情報処理学会論文誌 Vol.55 No.3 1201–1211 (Mar. 2014)

音声中の任意検索語検出のための未知語区間推定に基づく選択的インデックス統合法

神田 直之[1,2,a)] 糸山 克寿[1] 奥乃 博[1]

受付日 2013年6月11日, 採録日 2013年12月4日

概要：本研究では音声検索語検出のために複数の音声認識器から出力された認識結果を統合する手法において，未知語区間推定結果に基づいてインデックスを選択的に統合することで，検出精度の劣化を抑えつつインデックスサイズを削減する手法について提案する．提案する手法は 4 種類の音声認識器から得られた出力を 1 つのネットワークへと統合する．その際，未知語区間推定結果に基づきネットワーク中の有効なアークの選択や，インデックスに用いるサブワード単位の選択を行うことで，冗長なインデックスを削減する．日本語話し言葉コーパスを用いた評価の結果，提案法によって，検出精度の劣化を 1.4 ポイントに抑えたうえで音素 Transition Network から 22.7%のインデックスが削減できることを確認した．単一の音声認識結果から作成した音素単位のネットワークと比較した場合，提案法では，インデックスの統合による検出精度向上の効果（既知語で 16.3%，未知語で 16.0%の検出エラー削減）を保ちながら，単一の音声認識結果に基づくインデックスと同等以下の大きさまでインデックスサイズを抑えることができた．

キーワード：音声検索語検出，キーワードスポッティング，未知語検出，未知語区間推定

Selective Index Combination Method Based on Out-of-vocabulary Region Estimator for Open-vocabulary Spoken Term Detection

Naoyuki Kanda[1,2,a)] Katsutoshi Itoyama[1] Hiroshi G. Okuno[1]

Received: June 11, 2013, Accepted: December 4, 2013

Abstract: In this paper, a novel index combination method for spoken term detection is proposed. In our method, outputs from four different recognizers are combined into one confusion network. A novel index-selection method for the multiple index-combination method is then used to suppress the increase of the index size. Two methods are proposed to reduce index size: (1) arc selection and (2) unit selection, both of which are based on an Out-of-Vocabulary (OOV)-region estimator score. Experimental results with Japanese lecture recordings, Corpus of Spontaneous Japanese, showed that the index-selection method achieved a 22.7% reduction of index size of the best confusion network with only 1.4 points loss of its high accuracy. Compared with the best phoneme-based index from a single recognizer, the proposed method achieved smaller index size while keeping high accuracy of the index combination method (a 16.3% and 16.0% relative error reduction for IV and OOV queries).

Keywords: spoken term detection, keyword spotting, out-of-vocabulary detection, out-of-vocabulary region estimation

1. はじめに

近年の録音装置の普及およびストレージデバイスの大容量化にともない，コールセンタでの通話録音や大学の講義の録画など音声データを含むデータが容易に大量に蓄積されるようになっている．これらの蓄積された音声データを

[1] 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University, Kyoto 606–8501, Japan
[2] 株式会社日立製作所中央研究所
Hitachi Ltd., Central Research Laboratory, Kokubunji, Tokyo 185–8601, Japan
[a)] naoyuki.kanda.kn@hitachi.com

有効に活用するためには，音声データの効率的な検索技術が欠かせない．このため，音声データを文書と見立てた音声文書検索技術が研究されている [1], [2]．米国では国際会議 TREC（Text REtrieval Conference）において音声文書検索の評価が行われ，ニュース放送を用いた評価で高い検索性能が得られることを示した [1] ほか，2006 年には音声検索語検出のワークショップが開催されている [3]．日本でも 2011 年に NTCIR（NII Testbeds and Community for Information access Research）において音声文書検索の評価タスクが設定されるなど [2] 活発に研究が進められている．

上記を背景に，本研究では音声検索語検出技術 [3] を扱う．音声検索語検出とは，音声データの中で特定の検索語が発話された時刻を検出する技術であり，音声文書検索の基盤技術として用いられている．多くの場合，検索語検出システムはインデックス作成部と検出部の 2 つに分かれている．インデックス作成部は音声データに対して 1 度だけ動作し，音声データを高速な検索語検出に適したデータ形式（音声インデックス）へと変換する．検出部はユーザが検索語を入力するたびに動作し，音声インデックスに基づいて検索語を検出する．代表的な手法の 1 つは大語彙連続音声認識を用いて音声データを単語列もしくはラティス [4], [5] へと変換し，索引処理を行うことで音声インデックスを作成するものである．検出部では，単語単位のマッチングを行うことで検索語の検出を行う．この手法は言語モデルが検索対象の音声データと適合している場合には高い検出精度[*1]を示す [3] が音声認識辞書に含まれない検索語（未知検索語）については検出することができないという問題が存在する．

