

複数音声認識システムに基づいた音声中の検索語 検出手法の検討と CSJ テストコレクションでの評価

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に基づいた音声中の検索語検出手法 (STD) について述べる。本研究では、単独・複数の音声認識システムの出力から合計 6 種類のインデックスを構築し、それぞれのインデックスに対して STD の性能を比較した。インデックスの種類には、複数の音声認識システムの 1-Best 出力、10-Best 出力をそのまま利用するものや、出力を音節レベルにてコンフュージョンネットワーク化したものがある。これらの STD 用のインデックスを、日本語話し言葉コーパス (CSJ) に対するテストコレクション上で評価した。実験の結果、単独の音声認識システムの N-Best 出力を利用した場合と比較し、複数の認識システムの出力を利用した方が、STD 性能を改善させることに有効であることが示された。特に、未知語検出においては、複数の音声認識システムを組み合わせる利用することが、検出率向上に効果的であることが分かった。

Spoken Term Detection Using Multiple Recognition Systems and Evaluation with CSJ Test Collection

NATORI SATOSHI,^{†1} NISHIZAKI HIROMITSU^{†2}
and SEKIGUCHI YOSHIHIRO ^{†2}

This paper describes a spoken term detection (STD) technique by using multiple speech recognition systems' outputs. In this research, we built six kinds of indices from the single or multiple recognizers' outputs, and empirically evaluated the detection performance of these indices on the CSJ test collection for STD. To the kind of index, one index is made from ten sub-indices, each of which is from 1-Best output of a single recognizer. Another index is represented as syllable confusion network form which is constructed using 10 recognizers' outputs. Experimental results showed that using multiple recognizers' outputs improved the STD performance compared with N-Best outputs from a single recognizer's output. Furthermore, the syllable based confusion network made by 10 recognizers made the term detection rate to be improved especially on out-of-vocabulary term detection from much speech data.

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、今この瞬間も、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 の目的は SDR とは異なり、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⁾⁶⁾⁷⁾。

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によって STD 性能を向上させる手法について述べる。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を基に、数種類の検索用インデックスを構築し検索語の検出を行う。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

^{†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

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非常に効果的であることが知られている．例えば，Fiscus⁸⁾ 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⁹⁾ 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．

STD 性能の比較結果では，単一の音声認識システムの N-Best 出力を利用した場合と比較して，複数の音声認識システムの出力を利用することが，STD 性能を向上させることに有効であることが示された．特に未知語検出においては，複数の音声認識システムの出力を組み合わせることによって，探索の幅が大幅に広がることが示された．

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た，音声中の検索語を検出手法の検討を行う．

本稿では，複数の音声認識システムの出力を利用し，STD のための数種類の検索用インデックスを構築する．構築した数種類の検索用インデックスの STD 性能を比較する．

検索用のインデックスには，我々が提案してきた音節遷移ネットワーク (STN : Syllable Transition Network)¹⁰⁾ 以外に，複数の音声認識結果を単純に利用したインデックスと，単一の音声認識システムの 10-Best 出力をコンフュージョンネットワークとして統合したインデックスを用意した．これらのインデックスにおいて，複数の音声認識システムを利用することによって STD 性能が向上することを検証する．参考文献⁸⁾⁹⁾ 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．

3. 検索用インデックスの構築

3.1 複数の音声認識システム

本研究では，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. 4.1.3¹¹⁾ (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)¹²⁾ 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 種類である．

すべての LM は，単語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．

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から学習している．

2009 年 10 月に，CSJ の STD 用テストコレクションの β バージョンが公開されたが¹⁴⁾，文献でも述べられているように制定過程の段階である．現在のところ，AM，LM をオープンな条件で学習するように，以下のように学習，認識を行うことが検討されている．

