認識結果から生成したキーワード集合を用いた分類器による最良照合STD *

☆堂元健太郎, 宇津呂武仁 (筑波大), 澤田直輝, 西崎博光 (山梨大院)

1 はじめに

一般に，音声中の検索語検出 (Spoken Term Detection, 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 [1] 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．

そこで文献 [2] 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 [1]) 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「最良照合 STD によるキーワード集合の索引付け」方式を提案した．この方式の評価結果では，低再現率箇所での適合率を改善できたことから，誤検出の可能性が高い検出結果を抑制できることがわかった．しかし，この方式では正しい検出結果まで抑制してしまうため，高再現率部分での検索性能が低下していた．

この問題に対して本稿では，誤検出の可能性が高い検出結果を抑制するのではなく，「事前索引付け結果におけるクエリ検出個所への SVM による信頼度付与およびリランキング」方式を提案し (Fig. 1)，クエリ検出個所の候補に対して，信頼度推定結果の降順にクエリ検出個所を出力する．この方式においては，まず，「最良照合 STD によるキーワード集合の索引付け」方式を用いて，音声ドキュメントの音声認識結果から生成した補助的キーワード集合を事前索引付けし，また，この補助的キーワード集合事前索引付け結果と検索クエリキーワードの音素照合結果を併合する．そして，この事前索引付け結果の併合結果に対して SVM を適用することにより，クエリの検出個所の候補に対して信頼度の推定を行い，信頼度推定結果の降順にクエリ検出個所を出力する．本稿では，提案手法の評価結果において，従来手法である PTN 型インデックスを用いた STD [1] を上回ることを示す．

2 音素遷移ネットワーク型 STD

本稿では，PTN 型 STD として，文献 [3] における「ALPS-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．本稿では，SDPWS 講演 [4] を評価対象として STD を適用する．この講演音声を対象として単語認識率は 68%, 単語正解精度は 63%程度である [4].

3 キーワード集合を用いたPTNへの事前索引付け

本節では，音声認識結果から補助的キーワード集合を作成し，PTN に対して，補助的キーワード集合の最良照合結果とクエリの照合結果を索引付けする方式について述べる．

3.1 音声認識結果を用いた補助的キーワード集合の作成

次節で述べる「最良照合によるキーワード集合の索引付け」において索引付け対象として用いる補助的キーワード集合として，本稿では，音声ドキュメントの音声認識結果から生成した補助的キーワード集合を用いる．具体的には，音声ドキュメントの音声認識結果に含まれる全形態素を要素とする集合を作成する．ただし，この際，1 モーラの語を除外し，これを補助的キーワード集合 A とする．

3.2 PTN への補助的キーワード照合結果の競合集合の作成

補助的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 への照合を行い，補助的キーワード照合結果を併合する．この結果，音声中の各区間ごとに複数の補助的キーワード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，補助的キーワード照合結果の競合集合 C を作成する．

---

*STD using Classifier trained with Pre-indexed Keywords collected from ASR Outputs, by DOMOTO, Kentaro, UTSURO, Takehito (University of Tsukuba), SAWADA, Naoki, NISHIZAKI, Hiromitsu (University of Yamanashi)