そのため，サブワードに基づく検索語検出の研究が数多く行われ，未知検索語の検出において有効性が示されている [4], [5], [6], [7], [8], [9]．サブワードとは音素や音節など単語よりも細かい言語単位の総称である．この手法ではインデックス作成部は，音声データをサブワード認識器によってサブワード系列へと変換する．検出部は，検索語をサブワード系列に変換し，サブワード認識結果との類似度（編集距離など）が高い箇所を検索語の発話箇所として検出する．この手法により未知語でも検出可能となるが，似た音韻の発話箇所を誤検出することも多い．

これに対し近年になり，複数の検出結果もしくは複数のインデックスを統合することにより，既知語・未知語ともに検出精度を大幅に向上させる手法が提案されている．文献 [10] では，1/3 音素や SPS（Sub-Phonetic Segment）といった異なるサブワード単位を用いた手法からの検出結果を組み合わせることにより検出精度が向上することが示されている．文献 [11] は，単一の音声認識結果を音節や音素など異なるサブワード単位に変換した後でそれぞれのネットワークから得られた検出結果を統合することで検出精度が向上することを示した．さらに文献 [12], [13] では 10 種類の音声認識器の認識結果を Transition Network（TN）という形式に統合し，単一の音声認識器を用いた場合と比較して既知語・未知語ともに大幅な高精度化を達成した．

これらの手法はいずれも複数の音声認識結果をインデックスもしくは検出結果のレベルで統合することで検出精度の向上を達成している．一方で，統合する音声認識器の数が多いほど，インデックスサイズが大きくなるという問題がある．インデックスサイズの増大はストレージコストの増大に直結するほか，検出速度の低下にもつながるため，できる限り小さいことが望ましい．インデックスサイズを削減する方法としては，事後確率などによって求められた認識結果の信頼度 [14] を指標として，信頼度が低い区間のインデックスを削除する方法が考えられる [15], [16], [17]．しかしながら，一般に異なる認識器から得られる信頼度は異なる偏りを持ち，複数の認識結果を統合する手法において信頼度を比較することは容易ではない．加えて，未知語区間は通常は低信頼度となるため，信頼度が低いインデックスを削減することで未知検索語の検出精度が大幅に劣化する恐れがある．

本研究ではインデックス統合法において，その高い検出精度が劣化することを抑えつつインデックスサイズを削減するインデックス選択法を提案する．本研究ではインデックス統合法として文献 [12], [13] で提案された Transition Network 法を用いた．音声認識器ごと，もしくはインデックスのサブワード単位ごとに，既知語と未知語の検出特性が大幅に異なるという特性に着目し，未知語区間推定結果に基づいてネットワーク中の有効なアークの選択（アーク選択）や，インデックスに用いるサブワード単位の選択（サブワード単位選択）を行うことで，冗長なインデックスを削減する．文献 [5] では単語認識結果と音素認識結果のうち，単語信頼度が高い箇所の音素認識結果を削減する手法を提案している．また文献 [18] では，単語信頼度の代わりに未知語区間推定結果を用いる手法を提案している．これらの研究はいずれも基本的に音素認識結果は未知語の検出のために利用されており，単語信頼度が高い区間の音素認識結果を削除することは妥当である．一方インデックス統合法では，どのネットワークも既知語・未知語の検出で利用されており，適切にインデックスを削除する方法は自明ではない．提案法のうちアーク選択法は，インデックス統合法においても認識器ごとに既知語/未知語検出への寄与が大きく異なることに着目して上記の手法を拡張したものと位置づけられる．なお本研究は，文献 [19] の内容に，新たにアークの信頼度を正規化した手法との比較評価などの実験を追加したものである．

[*1] 本論文では「検出精度」という用語は検出の正確性（Precision）と網羅性（Recall）の総体を表す．いわゆる Precision に相当する用語としては「適合率」という表記を用いる．

以下ではまず 2 章で提案法のベースとなる音声検索語検出システムの構成について述べる．続いて 3 章で提案法である未知語区間推定に基づくインデックス選択法について述べる．4 章で日本語話し言葉コーパスを用いた一連の評価実験について述べる．

2. システム構成

本研究で提案する検索語検出システムの構成を**図 1** に示す．インデックス作成部ではまず 4 種類の異なる言語モデルを用いた 4 種類の音声認識器が動作する．それぞれの言語モデルの構成を以下に示す．

- **単語モデル（Word Model）**：単語を単位とした言語コーパスから学習された言語モデル．
- **音節モデル（Syllable Model）**：単語言語モデルで利用した言語コーパスを音節単位に変換した音節コーパスから学習された言語モデル．
- **単語–音節モデル（Word-Syllable Model）**：言語コーパスに含まれる単語のうち，低頻度語のみを音節単位に変換した単語–音節混合コーパスから学習された言語モデル．本研究では低頻度語の条件を出現頻度 2 回以下の単語とした．
- **フラグメントモデル（Fragment Model）**：音節フラグメントで構成されたコーパスから学習された言語モデル．音節言語モデルで利用した音節コーパスを初期コーパスとして，最も頻度が高い隣接 2 サブワードを接続して 1 つのサブワードとする操作を繰り返す[*2]．この操作により接続されたサブワードを音節フラグメントと呼ぶ．本研究では得られた音節フラグメントの平均音節数がもとの単語の平均音節数と一致するところで接続操作を停止した．

