Vol.2009-SLP-79 No.19
2009/12/22

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

名取　賢[†1]　西崎博光[†2]　関口芳廣[†2]

本稿では，複数の音声認識システムを用いた音声中の検索語検出について述べる．提案手法では，モデルの異なる複数の音声認識システムの認識結果から，コンフュージョンネットワークの形で，検索用のインデックスを構築する．このインデックスを利用することで，音声認識誤りや未知語に対して頑健な検索が期待できる．構築したインデックスでは単純な検索方法では湧き出し誤りが大量に発生してしまう．この湧き出し誤りを抑制するために簡単なフィルタを実装した．実験の結果，提案手法の潜在的な有効性を示すことができた．特に検索語が未知語である場合において 0.71 という高い再現率が得られた．

Spoken Term Detection Using Multiple Recognition Systems

Natori Satoshi,[†1] Nishizaki Hiromitsu[†2]
and Sekiguchi Yoshihiro [†2]

This paper describes a spoken term detection (STD) technique using syllable transition networks (STNs) derived from multiple speech recognition systems' outputs. In the proposed technique, an index for retrieving a query term consists of syllable-based confusion networks which are constructed by combining multiple outputs of speech recognizers. Although the index is robust to the speech recognition errors and the out-of-vocabulary problem, the index produces a number of false detection errors on the STD task. Therefore, we designs a simple filter which can control the false detection errors. The experiment of STD showed that our technique was very effective at detecting out-of-vocabulary terms, improving recall rate to 0.71 from the baseline.

1. はじめに

近年，ネットワークインフラの充実によって，動画コンテンツに代表される音声やマルチメディアコンテンツが急激に充実してきた．これらのコンテンツはネットワークストレージや動画共有サイトなどにアクセスすることで，容易に利用することができる．そして，いまこの瞬間も，コンテンツの量は急速に増加し続けている．

大量のマルチメディアコンテンツが充実するに従って，必要な情報を効果的に検索する技術が必要となってくる．しかし，現在のところ効果的な検索技術は確実性に欠けている．そのため，検索技術の進歩はますます重要となっている．

1990 年代の後半に，NIST と DARPA によって開かれた情報検索のコンテスト (TREC : Text REtrieval Conference) の音声ドキュメント検索 (SDR : Spoken Document Retrieval) 部門において，英語と標準中国語のニュースドキュメント検索に対する多くの研究成果が発表された．そして，NIST は 2006 年に音声中の検索語検出 (STD : Spoken Term Detection) プロジェクトの試験評価とワークショップを開始した．

STD は SDR とは異なる．STD の目的は，音声中の選択された用語の位置を見つけることにある．この検出において最も難しい問題は，検索語が認識不可な語 (未知語) の場合である．この場合，音声認識システムが正しく認識することができないため，単純な文字列検索による検出は困難となる．この STD に取り組む多くの研究が，既に発表されている[1)2)3)]．STD の研究の大部分は未知語と音声認識誤りの問題に焦点を合わせている．例えば，サブワードラティスやコンフュージョンネットワーク (CN) などを使用する STD の技術が提案されている[4)5)]．

本稿では，複数の音声認識システムの出力から得られた音節遷移ネットワーク (STN : Syllble Transition Network) を使用した，大学講義音声に対する STD について述べる．本研究が典型的な STD 技術と異なる点は，複数の音声認識システムを使用することにある．複数の認識システムの出力から得られる音節系列は，結合し STN に変換される．本研究では，同一のデコーダを使用した 10 種類の音声認識システムを利用する．使用するモデルは，2 種類の音響モデル (triphone ベースと syllable ベース) と 5 種類の言語モデル (単語ベースとサブワードベース) を用意した．

複数の認識システムとその出力を使用することは，音声認識性能を向上させることにおい

†1 山梨大学大学院医学工学総合教育部コンピュータ・メディア工学専攻
Dept. of Computer Science and Media Engineering, Educational Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi
†2 山梨大学大学院医学工学総合研究部
Dept. of Research Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi

