キーワード集合をクエリとする最良照合STD方式

堂元 健太郎[1]　宇津呂 武仁[2]　古屋 裕斗[3]　西崎 博光[3]

概要：本論文では，従来型の音素照合型 STD における過照合を回避するため，音声ドキュメントにおける出現が予測される全キーワード集合をクエリとした STD の後，照合音声区間が競合するキーワード集合に対して，照合コストを用いた順位付けを行い，照合コスト最小のキーワードのみを STD 結果として出力する「キーワード集合をクエリとする最良照合 STD」方式を提案し，その有効性を示す．

STD by Selecting the Best Match from Candidate Query Keywords

Domoto Kentaro[1]　Utsuro Takehito[2]　Furuya Yuto[3]　Nishizaki Hiromitsu[3]

Abstract: This paper proposes a novel framework of spoken term detection (STD) and keyword indices annotation to spoken documents. The underlying framework of our spoken term detection is that based on phoneme transition network (PTN) which is constructed from outputs of 10 continuous speech recognition systems. One of the major difficulties of the approach to STD based on phoneme matching is over detection of spoken terms caused by loose phoneme matching. Our approach avoids this over detection by selecting the best match from candidate query keywords that are collected from text data closely related to the spoken document. Experimental evaluation results show that the proposed framework is quite effective in annotating keyword indices to spoken documents.

1. はじめに

近年，音声や動画像などのマルチメディアコンテンツの増加に伴い，これらを効率的に検索する技術が期待されている．中でも，音声ドキュメントを対象として検索語が発話されている箇所を特定する，音声中の検索語検出 (Spoken Term Detection, STD) の研究が盛んである．一般に，STD においては，大語彙音声認識システムを用いて音声認識を行うため，音声認識誤りや未知語の対策が課題である．これらの問題に頑健な STD 手法として，10 種類の音声認識システムの認識結果から音素遷移ネットワーク (Phoneme Transition Network, PTN) 型のインデックスを構築し，これと音素列に変換した検索語の照合を行う方式 [10] が提案されている．しかし，音素照合型 STD においては，検索語と異なるキーワードの発話であっても音素列が類似していれば検出してしまうという，過照合による誤検出が重要な問題である．例えば，図 1 の例の場合，「バブルソート」,「クイックソート」,「ソート」等をクエリとして音素照合型 STD を適用した場合，「バブルソート」という音声区間が照合するため，類似音素列である「クイックソート」や「ソート」等が過照合してしまう．また，「二分探索」,「幅優先探索」,「深さ優先探索」等をクエリとして音素照合型 STD を適用した場合，「二分探索」という音声区間が照合するため，類似音素列である「幅優先探索」や「深さ優先探索」等が過照合してしまう．

そこで本論文では，当該分野の音声中に出現する可能性のあるキーワード集合をあらかじめ用意しておき，これら全てをクエリとして音素照合型 STD(従来法である PTN 型インデックスを用いた STD [10]) を適用した後，照合音声区間が競合するキーワード集合に対して，照合コストを用いた順位付けを行い，照合コスト最小のキーワードのみを STD 結果として出力する「キーワード集合をクエリと

1　筑波大学理工学群工学システム学類
College of Engineering Systems, School of Science and Engineering, University of Tsukuba
2　筑波大学システム情報系
Faculty of Engineering, Information and Systems, University of Tsukuba
3　山梨大学大学院医学工学総合教育部
Department of Education, Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi

する最良照合 STD」方式を提案する．この方式においては，図 1 左の「バブルソート」という音声区間の場合，競合するキーワード集合のうち最小コストで照合する「バブルソート」が優先され，図 1 右の「二分探索」という音声区間の場合も，競合するキーワード集合のうち最小コストで照合する「二分探索」が優先される．

2. 音素遷移ネットワークを用いた STD

本論文では，PTN 型 STD として，文献 [3] における「Vot+Acw1」[*1]を用いた．この方式においては，デコーダとして Julius rev. 4.1.3 を用い，2 種類の音響モデル (AM)，および，5 種類の言語モデル (LM) を用意して，AM と LM の組み合わせによって 10 種類の音声認識モデルを構築した．