続いて，それぞれの音声認識結果から得られた認識結果を同一のサブワードを単位とするサブワード系列へと変換した後，1 つのネットワークへと統合する．本研究ではサブワード単位として**音素**または**音節**を用いた．また認識結果の統合においては Transition Network（TN）法 [13] を用いた．TN 法では単語認識から得られた 1-best サブワード系列を基準系列とする．続いて残りのサブワード系列それぞれに対し，基準系列との間の編集距離が最小になるようなアライメントを行う．最終的に得られたネットワークは Confusionn Network [5] と同様の構成となるが，本研究では上記の手法によって得られたネットワークを Transition Network（TN）と呼び区別する．音素 TN が作成される様子を**図 2** に示した．図中の ε は入力シンボルなしで遷移可能な空アーク [13] を表す．TN は発話ごとに生成される．

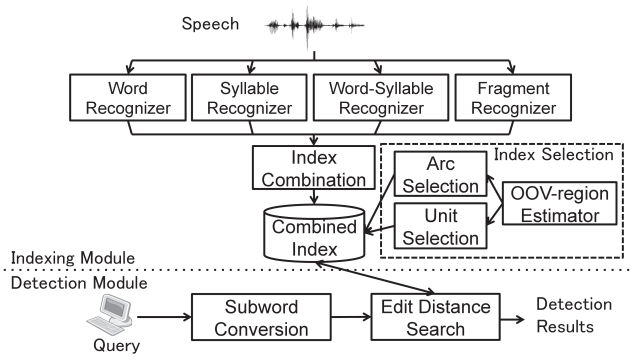


図 1 音声検索語検出システムの概要
Fig. 1 Overview of spoken term detection system.

[*2] この操作は簡易な教師なし形態素解析に相当する [20].

検出部ではまず入力された検索語（Query）をその読みに従ってサブワード系列へ変換する．続いて，検索語サブワード列が TN に含まれるようになるまでの編集回数を編集距離 E として計算し，検索語サブワード列に含まれるサブワード数 N_q で正規化したものを検出スコア S とした．

$$S = 1 - \frac{E}{N_q} \tag{1}$$

検出部は発話ごとの TN に対して検出スコアを算出し，スコアが高い順にソートして出力する．なお，編集距離 E は動的計画法を用いて効率良く算出することができる．具体的には，Transition Network の i 番目のアーク集合を $Arc(i)$（$i = \{1, \ldots, I\}$），アーク集合に含まれるアークのうち空アーク以外の要素を $a \in Arc(i)$，また検索語の j 番目のサブワードを q_j（$j = \{1, \ldots, N_q\}$）とする．このとき，Transition Network と検索語の間の編集距離 E は次の手続き 1〜3 によって求められる．

手続き 1：初期化（$i = \{0, \ldots, I\}, j = \{0, \ldots, N_q\}$）

$$D(i,j) = \begin{cases} 0 & \text{if } j = 0 \\ N_q & \text{otherwise} \end{cases} \tag{2}$$

手続き 2：反復（$i = \{1, \ldots, I\}, j = \{1, \ldots, N_q\}$）

$$D(i,j) = \min \begin{cases} \min\limits_{a \in Arc(i)} \{D(i-1, j-1) + sub(q_j, a)\} \\ D(i-1, j) + ins(Arc(i)) \\ D(i, j-1) + del \end{cases} \tag{3}$$

手続き 3：編集距離の算出

$$E = \min_{i \in \{1, \ldots, I\}} D(i, N_q) \tag{4}$$

Recognizer	Recognition Result
Word	o N g a k u o t a m e n i
Syllable	a N n a k a t a n a i n i
Word-Syllable	o N g a k u w a t a n e n i
Fragment	o N g a k u z a N n e N n i

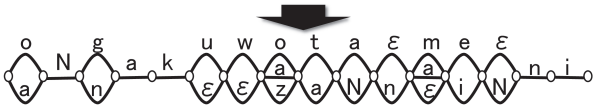


図 2 Transition Network の構築例
Fig. 2 Index combination as a transion network.

ここで $sub(q_j, a)$ はサブワード q_j とアーク a の入力シンボルの置換コストであり，本研究では簡単のため，a の入力シンボルが q_j と一致した場合に 0，それ以外の場合に 1 とした[*3]．また $ins(Arc(i))$ は，検索語サブワード列へアークの入力シンボルを挿入するコストであり，本研究ではアーク集合 $Arc(i)$ に空アーク ϵ が含まれていれば 0，それ以外の場合は 1 とした．最後に del は検索語サブワード列の削除コストであり，本研究では一律で 1 とおいた．このとき，手続き 2 の反復が全体の処理量の主となる．手続き 2 において実行される演算は，インデックスに含まれる総アーク数を M としたとき，式 (3) 右辺第 1 行と第 2 行の演算に $N_q \cdot M$ 回，第 3 行の演算に $N_q \cdot I$ 回の演算を要し，またそれらの比較に $N_q \cdot M$ 回の比較演算を必要とする．ただし実際上は第 1 行と 2 行の演算に要する処理量，および比較演算に要する処理量が律速となり処理量はおおむね $O(N_q \cdot M)$ となる．