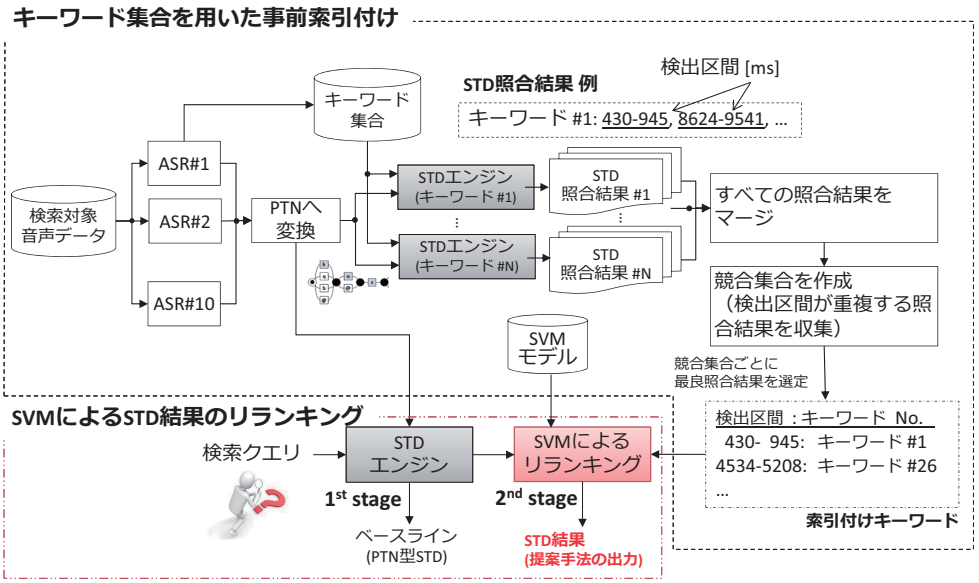

Fig. 1　事前索引付け結果におけるクエリ検出個所への SVM による信頼度付与およびリランキングの流れ

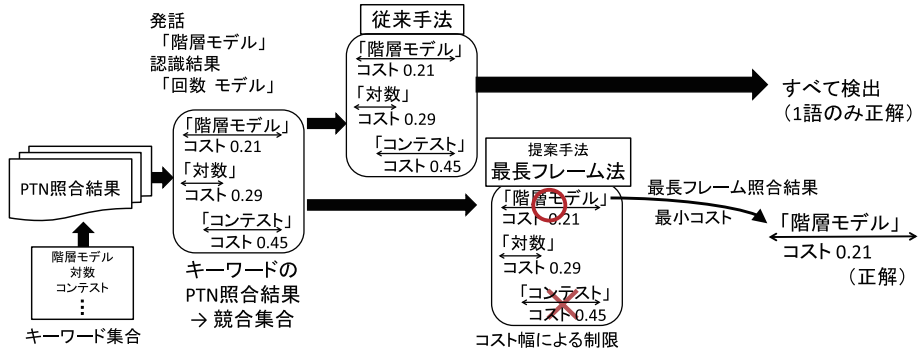

Fig. 2　競合集合における最長フレーム照合結果優先方式

3.3　競合集合における最長フレーム照合結果優先方式

競合集合中からキーワード索引結果を選定する方式として，本節においては，特に，「最長フレーム法」(最長フレーム照合結果優先方式) について述べる．

まず，キーワードを w，その照合開始フレームを t，照合終了フレームを t'，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．

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

このとき，従来手法による STD においては，Fig. 2 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．

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

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

をまず選定する．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．

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

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

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

を選定する．これが当該音声区間の最良照合結果となる．そして，この最良照合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．

この索引付け結果は，相互に重複しない最良照合結果の列

$$\langle w_{lg},\ t_b,\ t'_b,\ cost_{lg} \rangle_1,\ \ldots,\ \langle w_{lg},\ t_b,\ t'_b,\ cost_{lg} \rangle_m$$

として表現される．

3.4　補助的キーワード集合およびクエリの照合結果の併合

クエリ w_q の照合結果 $\langle w_q, t_q, t'_q, cost_q \rangle$ と，補助的キーワード集合の最良照合結果の列を併合する．この

Table 1　SVM に用いた素性

分類	重要な素性	その他の素性
クエリ素性	クエリのモーラ数 /クエリが既知語か未知語か /クエリの PTN 照合コスト	クエリの検出フレーム数
競合キーワード集合の素性	競合キーワードの文字種 /競合キーワードの検出フレーム数 /クエリとの重複検出フレーム率 /競合集合中の索引付けキーワードの数 /競合集合中の最小コストの値	競合キーワードのモーラ数 /競合キーワードの品詞 /競合キーワードの照合コスト /クエリとの重複検出フレーム数 /クエリとの PTN 照合コストの差 /競合キーワードとクエリの音素編集距離 /競合集合中でクエリの検出区間が最長か

際，クエリ w_q の照合結果 $\langle w_q, t_q, t'_q, cost_q\rangle$ の検出区間は，最良照合結果の列のうちの一つ以上の部分列

$$\langle w_{lg},\ t_b,\ t'_b,\ cost_{lg}\rangle_i,\ \ldots,\ \langle w_{lg},\ t_b,\ t'_b,\ cost_{lg}\rangle_j$$

と重複する．そこで，この場合の競合集合 C_q は，次式となる．

$$C_q = \Big\{\langle w_{lg}, t_{lg}, t'_{lg}, cost_{lg}\rangle_i,\ \ldots,\ \langle w_{lg}, t_{lg}, t'_{lg}, cost_{lg}\rangle_j,\ \langle w_q, t_q, t'_q, cost_q\rangle\Big\}$$

次節においては，この競合集合 C_q から得られる情報を素性として SVM を適用する。

4　クエリ検出個所への SVM による信頼度付与

本節では，競合集合 C_q に対して，クエリ w_q の検出結果が妥当であるか否かを判定するタスクに対して SVM を適用し，その判定結果における信頼度推定結果の降順にクエリ検出個所を出力する方式について述べる．このタスクにおいて用いる全 16 種類の素性を Table 1 に示す．これらの素性のうち，全素性から当該素性を除外することにより検索性能が低下する場合に，当該素性は「重要な素性」であるとした．