AM としては HMM を用い，日本語話し言葉コーパス (CSJ)[*2]のコア講演以外の講演音声から学習した Syl(syllable モデル：音節 124 種) と Tri(triphone モデル：音素 43 種) の 2 種類を用いた．一方，LM は，単語もしくは文字の tri-gram モデルとして，以下の 5 種類を用いた．

(1) 漢字，英数字，平仮名，片仮名から構成される単語の tri-gram モデル (WBC)，
(2) 単語 tri-gram モデル，ただし，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 (WBH)，
(3) すべて平仮名によって構成される文字の tri-gram モデル (CB)，
(4) 2 文字の平仮名によって構成される文字系列の tri-gram モデル (BM)，
(5) LM を使用しない．連続音素 (音節) 認識と等価となる (Non)．

Non 以外のすべての LM は，CSJ のコア講演以外の講演音声を書き起こしたテキストを用いて訓練した[*3]本論文では，CSJ の講演音声 (雑音が少ない良質な音声) および模擬講義 (雑音の多い低品質な音声) [11] を評価対象として STD を適用するが，これらの音声認識モデルを用いた場合，CSJ を対象とした単語認識率は 67∼76%程度，単語正解精度は 64∼72%程度である [3]．一方，模擬講義を対象とした単語認識率は 26%，単語正解精度は 9%程度である [11]．

[*1] 誤検出抑制パラメータとして，当該音素を出力する音声認識システム数，および，音素遷移ネットワーク中の 2 ノード間のアーク数を考慮し，脱落・置換誤りコストを 1，正解の場合のコストを 0，NULL 遷移コストを 0.1 とする設定での STD．
[*2] http://www.ninjal.ac.jp/csj/
[*3] AM と LM の訓練，認識用単語辞書，音声認識条件については，いずれも，2010 年 5 月に公開された CSJ の STD 用テストコレクション [6] の条件に基づいている．

3. キーワード集合をクエリとする最良照合 STD

「キーワード集合をクエリとする最良照合 STD」によるキーワード索引付けの流れの概略を図 2 に示す．

3.1 キーワード集合

本論文の評価実験においては，講演・講義ごとにキーワード集合を人手で作成した．その際には，講演・講義の書き起こし文書に対して，専門用語抽出ツール TermExtract[*4] を適用し，出力された専門用語候補に対して，検索語として利用するキーワードを手作業によって選定した．

3.2 STD 結果の競合集合の作成

次に，キーワード集合のすべてのキーワードをクエリとして PTN 型 STD を行い，STD 結果を併合する．この結果，音声中の各区間ごとに複数の STD 照合結果が重複して得られる．このうち，検出フレーム時間が重複している照合結果を推移的に収集することにより，STD 結果の競合集合 C を作成する．

3.3 最長フレーム照合結果優先方式

従来手法 [10]，提案手法のうちの最長フレーム照合結果優先方式 (以下，「最長フレーム法」)，および，提案手法のうちの競合集合内の最小コストを用いたリランキング (以下，「最長フレーム法+リランキング」) の 3 種類の手法による STD の具体例を図 3 に示す．本節では，特に，「最長フレーム法」について述べる．

まず，キーワードを w，その STD 検出開始フレームを t，STD 検出終了フレームを t'，STD 照合コストを $cost$ とすると，n 個の四つ組 $\langle w, t, t', cost\rangle$ から成る競合集合 C は，次式のように書ける．

$$C = \left\{ \langle w_1, t_1, t'_1, cost_1\rangle,\ \ldots,\ \langle w_n, t_n, t'_n, cost_n\rangle \right\}$$

図 3 の場合の C を次式に示す．

$$\begin{aligned} C\ =\ \Big\{\ & \langle\, \text{スペクトル},\ t_1,\ t_3,\ 0.06\,\rangle, \\ & \langle\, \text{スペクトル部},\ t_1,\ t_4,\ 0.22\,\rangle, \\ & \langle\, \text{スペクトルパラメータ},\ t_1,\ t_5,\ 0.12\,\rangle, \\ & \langle\, \text{特徴パラメータ},\ t_2,\ t_5,\ 0.28\,\rangle, \\ & \langle\, \text{パラメータ},\ t_3,\ t_5,\ 0.10\,\rangle\ \Big\} \end{aligned}$$