- 各モデルの学習は奇数/偶数講演に分けて学習を行う
- 音声認識対象の講演 ID が奇数講演の場合は偶数講演から学習したモデル，偶数講演の場合は奇数講演から学習したモデルによって認識を行う
- WBC の LM は奇数/偶数講演に共通に出現する 26,700 種の語彙によって学習を行う
この STD 用の音声認識条件に合うよう，音声認識対象の講演ごと認識に使用するモデルを切り替え音声認識を行った．

この 10 種類の音声認識システムを用い，CSJ コア講演音声を認識した．

LM に WBC を用いた場合の平均単語認識率を表 1 に示す．また，10 種類の音声認識シ

表 1 CSJ コア講演音声の平均単語認識率 [%]

Table 1 Word Recognition Rates for CSJ Core Lectures.[%]

認識システム	Corr.	Acc.
WBC/Tri	76.68	71.93
WBC/Syl	67.54	64.10

表 2 CSJ コア講演音声の平均音節認識率 [%]

Table 2 Syllable Recognition Rates for CSJ Core Lectures.[%]

認識システム数	1-Best		10-Best	
	Corr.	Acc.	Corr.	Acc.
1	86.46	83.01	89.96	44.88
10	94.19	-11.67	96.25	-241.04

システムの 1-Best 出力で最も認識率が良かった結果と、単一の音声認識システムの 10-Best 出力を時間同期で連結させたときに最も認識率が良かった結果、10 種類の音声認識システムの 1-Best 出力を時間同期で連結させたとき、10 種類の音声認識システムの 10-Best 出力を時間同期で連結させたときの音節認識率を表 2 に示す。

表 2 では、10 種類の音声認識システムを用意することで、94%という高い Corr. を示すことができている。また、単一の認識システムの 10-Best 出力を組み合わせた結果と、10 種類の音声認識システムの出力を組み合わせた結果を比較すると、10 種類の音声認識システムの出力を組み合わせた結果の方が Corr. が良い。つまり、単一の音声認識システムの出力より、複数の音声認識システムの出力を組み合わせの方が、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くなる。

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

検索用のインデックスは表 3 に示す 6 種類を構築した。STN(1-Best) 以外の全てのイン

表 3 検索用インデックスとサブインデックス数

Table 3 Indices for STD and the number of their sub-indices.

検索用インデックス	サブインデックス数	サブインデックスの構成
WRD(1-Best)	2	WBC/Tri, WBC/Syl の 1-Best の単語認識結果
WRD(10-Best)	20	WBC/Tri, WBC/Syl の 1~10-Best の単語認識結果
10SYL(1-Best)	10	10 種の 1-Best の音節認識結果
10SYL(10-Best)	100	10 種の 1~10-Best の音節認識結果
10CN(10-Best)	10	1 種の 1~10-Best の音節認識結果を CN 化 × 10 種
STN(1-Best)	1	10 種の 1-Best の音節認識結果を CN 化

デックスは、複数のサブインデックスから構成されている。

3.3.1 WRD/10SYL

WRD, 10SYL は認識結果をそのまま用いた単純なインデックスである。WRD/10SYL(1-Best) と WRD/10SYL(10-Best) の違いは、インデックスを構成するサブインデックスの数である。

1-Best では認識結果の 1-Best のみをサブインデックスとしているのに対し、10-Best では認識結果の 1 から 10-Best までの各 nBest 認識結果をそれぞれサブインデックスとしている。即ち、WRD(1-Best) は 2 種類のサブインデックスから構成されているのに対し、WRD(10-Best) では 20 種類のサブインデックスから構成されている。同様に、10SYL(1-Best) は 10 種類のサブインデックスから構成されているのに対し、10SYL(10-Best) では 100 種類のサブインデックスから構成されている。

3.3.2 10CN/STN

10CN は、単一の音声認識システムの 1~10-Best 出力を CN 化したサブインデックスから構成されている。即ち、10CN は 10 種類のサブインデックスから構成されている。