3. 未知語区間推定に基づくインデックス選択法

3.1 最適なインデックス選択によるインデックス削減

インデックス選択の目的はインデックス統合法の高い検出精度を損なうことなくインデックスサイズを削減することである．本研究では未知語区間推定に基づくインデックス選択法を提案する．ここで未知語区間推定とは音声データ中に未知語が存在するか否かを推定する技術である [21], [22], [23]．本研究では未知語区間推定が出力する推定スコアを用いてインデックスを削減するための 2 種類の手法：アーク選択法（Arc Selection）とサブワード単位選択法（Unit Selection）を提案する．

3.1.1 アーク選択法

アーク選択法では，認識器の種類によっては，その認識器の出力を統合しても未知語の検出精度は向上しない，という仮説に基づいている．このような認識器から生成されたアークは未知語を検出する際には不要であるため，未知語が存在している可能性が高い区間（これは未知語区間推定器（3.2 節）により推定される）では，当該アークを削除しても検出精度の劣化は生じないものと期待される．上記の議論は未知語と既知語を入れ替えても同様である．この考え方に基づき，検出精度の劣化を抑えつつ不要なアークを削除する手法を本研究ではアーク選択法と呼ぶ．

本研究では開発データ（4 章）を用いた予備実験の検討に基づき，以下の条件が成立するアークを削除対象とした．

(1) 単語認識器の出力のみに基づいて生成されたアークで，未知語区間推定スコアが上位 N%に含まれる．
(2) 音節認識器もしくは単語–音節認識器の出力に基づいて生成されたアークで，未知語区間推定スコアが下位 N%に含まれる．

条件 (1) は，単語認識器に基づいて生成されたアークは既知語の検出にのみ有効であるという仮説に基づく．一方条件 (2) は音節認識器もしくは単語–音節認識器に基づいて生成されたアークは未知語の検出には有効であるが既知語の検出精度向上には寄与しないという仮説に基づく．なお本研究では，フラグメント認識器は既知語，未知語双方の検出に有効であるとして，アーク選択の対象から除外した．

3.1.2 サブワード単位選択法

サブワード単位選択法は未知語区間推定スコアに従って最適なサブワード単位を選択する手法である．サブワード単位選択法は発話単位で動作する．本手法は発話に未知語が含まれていない場合にはより粗いサブワード単位を用いても検出精度の劣化がないという仮説に基づいている．具体的には下記の条件に従ってサブワード単位を選択する．

(1) 当該発話中の未知語区間推定スコアの最大値が閾値 θ 以下であった場合には音節を単位とした TN を用いる．
(2) それ以外の場合には音素 TN を用いる．

多くの日本語音節は 2 音素で構成されていることから音節 TN のアーク数は音素 TN のアーク数よりも大幅に少ない．

サブワード単位選択法を用いた場合には音節 TN で表現された発話と音素 TN で表現された発話が混在するため，検索語の検出には，発話ごとに音節 TN か音素 TN のいずれかに基づいて検出スコアを算出し，それらを比較することとなる．ここで一般に音素認識率は音節認識率より高い[*4]のと同様に，音素 TN に基づいて計算される検出スコアは音節 TN に基づいて計算される検出スコアよりも高い傾向がある．ここでサブワード単位選択法では，未知語が発話に含まれている可能性が高い場合に音素 TN が選択される．しかし，既知の検索語を検出する際には，当該発話に未知語が含まれているかどうかはその発話に検索語が含まれているかとは本質的に無関係であり，音素 TN で表現された発話の方が高い検出スコアを出すことは検出精度の劣化につながる．このような状況を避けるためそれぞれの TN から得られる検出スコア（式 (1)）を正規化する必要がある．本研究では開発データ（4 章）に対して単語認識器で音声認識を行ったときの音節認識率（83.6%）と音素認識率（88.5%）の比 $\alpha = 0.94$ を，音素 TN から得られた検出スコアに対し乗算することによりスコアを補正した．

3.2 未知語区間推定

本研究では文献 [22] で提案された CRF（Conditional Random Field）[24] に基づく手法をベースに未知語区間推定を作成した．文献 [22] の手法では単語とサブワードが

[*3] 本研究では，文献 [13] で提案された Voting や ArcWidth は用いていない．

[*4] 2 音素で構成された音節のうち 1 音素が間違えていた場合，音素認識率は 50%と算出されるのに対し，音節認識率は 0%となるため．

混合された Confusion Network に基づいて未知語区間推定を行う．ここでは，インデキシングに利用した認識器のうちで単語とサブワードが混合された結果が出力される，単語–音節認識器から得られる Confusion Network を利用した．Confusion Network の各区間（bin と呼ぶ）を識別単位として，IV，B-OOV，I-OOV の 3 種類のラベルを出力する CRF を構成する．ここで IV は既知語であることを示す．また B-OOV と I-OOV はそれぞれ未知語の開始と，未知語の継続区間であることを示す．