© 2009 Information Processing Society of Japan

て非常に効果的であることが知られている．例えば，Fiscus[6] は単語投票方式を採用する ROVER 法を提案している．また，宇津呂ら[7] は音声認識性能を向上させるのに，サポートベクタマシン (SVM) を使用することによって，複数の認識システムの出力を結合するための技術を見出した．複数の認識システムによる単語 (または，サブワード系列) 出力の適用は，各音声認識システムの特性が異なっているため，良い音声認識性能を示すことが可能となる．複数の認識システムの出力に基づく STN は，発話された用語に対するより多くのサブワード系列をカバーできる．

STD の実験結果では，STN が STD の性能を向上させることにおいて有効であることが示された．特に，未知語検出に対して非常に頑健な性能を示した．

2. 複数音声認識システムを用いた音声中の検索語検出

本研究では，複数の音声認識システムの出力を利用し，STD のための検索用インデックス (STN) を構築する．構築した STN から検索用語の検出を行う．

STN はサブワードが単一の音声認識システムによって生成されたコンフュージョンネットワーク (CN) と基本的に同様のものである．Gao ら[4] は，STD にサブワードベースの CN を使用することを提案した．この CN は単一の音声認識システムから生成されているが，我々が提案する STN は複数の音声認識システムの出力から変換された音節系列から生成される．参考文献[6][7] に示されているように，複数の出力は音声認識性能を向上させる．従って，認識誤りに対して頑健となり，STD の性能を向上させるために有効であると考えられる．

図 1 に本稿の STD の概要を示す．まず，検索対象の音声データを 10 種類の音声認識システムによって認識させる．これらの認識結果は音節系列に変換される．次に，音節系列間のマルチプルアライメントを DP マッチングによって行う．整合が取られた音節系列は STN に変換される．STN から検索用語の検出を行い，その結果から湧き出し抑制のためのフィルタを通し，最終的な検出結果とする．

3. 音節遷移ネットワークの構築

3.1 複数の音声認識システムの利用

図 1 に示されているように，音声データは 10 種類の音声認識システムによって認識される．デコーダには Julius rev. 4.1.2[8](LVCSR のためのオープンソースデコーダ) を使用した．このデコーダに対し，2 種類の音響モデル (AM) と 5 種類の言語モデル (LM) を用意し，AM と LM の組み合わせによって 10 種類の音声認識システムを構築した．AM は HMM モデルで，日本語話し言葉コーパス (CSJ)[9] の全講演 (2702 講演) から学習した syllable モデル (音節 124 種) と triphone モデル (音素 43 種) の 2 種類である．すべての LM は単語

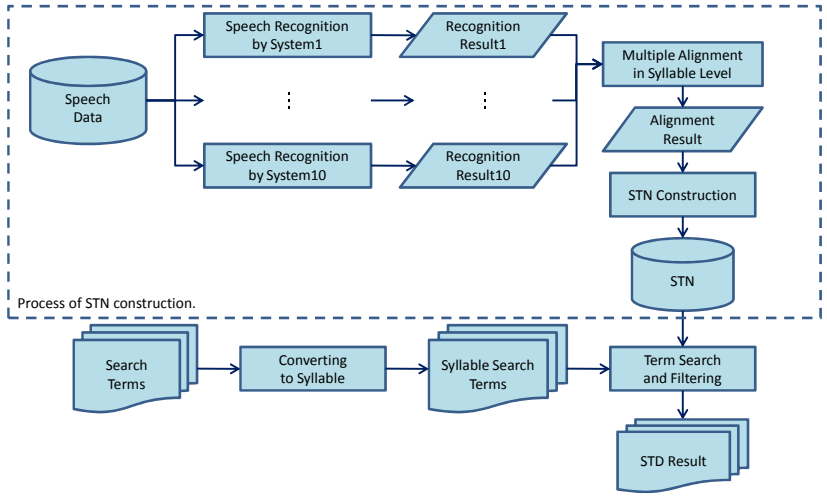

図 1 提案手法の処理の概要
Fig. 1 Overview of STD framework based on Syllble Transition Network.