STN は 10 種類の音声認識システムの 1-Best 出力を CN 化したものである。

10CN と STN の違いは、CN 化する音節認識結果が、10 種類の 1-Best 出力であるか 1 種類の 10-Best 出力であるかの違いである。

このとき、10 個の認識結果は全ての出力が、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⁸⁾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。

STN の構築は図 1 に示すように行われる。図 1 は STN の構築例を表しており、複数の出力の例と、認識結果全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

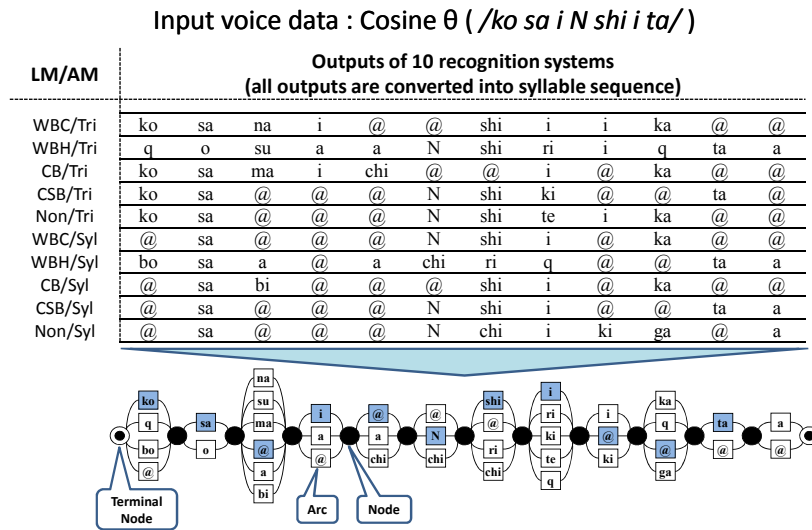


図1 音節遷移ネットワークの構築例

Fig. 1 Example of converting output of 10 recognition result into Syllble Transition Network.

列を CN に変換したイメージを表す。

10CN のサブインデックスを構築する手順も STN と基本的に同じであり、10 種類の認識結果が 1 種 10-Best の認識結果に変わるのみである。

最終的に、並べられた音節系列は CN/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/STN へ変換している。図1の“@”はヌル遷移を示す。

4. 用語検索エンジン

4.1 用語検索方法

各インデックスからの検索方法は以下の3種類を用いた。

Match : 文字列/音節系列の完全一致

Tol.1/4 : 検索語の音節数4に対して、ミスマッチを1つまで許容する音節系列の一致

Filter : Tol.1/4 の検出結果に対して湧き出し誤り抑制フィルタを適応

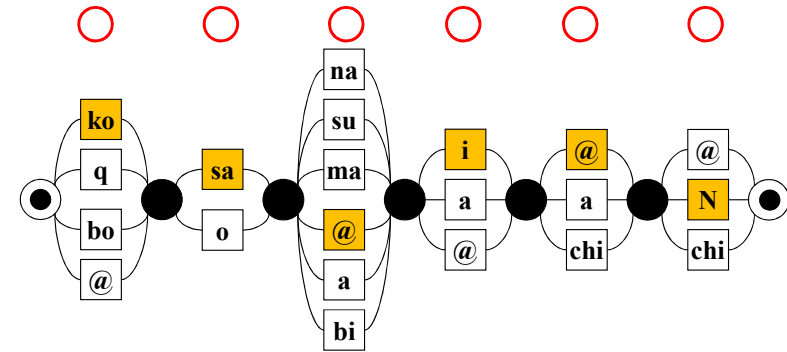


図2 コンフュージョンネットワークからの完全一致による検索例

Fig. 2 Example of perfect term matching on Confusion Network.