識別のための特徴量としては，文献 [22] を参考に，サブワード存在確率 $P_s(t_j)$，単語エントロピー $H_w(t_j)$ を用いた．

$$P_s(t_j) = \sum_{s \in t_j} p(s|t_j) \tag{5}$$

$$H_w(t_j) = -\sum_{w \in t_j} p(w|t_j) \log p(w|t_j) \tag{6}$$

ここで t_j は Confusion Network において着目している j 番目の bin を表す．また s および w はそれぞれ当該 bin に含まれる音節と単語を表す[*5]．上記に加えて，エントロピー計算において音節も計算に加えた単語–音節混合エントロピー $H_{ws}(t_j)$ も特徴量として用いた．

$$\begin{aligned} H_{ws}(t_j) = & -\sum_{w \in t_j} p(w|t_j) \log p(w|t_j) \\ & -\sum_{s \in t_j} p(s|t_j) \log p(s|t_j) \end{aligned} \tag{7}$$

また，単語–音節認識器における最尤単語とその信頼度，単語認識器と音節認識器の最尤認識系列における言語モデルスコアの差，単語認識器と音節認識器の最尤認識系列における音響モデルスコアの差，も特徴量として用いた．

なお，CRF において連続量を扱うためには，(a) 連続量を適切に設定した小区間で分割することで離散化し，それぞれの小区間を 0 か 1 の値を持つ 1 つの素性に対応づける（ほとんどの素性が 0 となる）方法 [22] や (b) 連続量を素性の持つ重みに乗算して用いる方法 [25] などがある．未知語区間推定のベースとした文献 [22] では (a) が用いられているが，我々の予備実験の結果 (b) の性能の方が高かったため，本論文では (b) を用いた．CRF の学習と適用には (b) を実装している CRFsuite [25] を用いた．

CRF の学習においては単語–音節認識器から得られる Confusion Network の各 bin に IV，B-OOV，I-OOV のラベルを割り当てたデータを用意する必要がある．このために開発データ（4 章）に対して単語–音節認識を行って Confusion Network を生成した．正解ラベルと音声データの強制アラインメントにより求めた正解未知語区間との着目している bin の重複が 80%以上の場合に B-OOV もしくは I-OOV ラベルを付与した．

CRF を用いる際には，単語–音節認識器によって生成された Confusion Network の各 bin ごとに IV，B-OOV，I-OOV のそれぞれのラベルに対する事後確率が推定される．未知語に関連するラベルが 2 種類あるため，当該 bin が未知語である確率は B-OOV と I-OOV のそれぞれの事後確率の和で求めた[*6]．

4. 評価実験

4.1 データセット

評価を行うにあたり，日本語話し言葉コーパス [26] に含まれる学会講演・模擬講演 2,702 講演を学習データ，開発データ，評価データの 3 種類に分類した．学習データとは音響モデルや言語モデルの学習用に用いたデータである．また，開発データは未知語区間推定器の学習や各種のパラメータ調整に，評価データは提案する検索語検出システムの評価にそれぞれ用いた．評価データは日本語話し言葉コーパスに含まれる学会講演・模擬講演のうちコアセットに含まれる 177 講演，39 時間の音声データを用いた．また開発データはコアセット以外から 200 講演，46 時間の音声データを選択した．残りの 2,325 講演，522 時間の音声データは学習データとして用いた．本評価での既知語は学習データ中で 3 回以上出現した単語から評価用の未知検索語セットで未知語と定義された単語（文献 [27] に従った）を除いた 33,337 単語とした．この結果，評価データと開発データにおける単語中の未知語率はそれぞれ 2.00%と 2.04%となった．なお，言語モデル学習時には未知検索語セットで未知語と定義された単語は，当該単語を含む発話ごと学習データから除外した．

本研究での言語モデルはすべて Witten-bell スムージングに基づく 3-gram 言語モデルを用いた．また音響モデルは 2,911 状態 32 混合状態共有トライフォンモデルを用いた．特徴量は 13 次元の MFCC（Mel Frequency Cepstral Coefficients）とその差分，2 次差分を合わせた 39 次元の特徴量に対して平均分散正規化を施したものを利用した．音声認識器は Julius [28] を用い，言語モデル重みは 10，挿入ペナルティは 0 として認識を行った．**表 1** に各音声認識

表 1 音声認識器ごとの単語認識率および音素認識率
Table 1 Word and phoneme-based accuracy rates of four speech recognizers.

