音声認識結果から生成した補助的キーワード集合を利用する最良照合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．この方式を一般化すると，Fig. 1 に示す「最良照合によるキーワード集合の索引付け」としてとらえることができる．Fig. 1 の索引付け方式においては，例えば，Fig. 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，Fig. 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．

ここで，この「最良照合によるキーワード集合の索引付け」方式においては，あらかじめ用意しておくクエリ・キーワード集合をどのように作成するか，という点が最も重要な問題である．この問題に対して，本論文では，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利用する方式を提案する．この方式においては,「最良照合によるキーワード集合の索引付け」方式において，まず，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の索引付けを行う．そして，この索引付け結果に対して，検索クエリキーワードを最良照合する二段階最良照合結果に対して STD を行う (Fig. 2)．本論文では，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．また，音声認識結果ではなく書き起こしテキストから生成した補助的キーワード集合を索引付けする方式との比較を行った結果においても，ほぼ同等かそれ以上の検索性能が達成できることを示す．

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 最良照合によるキーワード集合の索引付け

本節では，PTN に対してキーワード集合を照合し，その結果に対して最良照合結果を索引付けする一般的な方式について述べる．

3.1 PTN へのキーワード照合結果の競合集合の作成

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，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，キーワード照合結果の競合集合 C を作成する．

3.2 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

*STD by Selecting the Best Match using Keywords collected from ASR Outputs, by DOMOTO, Kentaro, UTSURO, Takehito (University of Tsukuba), SAWADA, Naoki, NISHIZAKI, Hiromitsu (University of Yamanashi)

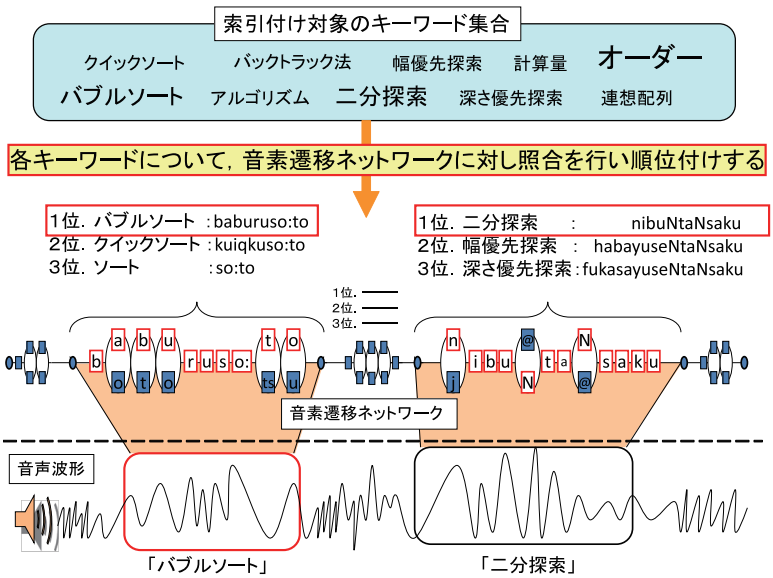

Fig. 1　最良照合によるキーワード集合の索引付け

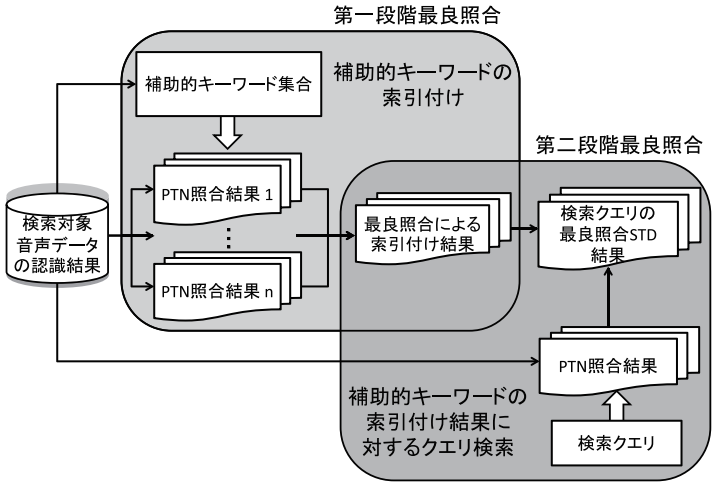

Fig. 2　補助的キーワード集合およびクエリを用いた二段階最良照合による STD の流れ

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. 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．