4.1.1 Match による用語検索

WRD, 10SYL は認識結果をそのまま用いた単純なインデックスである。このため、Match による用語検出には単純な文字列一致を利用した。

10CN, STN からの Match による用語検索例を図2に示す。

検出エンジンは、検出用語に対して音節系列が完全に一致している音節系列を見つけ出す。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。

4.1.2 Tol.1/4 による用語検索

完全一致による用語検索方法では図3の(a)に示すように、CN を利用したインデックスにおいても限界が存在する。

そこで、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行う。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。

- 検索用インデックスの音節系列の一部が異なる音節になってしまう置換誤り
- 検索用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
- 検索用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

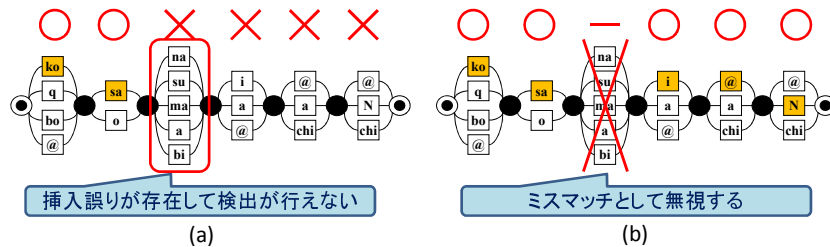


図3 コンフュージョンネットワークとミスマッチを許容した検索例

Fig. 3 Example of term matching when there is a mismatch between CN and term.

以上の3種類である。

この検索語の音節のミスマッチを許すことによって、図3の(b)に示すように検索の幅が広がり、より多くの用語が検出されるようになる。

4.1.3 Filter による用語検索

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、検索用インデックスにSTNを用いる場合、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誤りを抑制するためのフィルタを構成することで、湧き出し誤り検出を抑えることが出来ると考えられる。

今回使用した湧き出し抑制フィルタは単純なものであり、検索語の検出時に音節が一致 or スルの場合において、一つの検出箇所の多数決信頼度¹⁰⁾の相乗平均が既知語の場合は0.20以下、未知語の場合は0.25以下をフィルタリングする。さらに、一致した音節 or スルを認識していたシステムが全て1の場合にはフィルタリングを行う。

5. 音声の中の検索語検出実験

5.1 STD テストコレクション

音声データには、CSJのコア講演音声を用いた。検索語には、CSJのコア講演用テストコレクション¹⁴⁾を用いた。テストコレクションの概要を以下に示す。

- 既知検索語彙 (IV) 数: 50 種, 769 箇所
- 未知検索語彙 (OOV) 数: 28 種, 89 箇所

現在、テストコレクションに記載されている未知検索語は、既知語を含むものとなっている。これは、テストコレクション制定時から言語モデルと学習語彙が変更されたためである。

表4 検索用インデックスと検索方法の組み合わせ
Table 4 Combination of index and search method.

	Match	Tol.1/4	Filter
WRD(1-Best)	○	-	-
WRD(10-Best)	○	-	-
10SYL(1-Best)	○	○	-
10SYL(10-Best)	○	○	-
10CN(10-Best)	○	○	-
STN(1-Best)	○	○	○

る。本稿では、このテストコレクションから既知語のテストコレクションを除いたものを用いたため、OOVの種類が28種となっている。

5.2 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)¹⁵⁾を用いた。評価尺度は各検索語の平均となる。以下に、評価式を示す。

$$Recall(t) = \frac{N_{corr}(t)}{N_{true}(t)} \quad (1)$$

$$Precision(t) = \frac{N_{corr}(t)}{N_{corr}(t) + N_{spurious}(t)} \quad (2)$$

$$F-measure(t) = \frac{2 * Recall(t) * Precision(t)}{Recall(t) + Precision(t)} \quad (3)$$

$$ATWV(t) = 1 - (P_{miss}(t) + \beta P_{fa}(t)) \quad (4)$$

$$P_{miss}(t) = 1 - Recall(t), P_{fa}(t) = \frac{N_{spurious}(t)}{Total - N_{true}(t)} \quad (5)$$

N_{corr} は検出された適合検索語の出現数を表し、 $N_{spurious}$ は誤検出された検索語の出現数を表す。 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)を表し、 144×10^3 を設定した。 β は本稿では 14×10^3 を設定した。

5.3 検索用インデックスと検出方法

検索用インデックスと検索語の検出方法の対応を表4に示す。

WRD, 10SYL, 10CNではインデックス全体で湧き出し誤りと適合検出を評価しており、異なるサブインデックスで同じ位置を検出した場合は、一箇所の検出位置として評価している。

表 5 音声中の検索語検出実験結果 (IV)

Table 5 STD performances on spoken lectures (IV).