Recognizer	Word Acc. (%)	Phoneme Acc. (%)
Word	74.5	88.6
Syllable	-	84.9
Word-Syllable	70.8	88.7
Fragment	-	87.9

[*5] 単語–音節認識器を用いた場合，Confusion Network の bin は単語と音節を同時に含んだものとなる．また，$\sum_{s \in t_j} p(s|t_j) + \sum_{w \in t_j} p(w|t_j) = 1$ である．

[*6] これは “1-(IV ラベルの事後確率)” と等価である．

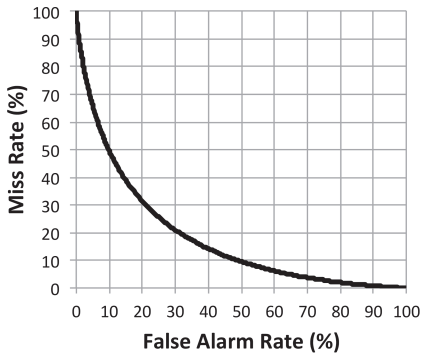


図 3 未知語区間推定結果
Fig. 3 Evaluation of OOV-region estimator.

器から得られた単語列もしくは音素列の認識率を示す．表から音素認識率において音節認識器は他の認識器より若干悪いことが分かる．その他のサブワード認識器は音素認識率の観点ではほぼ同等であった．

4.2 未知語区間推定の評価結果

図 3 に未知語区間推定の評価結果を示す．図では横軸が誤検出率（False Alarm Rate），縦軸が Miss Rate を表す．それぞれの定義を下記に示す．

$$Miss\ rate = 1 - \frac{\#\ of correctly\ detected\ OOV}{\#\ of\ actual\ OOV} \quad (8)$$

$$False\ alarm\ rate = \frac{\#\ of\ IV\ detected\ as\ OOV}{\#\ of\ actual\ IV} \quad (9)$$

図よりたとえば未知語区間の 70%を検出したとき（Miss Rate = 30%），誤検出率は約 20%であったことが分かる．

4.3 単一の音声認識器を用いた場合の検出精度

まず 4 種類の音声認識器それぞれの検出精度の評価を行った．検索語セットは文献 [27] で定義された既知語のみで構成された 50 個の検索語（既知検索語セット）と未知語を含む 50 個の検索語（未知検索語セット）を用いた．未知検索語セットに含まれる複合語の中には既知語を含むものもあるが，必ず 1 つ以上の未知語を含んでいる．評価データにおける平均出現回数はそれぞれ 14.5 回と 4.7 回である．検索語検出の検出精度の評価指標として F 値（F-measure）を用いた．F 値は適合率（Precision）と再現率（Recall）の調和平均であり，適合率と再現率は同一閾値での検索語平均によって求めた．ここでは閾値を変化させたときの F 値が最大となる点での F 値を評価尺度とした[*7]．

表 2 にそれぞれの音声認識器を単独で用いた場合の既知語セット（IV）と未知語セット（OOV）ごとの検出精度，およびネットワークのインデックスサイズ（Index Size）を示す．インデックスサイズは，正解データにおける 1 単語あたりのネットワークのアーク数として示している．表中の “Recognizer” は利用した音声認識器の種類を表す．“Index Unit” はインデックスを構成する際のサブワード単位を表す．1-best は最尤認識結果のみを用いた場合，CN は文献 [5] と同様の方法によって Confusion Network を構築した場合，TN は 2 章と同様の方法によって 5-best 認識結果を Transition Network へ変換した場合を表している．CN と TN は代表的な条件でのみ結果を示す．なお，表中で CN や TN の記載のないものはすべて 1-best の結果を表す．

表 2 単一の音声認識を用いた場合の F 値とインデックスサイズ
Table 2 F-measure and index size of single system.

Recognizer	Index Unit	IV(%)	OOV(%)	Index Size
Word	Word	74.9/	19.7/	1.03/
	(1-best/CN)	**78.2**	35.4	5.90
	Syllable	76.4	46.8	1.81
	Phoneme	75.9/	53.9/	3.21/
	(1-best/TN)	77.7	56.7	4.78
Syllable	Syllable	58.5	55.7	1.78
	Phoneme	61.5	62.8	3.16
Word-Syllable	Word	72.5	19.8	1.06
	Syllable	74.7	57.8	1.80
	Phoneme	74.9/	**63.3**/	3.20/
	(1-best/TN)	76.9	62.0	4.68
Fragment	Syllable	68.0	55.3	1.80
	Phoneme	71.2/	**63.3**/	3.19/
	(1-best/TN)	71.6	62.3	4.92

単語認識器（Word）を用いた場合，既知語の検出精度はどのインデックス単位でもおおむね高く，特に単語単位の Confusion Network を用いたときに，単体の音声認識器では最高の 78.2%の F 値が得られた．ただし，このときのインデックスサイズは 5.90 と大きかった．また，TN も 77.7%と高い F 値が得られたが，やはりインデックスサイズは 4.78 と大きかった．一方で，未知語の検出精度については最大でも 56.7%とその他の認識器と比較して 6 ポイント程度低かった．なお，単語単位のインデックスにおいても未知語検出の F 値が 0%でないのは，今回は単語単位の検索でも式 (1) の検出スコアを用いており，その結果，既知語を一部含む複合語が検出されためである．