ただし，

$$t_1 < t_2 < t_3 < t_4 < t_5$$

である．

このとき，従来手法による STD においては，図 3 に示すように，競合集合 C のすべての照合結果を出力する．

[*4] http://gensen.dl.itc.u-tokyo.ac.jp/

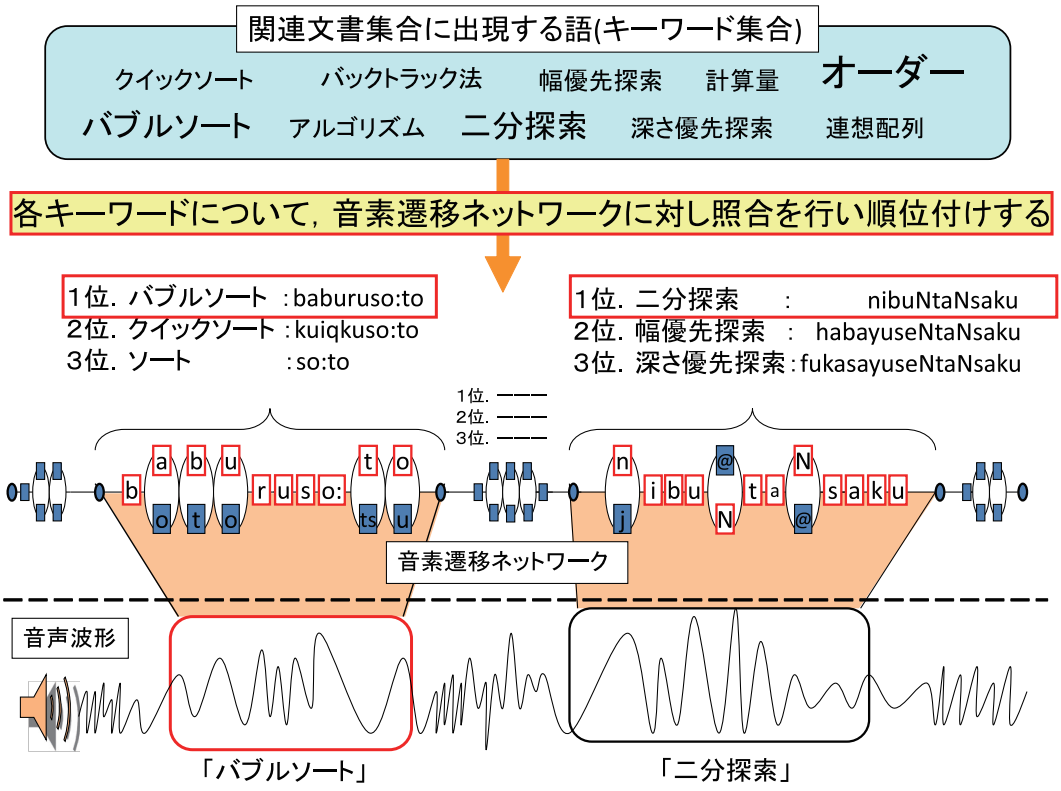

図 1 キーワード集合をクエリとする最良照合 STD

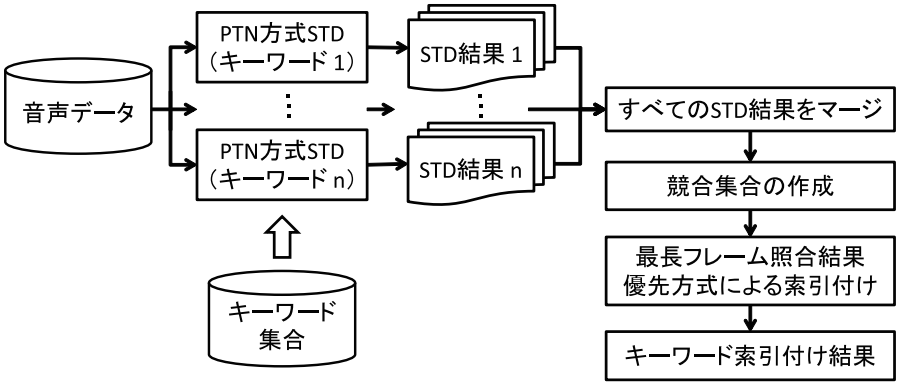

図 2 「キーワード集合をクエリとする最良照合 STD」によるキーワード索引付けの流れ

一方，「最長フレーム法」では，競合する n 個の照合結果のうち，最小コストとなる照合結果

$$\langle\, w_{min},\ t,\ t',\ cost_{min} \,\rangle$$

をまず選定する．例えば，図 3 の例では，

$$\langle\, \text{スペクトル},\ t_1,\ t_3,\ 0.06 \,\rangle$$

が最小コスト照合結果となる．

次に競合集合 C 内の照合結果について，最小コスト照合結果からコスト幅 Δ 以内にある照合結果を索引付けの候補集合 $C(\Delta)$ とする．

$$C\,(\Delta) = \left\{ \langle\, w, cost \,\rangle \in C \;\middle|\; cost \leq (cost_{min} + \Delta) \right\}$$

例えば，図 3 の例において，$\Delta = 0.10$ とすると，$C\,(\Delta = 0.10)$ は

$$C\,(\Delta = 0.10) = \left\{ \begin{array}{l} \langle\, \text{スペクトル},\ t_1,\ t_3,\ 0.06 \,\rangle, \\ \langle\, \text{スペクトルパラメータ},\ t_1,\ t_5,\ 0.12 \,\rangle, \\ \langle\, \text{パラメータ},\ t_3,\ t_5,\ 0.10 \,\rangle \end{array} \right\}$$