Search	Match				Tol.1/4				Filter			
Index	Recall	Precision	F-measure	ATWV	Recall	Precision	F-measure	ATWV	Recall	Precision	F-measure	ATWV
WRD(1-Best)	0.59	0.91	0.69	0.54	-				-			
WRD(10-Best)	0.61	0.90	0.70	0.55	-				-			
10SYL(1-Best)	0.68	0.90	0.75	0.56	0.84	0.34	0.41	-20.80	-			
10SYL(10-Best)	0.72	0.88	0.77	0.54	0.88	0.26	0.34	-42.15	-			
10CN	0.73	0.77	0.75	0.32	0.90	0.20	0.26	-83.26	-			
STN(1-Best)	0.70	0.82	0.73	0.32	0.84	0.17	0.23	-60.84	0.84	0.18	0.24	-53.78

表 6 音声中の検索語検出実験結果 (OOV)

Table 6 STD performances on spoken lectures (OOV).

	Match				Tol.1/4				Filter			
Index	Recall	Precision	F-measure	ATWV	Recall	Precision	F-measure	ATWV	Recall	Precision	F-measure	ATWV
WRD(1-Best)	0.00	0.00	0.00	0.00	-				-			
WRD(10-Best)	0.00	0.00	0.00	0.00	-				-			
10SYL(1-Best)	0.32	0.43	0.36	0.32	0.70	0.29	0.31	-6.03	-			
10SYL(10-Best)	0.40	0.46	0.42	0.40	0.76	0.26	0.27	-12.64	-			
10CN(10-Best)	0.41	0.46	0.42	0.40	0.76	0.16	0.17	-23.30	-			
STN(1-Best)	0.40	0.41	0.39	0.37	0.83	0.09	0.12	-19.23	0.82	0.09	0.13	-15.45

5.4 実験結果

表 5 に IV, 表 6 に OOV の STD 性能を示す.

実験結果から, 検索用のインデックスを構築する元となる認識結果が多い方が Recall は高くなる. 即ち, 単一の音声認識システムを利用するより, 複数の音声認識システムの出力を利用することによって, STD 性能が改善されることが示された.

OOV においては STN(1-Best)/Match と 10CN(10-Best)/Match の差が僅かであり, また, 検出方法が Tol.1/4 では STN(1-Best) が最も良くなっている. つまり, 異なる認識システムを組合せることによって探索の幅が広がることが示された. このことから, 未知検索語検出においては, STN が有効であることが改めて示された. さらに, STN を 10-Best から構築することによって, Recall の向上が図れると考えられる.

フィルタリングの効果については, ある程度の湧き出し誤り検出が抑えられ, F-measure, ATWV は向上している. しかし, やはり十分なフィルタリングが行えているとは言えない結果となった. この原因としては, フィルタリングに用いた情報が, 講義音声の結果をもと

に得られたものであるということが考えられる.

6. ま と め

本稿では, 単一の音声認識システムを利用するより, 複数の音声認識システムの出力を利用することが, STD 性能を向上させることに有効であることを示した.

特に, 未知検索語検出においては, 複数の音声認識システムの出力を組み合わせることによって大幅に探索の幅が広がることが示された.

しかしながら, 多くのサブインデックスを使用しているため, 湧き出し誤り検出が多く発生している. 湧き出し誤り検出を抑えるために, STN では単純なフィルタを適応したが, あまり効果はなかった. しかし, 湧き出し誤り検出を抑える可能性は示された.