もしくは文字ベースの trigram モデルである．以下に，LM の詳細を示す．

WBC ：単語ベースの trigram モデル．単語は，漢字と英数字，平仮名，片仮名で構成されている．
例：今回 / の / 実験 / の / 目的

WBH ：単語ベースの trigram モデル．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．
例：こんかい / の / じっけん / の / もくてき

CB ：文字ベースの trigram モデル．文字はすべて平仮名によって構成されている．
例：こ / ん / か / い / の / じ / っ / け / ん / の / も / く / て / き

CSB ：文字系列ベースの trigram モデル．文字系列は数文字の平仮名によって構成されている．
例：こん / かい / の / じっ / け / ん / の / もく / てき

non. ：LM を使用しない．連続音素 (音節) 認識と等価となる．

各 LM は，CSJ の全講演の書き起こしテキストから学習されている．

この 10 種類の認識システムを用い，5 つの大学講義音声に対して音声認識を行った．各認識システムの 1-Best 出力で最も認識率が良かった結果 (1-Best) と，各認識システムの 3-Best 出力を時間同期で連結させたとき (3-Best)，各認識システムの 1-Best 出力を時間同

© 2009 Information Processing Society of Japan

Vol.2009-SLP-79 No.19
2009/12/22

表 1 5 講義音声の音節認識率 [%]
Table 1 Syllble recognition rates for five spoken lectures.[%]

		1-Best		3-Best		10 Recognition Systems	
講義名	duration	Corr.	Acc.	Corr.	Acc.	Corr.	Acc.
オートマトン	40min.	75.81	67.04	78.27	55.42	90.31	-890.14
基礎物理学	36min.	63.24	55.74	65.92	43.31	83.49	-811.66
データベース	37min.	69.32	62.74	72.09	52.41	86.99	-816.47
ディジタル回路	59min.	60.87	46.51	64.87	29.89	77.59	-877.04
プログラミング	58min.	63.87	61.10	65.77	55.06	83.76	-695.98

期で連結させたとき (10 Recognition Systems) の音節認識率を表 1 に示す.

表 1 から，10 種類の音声認識システムを用意することで，ほとんどの講義において 80%以上の音節を認識できていることが分かる．また，単一の認識システムの 3-Best 出力を組み合わせた結果と比較すると，Corr. において大きな差があることが分かる．つまり，単一の認識システムの結果より，複数の認識システムの出力を組み合わせた方が，特定のキーワードを見つけられる可能性が高くなる．しかし，大量の挿入誤りが発生しているため，キーワードの検索において多くの湧き出し誤りが発生する可能性が高い.

3.2 コンフュージョンネットワークの利用

CN は，シンボルの順序関係を保持しながら，複数のシンボル系列を表現する最も効率的な方法といえる．この CN を用いることで，複数の音声認識結果を効率よく組み合わせることが可能となる．CN はヌル遷移を意味する特殊なシンボル “@” を持つ．“@” によって，ノードを飛ばしてシンボル列の検索を行える場合がある．これを利用し，複数の音節系列をうまく組み合わせることができると考えた．しかし，“@” の影響によりシンボル隣接性のチェックが難しくなるといった問題点が残る.

3.3 検索用インデックスの構築方法

認識システムからの 10 種類全ての出力が，音節レベルでアライメントが取れるよう，音節系列に変換される.

認識結果を音節系列に変換した後，ROVER 法[6] を参考に DP マッチングに基づくマルチプルアライメントを行うことで，音節系列間の時間的整合を取る.

整合が取られた音節系列の同一列をアーク群として CN に変換する.

CN は単体の音声認識システムからも得ることが可能である．本稿では，これらの CN と区別するために，複数の音声認識結果を組み合わせた CN を STN と呼称する.

図 2 は STN の構築例を表す．図 2 は各認識システムからの複数の出力の例と，認識システムすべての出力を音節ベースで整合を取った結果，整合が取られた音節系列を STN に変換したイメージを表す．図 2 の “Tri’ は triphone ベースの HMM を，“Syll’ は syllable