音節認識器（Syllable）を用いた場合既知語の検出精度が低く，最大でも 61.5%と単語認識器の場合と比較して 15 ポイント前後も低かった．一方で，未知語の検出精度は高く，単語認識器より約 6 ポイント高い 62.8%という精度が得られた．音節認識器の場合，未知語と既知語の F 値がほぼ同等であった．

単語–音節認識器（Word-Syllable）でインデックスを構築した場合，既知語，未知語ともバランス良く高い検出精度が得られた．特に未知語の F 値 63.3%は，4 種類の音声認識器の中で最大であった．既知語の検出精度に関しては

[*7] なお文献 [19] では 1 つも検出結果が得られなかった場合その検索語の適合率を 0 として評価していたが，本論文では NTCIR-9 [2] での評価に合わせて，この場合の適合率を 1（再現率は 0）として評価した．このため文献 [19] とは結果が一部異なる．

表 3 インデックス統合法における F 値とインデックスサイズ
Table 3 F-measure and index size of combined system.

Recognizer	Index Unit	IV(%)	OOV(%)	Index Size
All	Syllable	**81.6**	65.0	2.35
	Phoneme	80.6	**70.6**	3.87

単語認識器と比較して 1 から 2 ポイント前後の劣化がみられた．

フラグメント音声認識器（Fragment）も既知語，未知語ともバランス良く高い検出精度が得られた．未知語検出の F 値 63.3%であり，単語–音節認識器と同様に最高値であった．一方既知語の検出に関しては単語認識器から 6 から 8 ポイント程度低い結果となった．

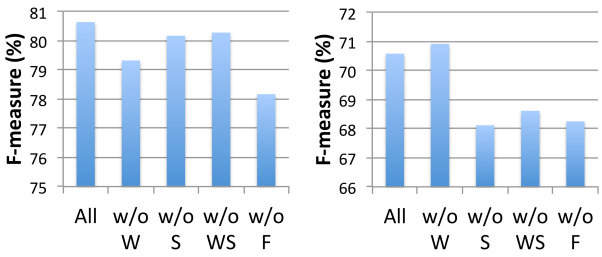
上記の結果から，表 1 のように音素認識率が同等であっても，言語モデルが異なるだけで既知語と未知語の検出特性が大幅に異なることが分かる．また，音節単位のインデックスと音素単位のインデックスを比較するとほとんどの場合[*8]に音素単位のインデックスの検出精度が良い反面，インデックスサイズは音節単位の方が小さくてすむことが分かる．

4.4 インデックスを統合した場合の検出精度

表 3 に 4 種類の音声認識結果を TN として統合した場合の検出精度とインデックスサイズを示す．音節 TN として認識結果を統合することで既知検索語に対し 81.6%と高い検出精度が得られたが，未知検索語に対しては 65.0%と十分高い検出精度は得られなかった．これに対し，音素 TN ではインデックスサイズは 3.87 と大きいものの，未知語の検出精度が向上し 70.6%の F 値が得られた．また既知語に対しても 80.6%と高い検出精度が得られた．この結果から，検出精度の観点からは音素 TN を作成することが望ましいがインデックスサイズは大きくなることが分かる．

上記をふまえ，以降は音素 TN をベースラインとし，このインデックスサイズをできる限り小さくすることを目指す．続いて，検出精度の向上にどの音声認識器の結果が寄与しているかを調べるために，音素 TN から，各音声認識の結果を 1 つだけ取り除いた結果を**図 4** に示す．図から，たとえば，未知語の検出においては単語認識器の結果を取り除くことでむしろ F 値が 70.9%に向上している．このことは，単語認識器に基づいて生成されたアークは未知語の検出においてなんら寄与していない（むしろ誤検出を増加させている）ことが分かる．同様の観察から既知語の検出においては音節認識器や単語–音節認識器に基づいて生成されたアークは既知語の検出にほとんど寄与していないことが分かる．なお既知語の検出において，単語–音節認識器を単体で用いた場合には検出精度が高いのに，インデックスを統合したときには単語–音節認識器の寄与が少ない

*8 単語認識器で既知語を検出する場合が唯一の例外であった．



(a) 既知検索語 (b) 未知検索語

W: Word, S: Syllable, WS: Word-Syllable, F: Fragment

図 4 特定の音声認識結果を除外した場合の F 値
Fig. 4 F-measure without a particular recognizer.