今後は, 検索用のインデックスの改善とインデックスからの検索パラメータの調査という 2 つの方向で進めていく. これらには, STN(10-Best) を構築することや Julius の CN 出力, 認識結果から得られる音響尤度を利用することを検討している.

参 考 文 献

- 1) D.Wang, S.King, and J.Frankel, “Stochastic Pronunciation Modelling for Spoken Term Detection,” in *Proc. of the INTERSPEECH 2009*, 2009, pp. 2135–2138.
- 2) T.J. Hazen, W.Shen, C.M. White, “Query-By-Example Spoken Term Detection Using Phonetic Posteriorgram Templates,” in *Proc. of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU’09)*, 2009, pp. 421–426.
- 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) 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.
- 5) J. Tejedor, D.Wang, S.King, J.Frankel, and J.Colas, “A Posterior Probability-Based System Hybridisation and Combination for Spoken Term Detection,” in *Proc. of the INTERSPEECH 2009*, 2009, pp. 2131–2134.
- 6) C.Parada, A.Sethy, B.Ramabhadran, “Query-By-Example Spoken Term Detection For OOV Terms,” in *Proc. of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU’09)*, 2009, pp. 404–409.
- 7) H.Y. Lee, Y.L. Tang, H.Tang, and L.S. Lee, “Spoken Term Detection from Bilingual Spontaneous Speech Using Code-switched Lattice-based Structures for Word and Subword Units,” in *Proc. of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU’09)*, 2009, pp. 410–415.
- 8) 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 and Understanding (ASRU’97)*, 1997, pp. 347–354.
- 9) 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.
- 10) 名取賢, 西崎博光, 関口芳廣, “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 Vol.2009-SLP-079, 2009.
- 11) 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.
- 12) K.Maekawa, “Corpus of spontaneous Japanese: Its design and evaluation,” in *Proc. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.
- 13) 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.
- 14) 伊藤慶明, 西崎博光, 胡新輝, 南條浩輝, 秋葉友良, 相川清明, 河原達也, 中川聖一, 松井知子, 山下洋一, “音声中の検索語検出のためのテストコレクション構築 -中間報告-”, 情報処理学会研究報告 Vol.2009-SLP-078, 2009.
- 15) 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>

音節認識性能

10種類の音声認識システムの音節認識率[%]				
	1Best		10Best	
	Corr	Acc	Corr	Acc
WBC/Tri	86.46	83.01	89.96	44.88
WBH/Tri	86.27	81.42	89.95	35.06
CB/Tri	81.83	77.42	85.99	41.74
CSB/Tri	85.66	80.96	89.26	37.16
non/Tri	71.00	51.20	74.56	21.06
WBC/Syl	79.11	76.35	84.19	35.73
WBH/Syl	79.32	75.83	84.29	29.90
CB/Syl	73.84	71.18	79.47	42.10
CSB/Syl	78.58	75.36	83.55	33.03
non/Syl	63.68	45.43	67.96	21.57