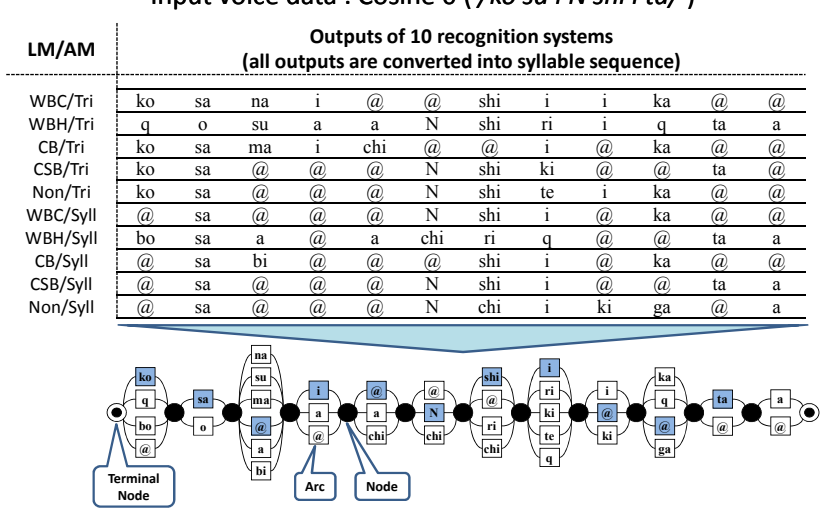

図 2 音節遷移ネットワークの構築例
Fig. 2 Example of converting output of ten recognition systems into Syllble Transition Network.

ベースの HMM を示す.

最終的に，並べられた音節系列は STN に変換される．本稿では，各アークに存在する音節に対する事後確率などの重み付けなどは一切考慮せず，CN へ変換している．図 2 の “@” はヌル遷移を示す.

3.4 用語検索エンジン

検索語は音声認識辞書をもとに音節系列に変換され，用語検索エンジンに入力される.

提案手法である複数の音声認識システムを使用する利点を調査するために，基本的な用語検索方法を採用した．検索エンジンは，STN の検索語に対して，すべての音節系列が一致している音節系列を見つけだす．エンジンは，エントロピーや生起確率などのようなスコアも使用せずに，検索語が STN に存在するかどうかによって判断する.

図 2 では，“Cosine θ (/ko sa i N shi i ta/)” という実際の音声を提案手法に入力した際の例を示している．この入力音声に対して，10 種類の認識システム個々では一つとして正しく認識できていない．しかし，STN に変換することによって，正しく認識できる経路が出現している.

© 2009 Information Processing Society of Japan

しかし，この用語検索方法では STN を利用しても限界が存在する．そこで，STN の検索インデックスとしての潜在的な性能を調査するために，検索語の音節数 “4” に対して “1” つまでのミスマッチを許すような，制約をゆるめた用語検索方法を採用した．例えば，検索語の音節数が 1 から 4 の場合には 1 つまでのミスマッチを，5 から 8 の場合には 2 つまでのミスマッチを許容するというものである．ここで挙げているミスマッチとは，検索語の音節系列に対して次の認識誤りが発生することを指す．

- インデックスの音節系列の一部が異なる音節となってしまう置換誤り
- インデックスの音節系列に余計な音節が入ってしまう挿入誤り
- インデックスの音節系列の一部が認識されていない脱落誤り

以上の 3 種類である．

4. 湧き出し抑制フィルタ

認識誤りへの対処を行うことによって，大量の湧き出し誤り検出が発生してしまう．そこで，STN を構築過程で得られる音節の生起確率や音響尤度，また，認識方法の組合せやエントロピーを用いて湧き出しを抑制するためのフィルタを構成する．

本稿では，認識システムの組合せによる信頼度と，認識システムの多数決による信頼度によってフィルタを構成した．

この信頼度 (Reliability) は，式 (1) によって算出する．

$$Reliability = \frac{R}{N} \tag{1}$$

ここで，N は n 種類の認識システムで共通に出力された音節数を，R は n 種類の認識システムで共通に出力された正解音節数を示す．つまり，n 種類の認識システムで共通に出力された音節の内，正解の割合となる．