となる．

最後に，$C\,(\Delta)$ の要素のうち，検出フレーム長 $t' - t$ が最大となる照合結果

$$\langle\, w_{lg},\ t,\ t',\ cost_{lg} \,\rangle$$

を選定する．これが当該音声区間の STD 結果となる．例えば，図 3 の例では，

$$\langle\, \text{スペクトルパラメータ},\ t_1,\ t_5,\ 0.12 \,\rangle$$

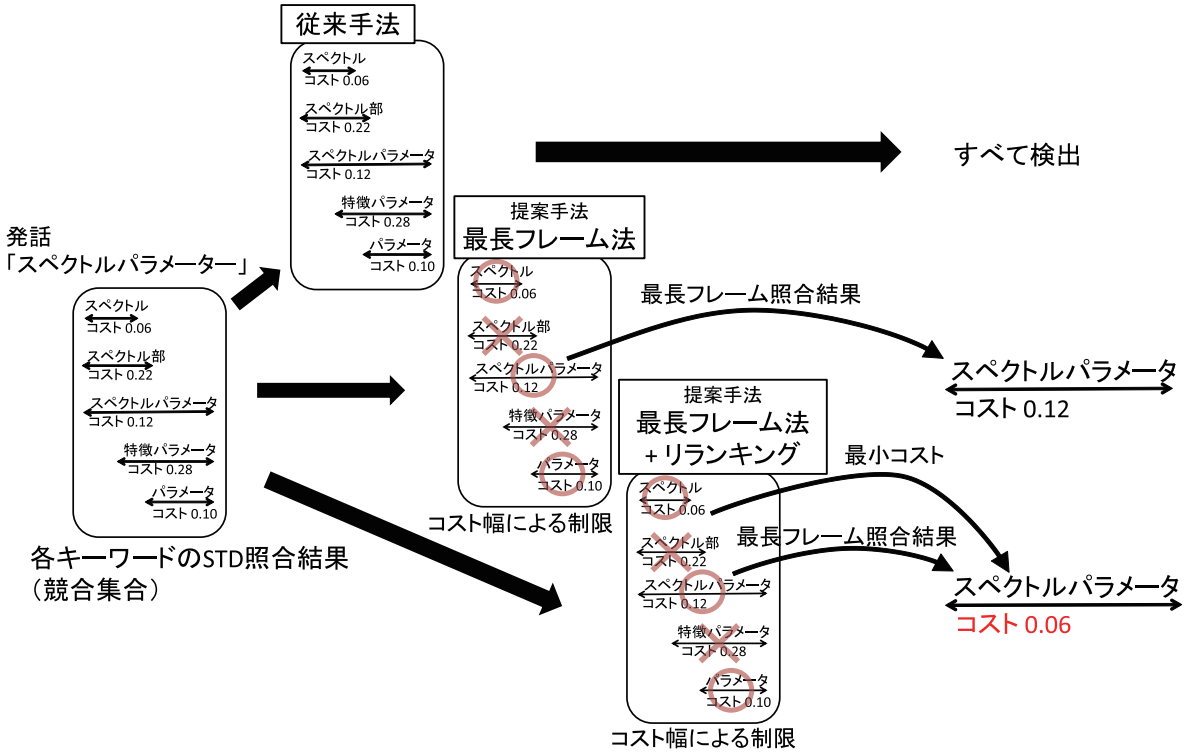

図 3　従来手法，提案手法「最長フレーム法」および「最長フレーム法+リランキング」による STD 結果

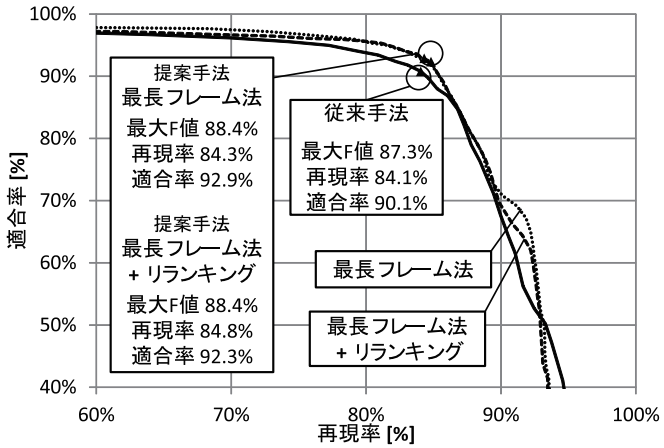

図 4　評価結果: CSJ11 講演

が STD 結果として出力される．そして，この STD 結果と検出フレーム時間が重複している照合結果を削除する．

競合集合が空になるまで以上の処理を繰り返す．

3.4　競合集合内の最小コストを用いたリランキング

一般に，音素数の多いキーワードは，STD における照合コストが大きくなる傾向がある．そこで，提案手法「最長フレーム法」における照合コストを，競合集合内の最小コストに置き換える「競合集合内の最小コストを用いたリランキング」方式を導入する．この手法による STD 結果は

$$\langle\ w_{lg},\ t,\ t',\ cost_{min}\ \rangle$$

と表記される．例えば，図 3 の例では，

$$\langle\ \text{スペクトルパラメータ},\ t_1,\ t_5,\ 0.06\ \rangle$$

となる．

4.　評価

評価対象として，CSJ に収録されている 11 講演[*5]，および，模擬講義 [4] 1 講義を対象として，従来手法 [10]，提案手法「最長フレーム法」，提案手法「最長フレーム法+リランキング」の 3 種類の手法の評価を行った．提案手法においては，講演ごとにキーワード集合を作成するため，評