未知検索語に対するSTD性能の内訳

未知検索語における音節インデックス(10SYL)のサブインデックスに対するSTD性能の内訳【10SYL:OOV】

	1Best								10Best							
	Match				Tol.1/4				Match				Tol.1/4			
	Recall	Precision	F値	ATWV	Recall	Precision	F値	ATWV	Recall	Precision	F値	ATWV	Recall	Precision	F値	ATWV
WBC/Tri	0.0537	0.1429	0.0661	0.0537	0.3476	0.1440	0.1271	-2.3157	0.1053	0.2143	0.1255	0.1053	0.3578	0.1245	0.1145	-3.9791
WBH/Tri	0.1725	0.2857	0.2036	0.1725	0.3952	0.1612	0.1673	-2.3445	0.2278	0.3571	0.2629	0.2278	0.4411	0.1542	0.1653	-4.5348
CB/Tri	0.1005	0.2143	0.1270	0.1005	0.5248	0.3419	0.3120	-1.6280	0.1704	0.2500	0.1941	0.1704	0.5987	0.3111	0.2999	-3.4570
CSB/Tri	0.1743	0.2857	0.2087	0.1743	0.5506	0.3391	0.3129	-2.0814	0.1885	0.2857	0.2200	0.1885	0.6246	0.3124	0.3082	-3.8062
non/Tri	0.0862	0.1786	0.1048	0.0862	0.3763	0.2137	0.2033	-1.6377	0.1021	0.1786	0.1190	0.1021	0.4668	0.1978	0.2107	-2.6479
WBC/Syl	0.0000	0.0000	0.0000	0.0000	0.3074	0.1554	0.1279	-1.9739	0.0071	0.0357	0.0119	0.0071	0.3356	0.1283	0.1219	-3.6784
WBH/Syl	0.0434	0.1429	0.0620	0.0434	0.3027	0.1368	0.1317	-2.0966	0.0434	0.1429	0.0620	0.0434	0.3168	0.1011	0.1119	-4.2771
CB/Syl	0.0857	0.1429	0.1026	0.0857	0.3735	0.1975	0.1842	-1.3627	0.0976	0.1429	0.1133	0.0976	0.4640	0.1957	0.1971	-3.0604
CSB/Syl	0.0621	0.1429	0.0826	0.0621	0.3535	0.1823	0.1666	-1.8202	0.1031	0.2143	0.1257	0.1031	0.3956	0.1484	0.1560	-3.5386
non/Syl	0.0280	0.1071	0.0440	0.0245	0.2071	0.0773	0.0694	-1.2374	0.0423	0.1071	0.0589	0.0388	0.3535	0.1461	0.1363	-2.1639

未知検索語における10種類のCNインデックス(10CN(10-Best))のサブインデックスに対するSTD性能の内訳【10CN(10-Best):OOV】

	Match				Tol.1/4			
	Recall	Precision	F値	ATWV	Recall	Precision	F値	ATWV
WBC/Tri	0.5882	0.9402	0.7002	0.5162	0.7752	0.3727	0.4170	-17.5003
WBH/Tri	0.5882	0.8698	0.6790	0.4968	0.7716	0.3423	0.3911	-21.3854
CB/Tri	0.2865	0.8064	0.3965	0.2496	0.6817	0.3726	0.3713	-16.8453
CSB/Tri	0.4302	0.8564	0.5441	0.3699	0.7557	0.3561	0.3966	-20.6974
non/Tri	0.0647	0.3788	0.1060	0.0550	0.4714	0.4533	0.3146	-16.0345
WBC/Syl	0.4450	0.8236	0.5462	0.3264	0.6098	0.2924	0.3127	-19.1803
WBH/Syl	0.4528	0.8254	0.5537	0.3186	0.6186	0.2864	0.3096	-21.2192
CB/Syl	0.1606	0.5931	0.2373	0.1198	0.5183	0.3160	0.2855	-13.4809
CSB/Syl	0.3154	0.7770	0.4144	0.2318	0.5938	0.2804	0.3040	-19.9157
non/Syl	0.0337	0.2800	0.0560	0.0317	0.3403	0.4390	0.2544	-8.4822

既知検索語に対するSTD性能の内訳

既知検索語における文字列インデックス(WRD)のサブインデックスに対するSTD性能の内訳【WRD:IV】

	1Best				10Best			
	Recall	Precision	F値	ATWV	Recall	Precision	F値	ATWV
WBC/Tri	0.5412	0.9360	0.6613	0.5237	0.5701	0.9257	0.6824	0.5429
WBC/Syl	0.4106	0.8132	0.5163	0.3794	0.4380	0.7925	0.5380	0.3855