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．

4　音声認識結果を用いた補助的キーワード集合の作成

前節で述べた「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キーワード集合として，本論文では，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

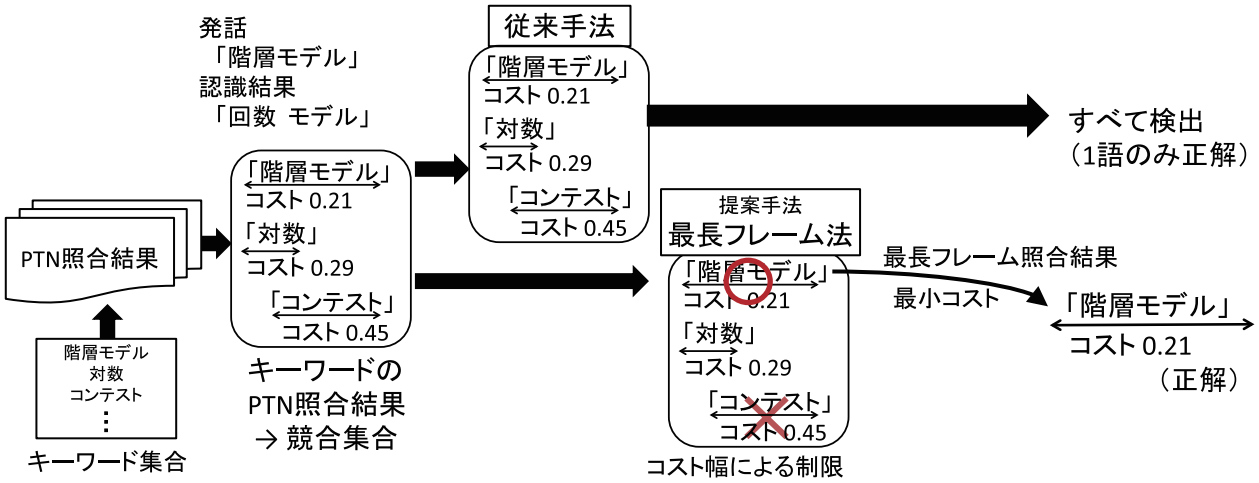

Fig. 3　競合集合における最長フレーム照合結果優先方式

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．

5　補助的キーワード集合およびクエリを用いた二段階最良照合による STD

本節では，前節で作成した補助的キーワード集合 A およびクエリ集合 Q 中のクエリ $q(\in Q)$ を用いた二段階最良照合方式およびこの方式を用いた STD について述べる．

二段階最良照合のうちの一段階目においては，3 節で述べた「最良照合によるキーワード集合の索引付け」方式に基づき，補助的キーワード集合 A 中のキーワードを検索対象音声ドキュメントに対して索引付けする．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．

次に，二段階目においては，四つ組で表現されたクエリ $\langle w_q, t_q, t'_q, cost_q\rangle$ と一段階目の最良照合結果の列との間で再度最良照合を行う．その際，クエリ $\langle w_q, t_q, t'_q, cost_q\rangle$ は，一段階目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 は，次式となる．

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

この競合集合 C に対して，3.2 節の最長フレーム照合結果優先方式を適用し，クエリ $\langle w_q, t_q, t'_q, cost_q\rangle$ もしくは，補助的キーワード集合中の部分列

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

のうちのいずれか一つだけが，二段階最良照合結果として索引付けされる．

6　評価

本論文の評価においては，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．提案手法の評価においては，補助的キーワード集合を作成するにあたって，以下の三種類の情報源から生成されたキーワード集合を用いて STD 性能の比較評価を行った．

(1) 検索対象の講演音声を音声認識した結果 (ASR)
(2) 検索対象の講演音声の書き起こし
(3) (1) および (2) の共通のキーワード (書き起こし $\bigcap$ ASR)

これら三種類の補助的キーワード集合中のキーワードの異なり数，および，第一段階最良照合による索引付け結果中の延べ数を Table 1 に示す．

提案手法におけるコスト幅 Δ を，$\Delta = 0.20$ として，既知語クエリ+未知語クエリ，既知語クエリ，および，未知語クエリを対象として行った評価結果を，それぞれ，Fig. 4 ～ 6 に示す．三種類のいずれの補助的キーワード集合を用いた場合でも，高適合率部分において，提案手法は従来手法より上回っている．ただし，提案手法においては，各音声区間の競合集合において最良照合 STD の結果一つのみを出力するた