本稿では SVM 分類器として LIBSVM[5] を用いた．カーネル関数として RBF カーネルを用い，全ての素性の値を [0,1] の範囲にスケーリングした．また，5 分割交差検定でのグリッドサーチによりパラメータを調整した．さらに，LIBSVM において，判定結果の推定信頼度を出力する機能を用いることにより，クエリ検出個所の候補に対して，信頼度推定結果の降順にクエリ検出個所を出力した．

5　評価

本稿の評価においては，NTCIR-10 SpokenDoc-2 Task [4] の SDPWS STD タスクで用いられた評価用データ集合を用いた．評価用データ集合における講演数は 104 講演，クエリは既知語 47 語と未知語 53 語の計 100 語である．ベースラインとしての従来手法として，2 節で説明した PTN 型 STD [3] を比較対象とした．提案手法のうち，事前索引付けの「最長フレーム照合結果優先方式」においては，コスト幅 $\Delta = 0.20$ とした．SVM 適用の際には，以下の 2 種類の交差検定を行った．

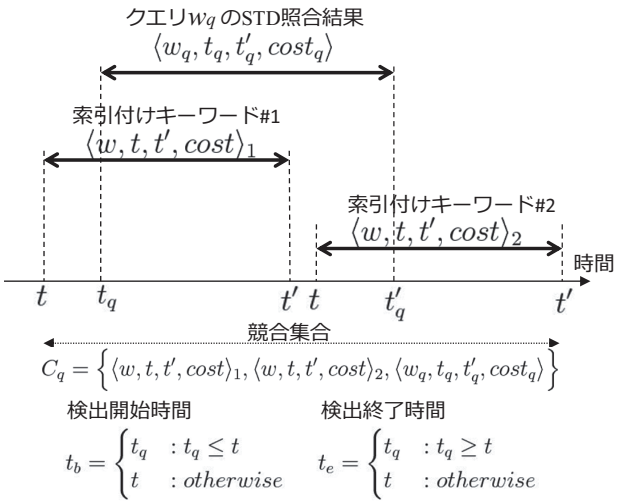

Fig. 3　競合集合 C_q 中のクエリと競合キーワードの間での重複検出フレームの定義

クエリ単位交差検定 ：100 語のクエリを 10 分割
講演単位交差検定 ：104 講演を 10 分割

そして，Table 1 に示す素性のうち，以下の 3 種の素性集合を評価対象とした．

全素性 ：16 種類すべての素性
最高性能かつ素性数最小の素性集合 ：Table 1 の「重要な素性」8 種類
クエリ素性 ：競合キーワード集合の素性以外のクエリ素性 4 種類．事前索引付け結果における競合キーワード集合から得られる素性の有効性を検証するために用意した．

SVM により付与された信頼度に対して下限値を設けて，特定の下限値以上の信頼度となるクエリ検出個所に対して再現率・適合率を評価し，その推移をプロットした結果を Fig. 4， Fig. 5 に示す．() 内の数値は F 値の最大値である．いずれの設定においても，高適合率部分において，提案手法は従来手法を上回っている．また，クエリ単位交差検定と比較すると，講

