpokenDoc Task. In Proceedings of International Conference on Language Resources and Evaluation, 2012.
38) NTCIR Workshop 9: Online Proceedings, Evaluation Results
http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/toc_ntcir.html
39) NTCIR-10 Core Task: 2nd round of IR for Spoken Documents (SpokenDoc-2)
http://www.cl.ics.tut.ac.jp/~sdpwg/index.php?ntcir10

Special feature: Futures of information retrieval. Spoken document retrieval: Searching spoken language information from multimedia data. Tomoyosi AKIBA (Toyohashi University of Technology, 1-1 Hibarigaoka, Tenpaku, Toyohashi, Aichi 441-8580 JAPAN)

Abstract: 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 sub-tasks for SDR, namely spoken term detection (STD) and spoken content retrieval (SCR) are introduced. Then, the several technological challenges 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. Especially, the NTCIR SpokenDoc task, which is currently conducted for evaluating both STD and SCR tasks, is introduced.

Keywords: spoken document / multimedia contents / spoken term detection / content retrieval / speech recognition / out-of-vocabulary words / recognition error / indexing / test collection