Table 1　補助的キーワードの数

キーワードの情報源	異なり数	第一段階最良照合による索引付け結果中の延べ数
ASR (音声認識結果)	6,575	278,197
書き起こし	8,493	260,149
書き起こし ⋂ ASR	4,162	284,058

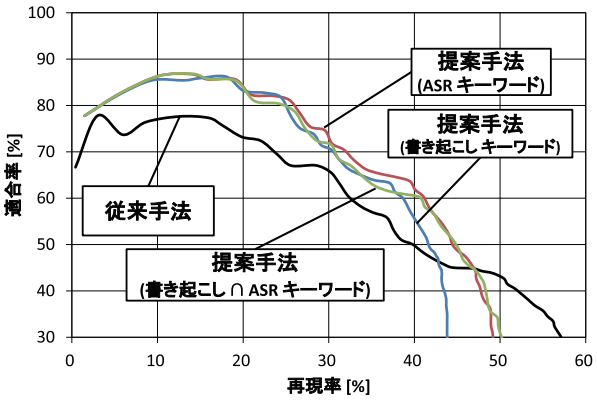

Fig. 4　評価結果: 既知語クエリ+未知語クエリ

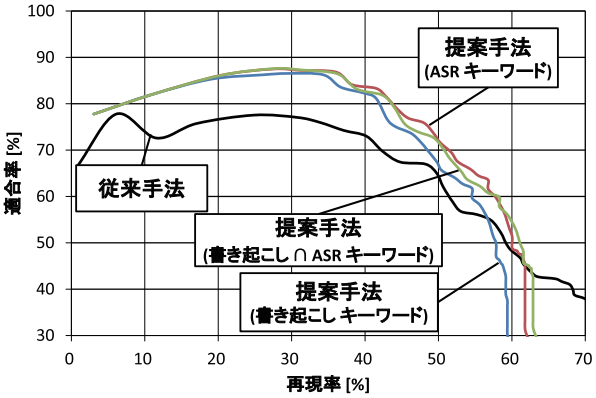

Fig. 5　評価結果: 既知語クエリ

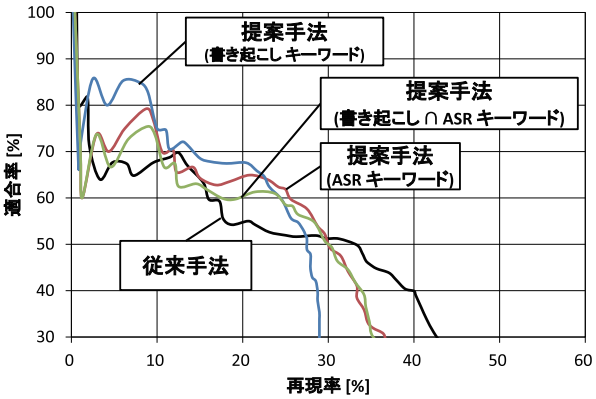

Fig. 6　評価結果: 未知語クエリ

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．また，音声認識結果ではなく書き起こしテキストから生成した補助的キーワード集合(「書き起こしキーワード」)を索引付けする方式との比較を行った結果においては，ほぼ同等かそれ以上の検索性能が達成できた．また,「書き起こし ⋂ ASR キーワード」が「ASR キーワード」とほぼ同等かやや劣る性能を示し,「書き起こしキーワード」よりも高い性能を示すことから，書き起こしにのみ含まれ，音声認識結果に含まれない語が性能を下げる要因となっていると推測される．これらの語の中には，機能語等，音声認識が相対的に困難で，かつ，STD における有用性も低い語が多く含まれている．さらに，それらの語のモーラ数が相対的に大きめであることから，検索対象音声データとの過照合を引き起こす可能性が高く，特に高再現率の領域において性能を下げる要因であると考えられる．

7　おわりに

本論文では,「最良照合によるキーワード集合の索引付け」方式において，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利用する方式を提案した．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] を上回ることを示した．さらに，音声認識結果ではなく書き起こしテキストから生成した補助的キーワード集合を索引付けする方式との比較を行った結果においても，ほぼ同等かそれ以上の検索性能が達成できることを示した．

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”，第 8 回音声ドキュメント処理ワークショップ SDPWS2014-09 (2014).

[3] Y. Furuya, et al.: “STD and SCR Techniques and Their Evaluations on the NTCIR-10 SpokenDoc-2 task”, Proc. 10th NTCIR Workshop Meeting, 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).