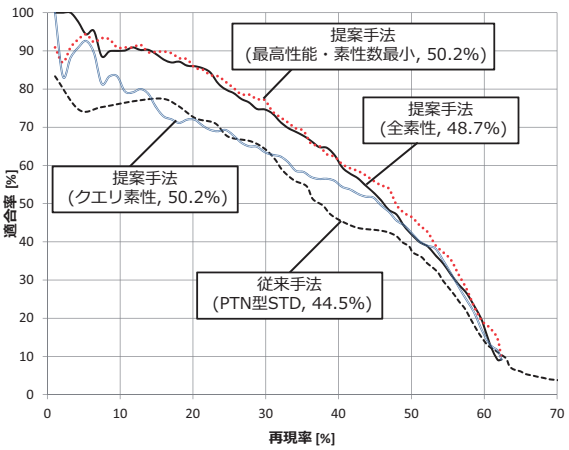

Fig. 4　評価結果：クエリ単位交差検定

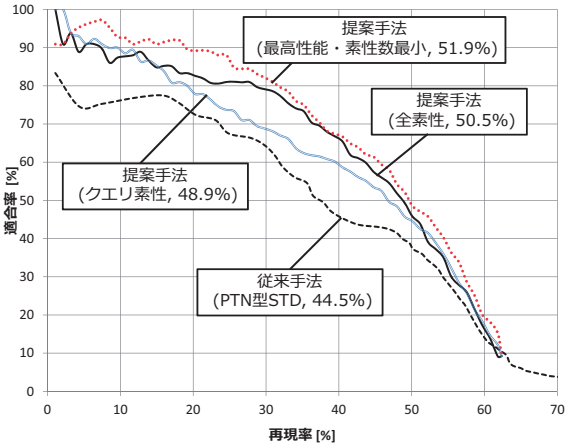

Fig. 5　評価結果：講演単位交差検定

演単位交差検定の方が高い検索性能となった．これは，用意した素性のうち，クエリに依存した素性はあるが講演に依存した素性は無いため，講演単位交差検定において，クエリに依存した素性の特性がクローズドとなったためであると考えられる．一方，「最高性能かつ素性数最小の素性集合」においては，全素性を利用した場合とほぼ同等以上の検索性能となった．さらに，「クエリ素性」のみを用いた場合の検索性能が，他の素性集合を用いた場合の検索性能を大きく下回る結果となったことから，本稿における主たる提案の一つである，「事前索引付け結果における競合キーワード集合から得られる素性」が有効であることが示された．

6　関連研究

STD に対して機械学習手法を適用する研究は近年増加している．例えば，深層学習や重回帰分析，SVM，多層パーセプトロンを適用した研究では，検出候補から出力を決定する過程 [6, 7, 8] や，出力のリランキングを行う過程 [9, 10] における信頼度の推定に用いられている．その他，音響的な素性 [11, 12] や，音声認識に関連した素性 [13] が提案されている．提案手法においては，SVM によるリランキングを行う点で既存研究と関連するが，事前索引付け結果を情報源とした素性を提案し，これにより検索性能が改善することを示した．今後，本稿の素性，および，関連研究における素性を併用することにより，検索性能を更に改善することが期待できる．

7　おわりに

本稿では，音声ドキュメントの音声認識結果から生成した補助的キーワード集合と検索クエリキーワードを用いて事前索引付けを行い，その結果に対して SVM を適用することにより，検索クエリの検出個所に対して信頼度を付与する手法を提案した．評価実験の結果，従来手法である PTN 型インデックスを用いた STD [1] を上回ることを示した．

参考文献

[1] S. Natori, et al.: “Spoken term detection using phoneme transition network from multiple speech recognizers’ outputs”, Journal of Information Processing, **21**, 2, pp. 176–185 (2013).

[2] 堂元他：“キーワード集合をクエリとする最良照合 STD 方式”, 8th SDPWS (2014).

[3] Y. Furuya, et al.: “STD and SCR Techniques and Their Evaluations on the NTCIR-10 SpokenDoc-2 task”, Proc. 10th NTCIR, pp. 626–633 (2013).

[4] T. Akiba, et al.: “Overview of the NTCIR-10 SpokenDoc-2 task”, Proc. 10th NTCIR Workshop Meeting, pp. 573–587 (2013).

[5] C.-C. Chang, et al.: “LIBSVM: A library for support vector machines”, ACM Transactions on Intelligent Systems and Technology, **2**, 3, pp. 27:1–27:27 (2011).

[6] X. Wang, et al.: “Improved Mandarin spoken term detection by using deep neural network for keyword verification”, Proc. 10th ICNC, pp. 144–148 (2014).

[7] D. Wang, et al.: “Term-dependent confidence for out-of-vocabulary term detection”, Proc. 10th INTERSPEECH, pp. 2139–2142 (2009).

[8] J. Tejedor, et al.: “An evolutionary confidence measurement for spoken term detection”, Proc. 9th CBMI, pp. 151–156 (2011).

[9] T.-W. Tu, et al.: “Improved spoken term detection using support vector machines with acoustic and context features from pseudo-relevance feedback”, Proc. 12th ASRU, pp. 383–388 (2011).

[10] N. Sawada, et al.: “Re-ranking of spoken term detections using CRF-based triphone detection models”, Proc. 6th APSIPA, pp. 1–4 (2014).

[11] H. Wang, et al.: “An acoustic segment modeling approach to query-by-example spoken term detection”, Proc. 37th ICASSP, pp. 5157–5160 (2012).

[12] S.-R. Shiang, et al.: “Spoken term detection and spoken content retrieval: Evaluations on NTCIR-11 SpokenQuery&Doc task”, Proc. 11th NTCIR, pp. 371–375 (2014).

[13] L. Mangu, et al.: “Exploiting diversity for spoken term detection”, Proc. 38th ICASSP, pp. 8282–8286 (2013).