[*5]　学会講演 5 講演，模擬講演 6 講演．

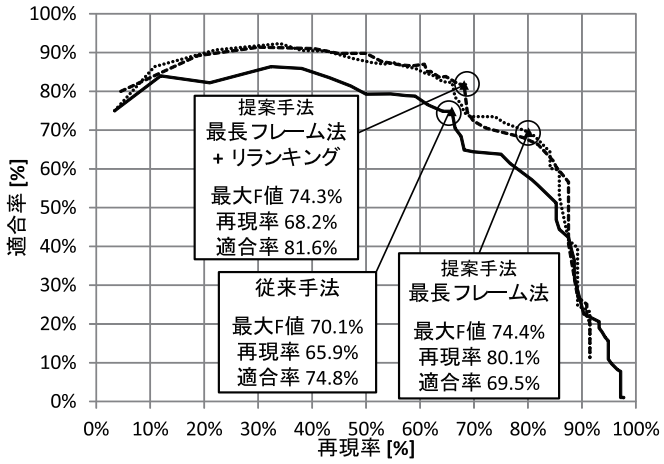

図 5 評価結果: CSJ11 講演中の模擬講演 S01F0050

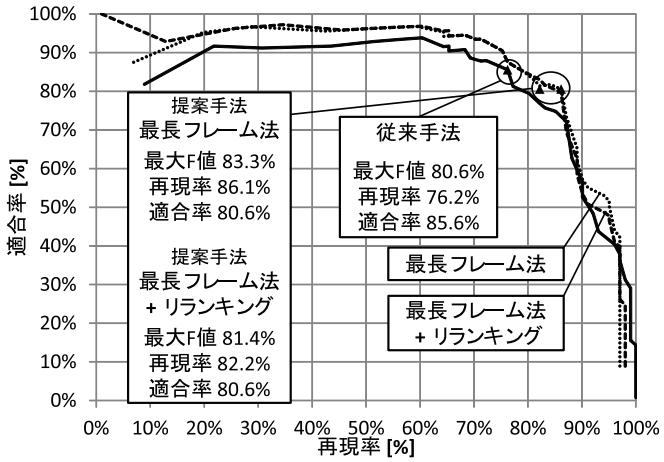

図 6 評価結果: CSJ11 講演中の模擬講演 S03F0184

価タスクは，1 講演ごとに個別に STD を行うタスクとした．また，3.1 節の手順によりあらかじめ作成したキーワード集合に含まれる全キーワードを検索語として評価を行った．各キーワード集合のキーワード種類数は，CSJ では，平均約 68 キーワード/講演，模擬講義では 96 キーワードであった．各講演におけるキーワード出現箇所数は，CSJ では平均約 253 箇所/講演，模擬講義では 587 箇所であった．提案手法におけるコスト幅 Δ は，$\Delta = 0.10$ とした．

CSJ11 講演を対象とした評価結果，11 講演中の 1 講演についての評価結果，および，模擬講義を対象とした評価結果をにおける再現率・適合率の推移を，図 4 ～ 7 に示す．

これらの結果から分かるように，提案手法では，各音声区間において，最良照合 STD 結果一つのみを出力するため，高再現率部分においては，従来手法を下回る評価結果となっているが，高適合率部分において従来手法を上回っている．また，提案手法のうちの「最長フレーム法+リランキング」は，特に品質の低い模擬講義において「最長フレーム法」の適合率を改善できている．その他に，CSJ11 講演のうちの 5 講演に対する評価結果を分析したところ，検索語が既知語の場合と比較して，特に検索語が未知語の場合において顕著な改善を示すことが分かった．

5. 関連研究

近年のマルチメディアコンテンツの増加に伴い，音声を含む大量のデータから，見たい・聴きたい箇所を検索することへのニーズが高まっている．このニーズに応えるための研究として，複数の音声ドキュメントの中から検索語が発話されているものを特定する技術である，音声ドキュメント検索 (Spoken Document Retrieval: SDR) の研究が行われている．ここで，一般的な SDR の手法では，音声ドキュメントを音声認識システムに入力し，得られた単語列に対してテキスト検索を行うことにより，音声ドキュメントを特定する．

これまでの SDR の研究の進展に関連して，米国で開催された情報検索評価型ワークショップである TREC (Text RE-trieval Conference) においては，1996 年から 2000 年まで音声ドキュメント検索が評価対象タスクとして取り上げられた [5]．一方，国内においても，情報処理学会音声

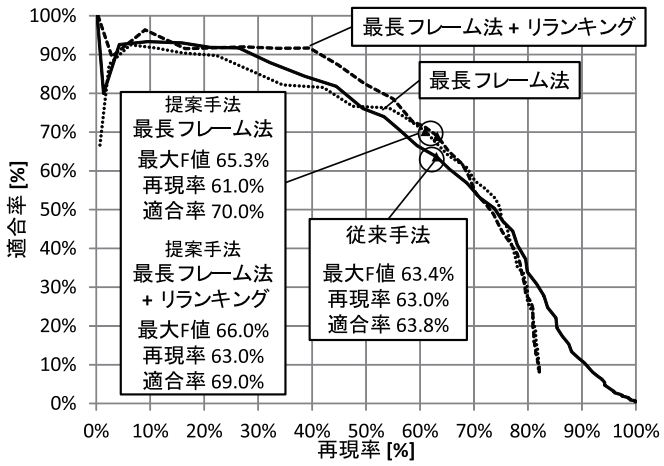

図 7 評価結果: 模擬講義 [4]

言語情報処理研究会 (SIG-SLP) において，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) が立ち上げられ，SDR 評価用テストコレクションの構築，公開を行っている [1].