4.1 認識システムの組合せによる信頼度

認識システムの組合せによる信頼度とは，2 種類の認識システムの組合せごとに信頼度を算出したものである．認識システムの組合せによる信頼度の一覧を表 2 に示す．

4.2 多数決による信頼度

多数決による信頼度とは，10 種類の認識システムで信頼度を算出したものである．多数決による信頼度の一覧を表 3 に示す．

5. 音声中の検索語検出実験

5.1 音声データ

音声データには，山梨大学工学部コンピュータ・メディア工学科で開講された 5 講演を用いた．音声データは，表 1 の音節認識率を求めたデータと同一のものである (この講義音声は Classroom Lecture Speech Contents(CJLC)[10] のコーパスの一部を含んでいる)．すべての音声が，16kHz と 16bit サンプリングでピンマイクによって録音されている．さらに，講師の半数がスライドの代わりに黒板を使用していたため，これらの音声にはチョーク音などのバックグラウンドノイズが含まれている．黒板を使用することは，講義における音声認識性能を低下させる要素の一つである．使用した音声は，全体で約 3.8 時間のデータとなっ

表 2 認識システムの組合せによる信頼度 [%]
Table 2 Reliability by recognition model's combination.[%]

Combination	Reliability	Combination	Reliability
WBH/Syll : WBC/Tri	87.43	CB/Syll : CSB/Tri	80.86
WBC/Syll : WBH/Tri	85.83	WBC/Tri : CSB/Tri	80.75
WBH/Syll : CSB/Tri	85.63	Non/Syll : WBH/Syll	80.65
WBH/Syll : CB/Tri	84.95	CB/Syll : Non/Tri	80.18
WBH/Syll : Non/Tri	84.64	WBH/Tri : CB/Tri	79.71
CSB/Syll : WBH/Tri	84.62	Non/Syll : CSB/Tri	79.69
WBH/Syll : WBH/Tri	84.56	Non/Tri : CSB/Tri	79.33
CSB/Syll : WBC/Tri	84.51	CB/Syll : CB/Tri	79.12
CB/Syll : WBH/Tri	84.19	WBH/Tri : CSB/Tri	78.70
WBC/Syll : CSB/Tri	84.02	WBC/Syll : Non/Syll	78.69
WBC/Syll : WBC/Tri	83.98	WBH/Syll : CB/Syll	78.53
CB/Syll : WBC/Tri	83.97	Non/Syll : CB/Tri	78.02
WBC/Syll : CB/Tri	83.48	WBC/Syll : WBH/Syll	77.80
WBC/Syll : Non/Tri	82.84	Non/Tri : CB/Tri	77.59
Non/Syll : WBH/Tri	82.27	WBC/Syll : CB/Syll	77.44
WBC/Tri : Non/Tri	82.11	Non/Syll : CSB/Syll	77.35
Non/Syll : WBC/Tri	81.99	WBH/Syll : CSB/Syll	76.29
WBC/Tri : WBH/Tri	81.85	WBC/Syll : CSB/Syll	76.28
CSB/Syll : Non/Tri	81.68	CB/Tri : CSB/Tri	75.69
Non/Tri : WBH/Tri	81.53	Non/Syll : CB/Syll	75.24
CSB/Syll : CSB/Tri	81.36	CB/Syll : CSB/Syll	73.65
WBC/Tri : CB/Tri	81.23	Non/Syll : Non/Tri	66.81
CSB/Syll : CB/Tri	80.94		

表 3 多数決による信頼度 [%]
Table 3 Reliability by voting.[%]

Recognitions	Reliability	Recognitions	Reliability	Recognitions	Reliability
10	94.33	7	85.57	4	73.16
9	91.89	6	82.18	3	70.15
8	88.46	5	78.02	2	69.61

© 2009 Information Processing Society of Japan

表 4 音声中の検索語検出実験結果
Table 4 STD performances on spoken lectures.

Term	In Vocabulary						Out of Vocabulary						The Whole					
Terms	324						105						429					
Index	WORD	10SYL	STN				WORD	10SYL	STN				WORD	10SYL	STN			
Search	Word	Syllble	Tol.1/4	Match	Filter1	Filter2	Word	Syllble	Tol.1/4	Match	Filter1	Filter2	Word	Syllble	Tol.1/4	Match	Filter1	Filter2
Recall	0.43	0.57	0.88	0.64	0.85	0.52	0.00	0.12	0.71	0.27	0.63	0.14	0.32	0.45	0.84	0.55	0.79	0.42
Precision	0.69	0.74	0.09	0.60	0.12	0.35	0.00	0.30	0.20	0.34	0.23	0.15	0.51	0.63	0.12	0.53	0.15	0.30
F-measure	0.48	0.58	0.12	0.54	0.15	0.31	0.00	0.15	0.19	0.26	0.22	0.12	0.35	0.47	0.14	0.47	0.17	0.26
ATWV	0.37	0.42	-18.70	0.15	-12.73	-0.76	-0.01	0.08	-7.72	0.12	-5.60	-0.27	0.28	0.33	-15.89	0.14	-10.90	-0.63