既知検索語における音節インデックス(10SYL)のサブインデックスに対するSTD性能の内訳【10SYL:IV】

	1Best								10Best							
	Match				Tol.1/4				Match				Tol.1/4			
	Recall	Precision	F値	ATWV	Recall	Precision	F値	ATWV	Recall	Precision	F値	ATWV	Recall	Precision	F値	ATWV
WBC/Tri	0.5463	0.9686	0.6712	0.5210	0.7179	0.4989	0.5064	-6.3954	0.5810	0.9608	0.6977	0.5440	0.7564	0.4304	0.4653	-10.5554
WBH/Tri	0.5362	0.8987	0.6471	0.5187	0.7125	0.4834	0.4927	-6.7431	0.5759	0.8752	0.6724	0.5312	0.7609	0.4213	0.4596	-11.5485
CB/Tri	0.2242	0.7951	0.3243	0.2126	0.6222	0.4999	0.4457	-6.2753	0.2731	0.8125	0.3822	0.2478	0.6786	0.4137	0.4052	-11.1642
CSB/Tri	0.3779	0.8545	0.4926	0.3623	0.6937	0.5080	0.4955	-6.5558	0.4084	0.8505	0.5176	0.3676	0.7501	0.4176	0.4528	-11.8879
non/Tri	0.0403	0.2985	0.0684	0.0384	0.3951	0.4344	0.2948	-7.6069	0.0549	0.3585	0.0911	0.0530	0.4595	0.4492	0.3151	-13.2015
WBC/Syl	0.4006	0.8293	0.5112	0.3676	0.5400	0.4487	0.4068	-5.7740	0.4335	0.8113	0.5384	0.3771	0.6042	0.4023	0.3905	-10.5772
WBH/Syl	0.4080	0.8301	0.5132	0.3575	0.5706	0.4282	0.4138	-5.9204	0.4399	0.8466	0.5428	0.3641	0.6105	0.3785	0.3774	-11.1854
CB/Syl	0.1056	0.4592	0.1601	0.0881	0.4226	0.4093	0.3055	-5.1855	0.1537	0.6033	0.2280	0.1304	0.5076	0.3600	0.3091	-9.7910
CSB/Syl	0.2660	0.7355	0.3642	0.2329	0.5204	0.4420	0.3901	-5.6362	0.3033	0.7433	0.3996	0.2567	0.5859	0.3621	0.3645	-11.1691
non/Syl	0.0174	0.1800	0.0305	0.0155	0.2702	0.4378	0.2237	-4.0700	0.0324	0.2600	0.0536	0.0305	0.3364	0.4392	0.2571	-7.4750

既知検索語における10種類のCNインデックス(10CN(10-Best))のサブインデックスに対するSTD性能の内訳【10CN(10-Best):IV】

	Match				Tol.1/4			
	Recall	Precision	F値	ATWV	Recall	Precision	F値	ATWV
WBC/Tri	0.5882	0.9402	0.7002	0.5162	0.7752	0.3727	0.4170	-17.5003
WBH/Tri	0.5882	0.8698	0.6790	0.4968	0.7716	0.3423	0.3911	-21.3854
CB/Tri	0.2865	0.8064	0.3965	0.2496	0.6817	0.3726	0.3713	-16.8453
CSB/Tri	0.4302	0.8564	0.5441	0.3699	0.7557	0.3561	0.3966	-20.6974
non/Tri	0.0647	0.3788	0.1060	0.0550	0.4714	0.4533	0.3146	-16.0345
WBC/Syl	0.4450	0.8236	0.5462	0.3264	0.6098	0.2924	0.3127	-19.1803
WBH/Syl	0.4528	0.8254	0.5537	0.3186	0.6186	0.2864	0.3096	-21.2192
CB/Syl	0.1606	0.5931	0.2373	0.1198	0.5183	0.3160	0.2855	-13.4809
CSB/Syl	0.3154	0.7770	0.4144	0.2318	0.5938	0.2804	0.3040	-19.9157
non/Syl	0.0337	0.2800	0.0560	0.0317	0.3403	0.4390	0.2544	-8.4822