ところが，SDR によって検索語と関連するドキュメント群が特定されたとしても，検索された全ドキュメント群の音声を全て聴取しなければ，音声中のどの箇所で検索語が発話されているかを特定することはできない．また，SDR において誤検出が起こる場合には，実際には検索語が一度も発話されていない可能性もある．このような背景から，音声ドキュメント中において，実際に検索語が発話されている箇所を特定する技術である検索語検出 (Spoken Term Detection: STD) の技術の研究が盛んに行われるようになった．実際に，2006 年に米国の NIST が STD を新しい研究対象として設定して以降[*6][*7]，音声関係の主要な国際会議において，多数の STD 技術についての研究報告がなされている．

ここで，一般に，音声認識結果として得られる単語系列は，音声認識システムの認識辞書に登録されている既知語の系列として構成される．したがって，検索語が認識辞書に登録されていない未知語の場合，音声認識結果に対して検索語をテキスト検索することはできない．しかし，未知語が検索語として用いられる可能性は十分考えられる [8], [9] ため，未知語を適切に検索するための手法として，例えば，音素単位で音声認識を行い，その認識結果と音素列に変換した検索語を照合することによって検索語 (未知語) の検出を行う方式等 [7], [10] が提案されている．

一方，音素単位の照合を行う STD 手法では，検索語と異なるキーワードの発話であっても認識した音素列が類似していれば検出してしまうという，過照合による誤検出が重要な問題である．そこで，本論文では，この問題に焦点を当て，キーワード集合を用いた索引付け手法を提案した．

6. おわりに

本論文では，当該分野の音声中に出現する可能性のあるキーワード集合をあらかじめ用意しておき，これら全てをクエリとして音素照合型 STD を適用した後，照合音声区間が競合するキーワード集合に対して，照合コストを用いた順位付けを行い，照合コスト最小のキーワードのみを STD 結果として出力する「キーワード集合をクエリとする最良照合 STD」方式を提案し，その有効性を示した．

今後は音声ドキュメント処理ワークショップの講演音声 (SDPWS) [2] を対象とした評価を行う．また，本論文においては，講演・講義の書き起こし文書を参照して，手作業でキーワード集合を作成したが，今後は，音声認識結果に含まれる専門用語をクエリとした Web 検索を行うことによって，当該分野の専門文書を収集し，そこからクエリ候補となるキーワード集合を自動生成する手法を確立する．

[*6] NIST, 2006 Spoken Term Detection Evaluation, available from `http://www.itl.nist.gov/iad/mig/tests/std/2006/`.

[*7] NIST, 2006 Spoken Term Detection Evaluation Plan, available from `http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf`.

参考文献

[1] Akiba, T., Kiyoaki, A., 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, *IPSJ Journal*, Vol. 50, No. 2, pp. 1234–1245 (2009).

[2] Akiba, T., Nishizaki, H., Aikawa, K., Hu, X., Itoh, Y., Kawahara, T., Nakagawa, S., Nanjo, H. and Yamashita, Y.: Overview of the NTCIR-10 SpokenDoc-2 Task, *Proc. 10th NTCIR Workshop Meeting*, pp. 573–587 (2013).

[3] 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，第 6 回音声ドキュメント処理ワークショップ SDPWS2012-11 (2012).

[4] 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 (2014).
[5] Garofolo, J. S., Auzanne, C. G. P. and Voorhees, E. M.: The TREC Spoken Document Retrieval Track: A Success Story, *Proc. 9th TREC*, NIST (2000).
[6] 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の構築と分析，情報処理学会論文誌， Vol. 54, No. 2, pp. 471–483 (2013).
[7] Iwata, K., Shinoda, K. and Furui, S.: Robust spoken term detection using combination of phone-based and word-based recognition, *Proc. 9th Interspeech*, pp. 2195–2198 (2008).
[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] Logan, B. and Thong, J. M. V.: Confusion-based Query Expansion for OOV words in Spoken Document Retrieval, *Proc. ICSLP* (2002).
[10] Natori, S., Furuya, Y., Nishizaki, H. and Sekiguchi, Y.: Spoken Term Detection using Phoneme Transition Network from Multiple Speech Recognizers' Outputs, *Journal of Information Processing*, Vol. 21, No. 2, pp. 176–185 (2013).
[11] 米倉千冬，太田晃平，古屋裕斗，西崎博光，関口芳廣：電子ノート作成支援システムで利用する音声からのキーワード検出技術，電子情報通信学会技術研究報告, NLC2013-3, Vol. 113, No. 83, pp. 13–18 (2013).