ている．

5.2 検索テストセット

検索語には名詞 unigram と名詞 bigram を用意した．これらの名詞は一般的な用語から専門的な用語まで含まれている．この検索語は，すべてテストデータ内で発話されているものである．検索語は，全体で 428 個 (82 種類) 用意した．内，61 種類の検索語は LM(WBC) に登録された既知語 (IV) である．他の 21 種類は未知語 (OOV) となる．

5.3 評価尺度

評価尺度には，Recall，Precision，F-measure，そして NIST で STD 性能を評価するのに使用されている ATWV(Actual Term-Weighted Value)[11] を用いた．Recall，Precision，F-measure は各検索語の平均となる．以下に，評価式を示す．

$$Recall(t) = \frac{N_{corr}(t)}{N_{true}(t)} \tag{2}$$

$$Precision(t) = \frac{N_{corr}(t)}{N_{corr}(t) + N_{spurious}(t)} \tag{3}$$

$$F - measure(t) = \frac{2 * Recall(t) * Precision(t)}{Recall(t) + Precision(t)} \tag{4}$$

$$ATWV(t) = 1 - (P_{miss}(t) + \beta P_{fa}(t)) \tag{5}$$

$$P_{miss}(t) = 1 - Recall(t), P_{fa}(t) = \frac{N_{spurious}(t)}{Total - N_{true}(t)} \tag{6}$$

N_{corr} は検出された適合検索語の出現数を表し，$N_{spurious}$ は検出された検索語の総数を表す．N_{true} は音声データ中に本来存在する検索語の出現総数を表す．Total は音声データの持続時間 (秒) を表す．β は本稿では 10^3 に設定している．

5.4 検索用インデックスと検出方法

STD の検索性能の比較のため，3 種類の検索用インデックスを用意した．

- **WORD：** WBC/Tri と WBC/Syll の Best1 出力のみから得られる単語ベースのインデックス
- **10SYL：** 10 種類の Best1 出力を音節に変換した結果から得られる，10 種類の音節レベルインデックス
- **STN：** 提案手法の STN に基づいたインデックス

また，検索語の検出方法はインデックスによって異なる．WORD からの検出は，検索語を単語ベースで検出する．10SYL からは，検索語の音節レベルでの完全一致によって検出を行い，10 種類の音節インデックスのどれか 1 つにでも検索語が出現していれば適合として判断する．STN からの検出方法は以下の 3 種類用意した．なお，STN の特性であるヌル遷移は機能する．

- **Tol.1/4：** 検索語の音節数 4 に対して，誤りを 1 つまで許容する音節系列の一致による検出
- **Match：** 検索語の音節系列の完全に一致による検出
- **Filter1：** Tol.1/4 の検出結果に対して，認識システムの組合せによる信頼度の上位 5 件までの組合せであれば適合音節として判断する
- **Filter2：** Tol.1/4 の検出結果に対して，多数決で過半数で認識しているのであれば適合音節として判断する

なお，フィルタはキーワードの探索時に適応する．

5.5 実験結果

表 4 に IV，OOV，検索語全体での STD の性能を示す．表 4 に示されるように，STN を利用することで，他のインデックスと比較して Recall を改善することができた．しかし，誤検出が増加しているため，他のインデックスと比較して Precision と F-measure，ATWV

© 2009 Information Processing Society of Japan

は低かった．

WORD と 10SYL は OOV を検出するのが難しかったが，STN(Match) は良い性能を示すことができた．Recall と Precision，F-measure，ATWV のすべての評価項目に対して，最も良い検索性能を示している．

IV に対しては，10SYL の検出用インデックスが最も良い検索性能を示した．これらのことから IV に対しては 10SYL を利用し，OOV に対しては STN を利用するといった組合せ手法を導入することで，検索性能を向上させることが可能となる．

STN(Tol.1/4) の結果が示すように，STN の潜在的な検索性能はすべての検索語に対して非常に有効であった．全体の Recall は 0.8 を超え，OOV に関しても 0.7 を超えている．さらに，OOV に対しては，F-measure において 10SYL を超えている．すなわち，OOV に対しては STN は非常に頑健な検索用インデックスであることが示された．

STN(Filter1) や STN(Filter2) の結果から，簡単な湧き出し抑制フィルタを STN(Tol.1/4) にかけることによって，F-masure や ATWV をわずかではあるが改善することができた．しかし，これらのフィルタは十分な機能を果たしていない．とくに，STN(Filter2) では Recall が大幅に低下している．

しかし，なんらかのフィルタリングを行うことで，湧き出しを抑えた検出が可能になることが示された．また，IV と OOV で F-masure や ATWV の変化に違いがあることから，検索語に応じて適切なフィルタリングを行うことで，STN を利用した STD の性能は改善されることが予想される．

6. ま と め

本稿では，複数の音声認識システムの出力を利用することで，講義音声に対する STD 性能を向上させる手法を提案した．

提案手法では，複数の音声認識システムの出力を，音節ベースのコンフュージョンネットワークに変換した検索用インデックスを構築する．この検索用インデックスは未知検索語に対して，STD 性能を向上させるのに有効であった．また，STN の潜在的な検索性能はどのような検索語に対しても有効であることが判明した．

本稿で行った STN のフィルタはあまり効果的なものではなかった．しかし，何らかのパラメータを利用することによって，湧き出し誤りを抑えた STD が可能となる可能性が示された．

今後は，コンフュージョンネットワークの生起確率や音響尤度などのスコアを導入し，STD 性能の向上を図る予定である．

参 考 文 献

1) D.Vergyri, I.Shafran, A.Stolcke, R.R. Gadde, M.A. adn B.Roark, and W.Wang, “The SRI/OGI 2006 spoken term detection system,” in *Proc. of the INTERSPEECH 2007*, 2007, pp. 2393–2396.
2) S.Meng, J.Shao, R.P. Yu, J.Liu, and F.Seide, “Addressing the out-of-vocabulary problem for large-scale chinese spoken term detection,” in *Proc. of the INTERSPEECH 2008*, 2008, pp. 2146–2149.
3) K.Iwata, K.Shinoda, and S.Furui, “Robust spoken term detection using combination of phone-based and word-based recognition,” in *Proc. of the INTERSPEECH 2008*, 2008, pp. 2195–2198.
4) J.Gao, J.Shao, Q.Zhang, Q.Zhao, and Y.Yan, “Spoken term detection using dynamic match subword confusion network,” in *Proc. of the 2008 4th ICNC*, 2008, pp. 250–254.
5) 堀 貴明，リー・ハセリントン，ティモシー・ヘイゼン，ジェームズ・グラス，“コンフュージョンネットワークを用いたオープン語彙発話検索法とその評価”，信学技法 SP11-8, 2007，pp.43–48.
6) J.G. Fiscus, “A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER),” in *Proc. of the 1997 IEEE Workshop on Automatic Speech Recognition adn Understanding (ASRU’97)*, 1997, pp. 347–354.
7) T.Utsuro, Y.Kodama, T.Watanabe, H.Nishizaki, and S.Nakagawa, “An empirical study on multiple lvcsr model combination by machine learning,” in *Proc. of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, 2004, pp. 13–16.
8) A.Lee, T.Kawahara, and K.Shikano, “Julius — an open source real-time large vocabulary recognition engine,” in *Proc. European Conference on Speech Communication and Technology (EUROSPEECH 2001)*, 2001, pp. 1691–1694.
9) K.Maekawa, “Corpus of spontaneous japanese: Its design and evaluation,” in *Proc. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.
10) M.Tsuchiya, S.Kogure, H.Nishizaki, K.Yamamoto, K.Ota, and S.Nakagawa, “Developing corpus of japanese classroom lecture speech contents,” in *Proc. of the 6th edition of the Language Resources and Evaluation Conference (LREC)*, 2008.
11) NIST. (2006) The spoken term detection (STD) 2006 evaluation plan. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf

© 2009 Information Processing Society of Japan