要因は，既知語に関しこの認識器が単語認識器と同じような認識結果を出力しているためと考えられる．なお，音節 TN に関しても上記と同様の結果が観察された．

4.5 インデックス選択法の評価

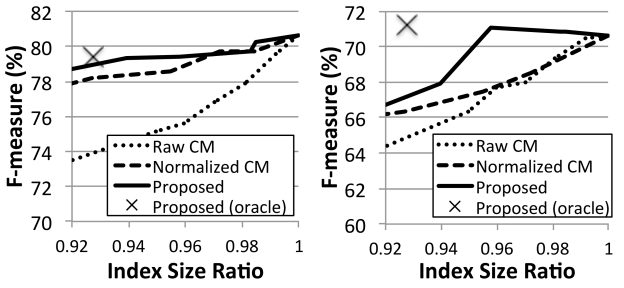
4.5.1 アーク選択法の評価

ここではまずアーク選択法の評価を行う．下記の 4 種類の手法を比較した．

- **Raw CM**：当該アークの元となる認識器が出力する認識信頼度[*9]が閾値を下回った場合に当該アークを削除する．あるアークが複数の音声認識器の出力に基づいて生成されている場合には，そのうちの最大値を当該アークの信頼度とした．
- **Normalized CM**：アークの信頼度をロジスティック回帰モデル [24] によって算出しなおし，そこで得られた信頼度（Normalized CM）が閾値を下回った場合に当該アークを削除する．ロジスティック回帰の特徴量として，当該アークに対して各音声認識器から得られる信頼度[*10]，そのアークの元となる音声認識の種類数の逆数，および当該アークと同一区間にあるアークの数とした．またロジスティック回帰モデルは開発データを用いて Transition Network を構成した後，アークに正解/不正解ラベルを付与することで作成したデータに基づき学習した．
- **Proposed**：未知語区間推定に基づくアーク選択法．
- **Proposed（Oracle）**：アーク選択法において未知語区間推定の正解ラベルを与えた場合の性能．

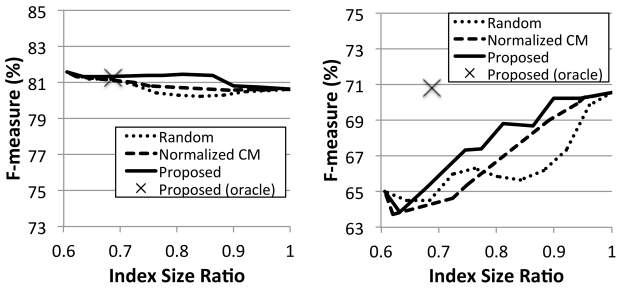
いずれの手法も音素 TN をベースとして，閾値を変えて少しずつアークを削除していった．結果を**図 5** に示す．図において縦軸は F 値を表す．また横軸は音素 Transition Network のアーク数を 1 としたときのアーク数を表しており，図の右端での F 値が音素 TN の検出精度に相当する．図よりまず音声認識器から出力される信頼度に基づきアークを削除する手法（Raw CM）では既知検索語セット，未

*9 本研究では Julius [28] が出力する信頼度 [14] を用いた．
*10 その音声認識器から生成されたアークではない場合は 0 とする．



(a) 既知検索語　(b) 未知検索語

図 5 アーク単位選択法の評価結果
Fig. 5 Evaluation of arc selection method.



(a) 既知検索語　(b) 未知検索語

図 6 サブワード単位選択法の評価結果
Fig. 6 Evaluation of unit selection method.

知検索語セットのいずれにおいても，インデックスサイズを小さくする（アークを削除する）ことにより急激に検出精度が劣化していることが分かる．既知語においても検出精度が劣化しており，異なる音声認識器から出力される信頼度を比較することが難しいことが示唆されている．ロジスティック回帰により求めた信頼度に基づいてアークを削除する手法（Normalized CM）では Raw CM と比較して，インデックスを小さくした際の検出精度の劣化が緩やかになっている．しかしながら，未知語検索セットではやはり急激な検出精度の劣化がみられた．これは信頼度の低いアークから削除することにより，未知語を含む区間のアークが多く削除されたためと思われる．未知語区間推定に基づく提案法（Proposed）はいずれの場合も最も良いインデックス削減性能を示した．ただし，未知語の検出においては 4.2%以上のアークを削った段階で検出精度の劣化が生じており，本手法によるインデックス削減効果は限定的であることも分かった．正しい未知語区間推定結果を与えた場合には既知語で 79.4%，未知語で 71.2%とほぼ検出精度を保ったまま 7.2%のアークが削除された．これらの結果から未知語区間推定に基づくアーク選択は，インデックス削減効果は小さいものの，インデックス削減に有効であることが確認された．信頼度を用いたアーク削減法では特に未知語の検出精度劣化が大きいことも確認された．

4.5.2 サブワード単位選択法の評価

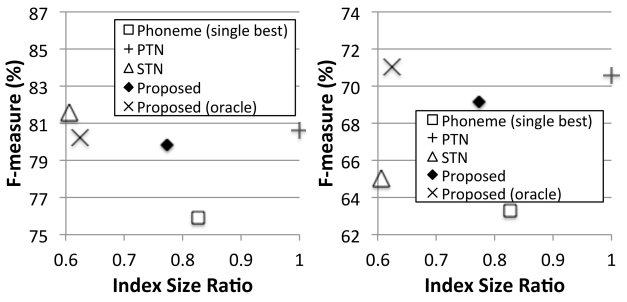
続いてサブワード単位選択法の評価を行った．ここでは下記の 4 種類の手法を比較評価した．

- **Random**：音節 TN を用いるか音素 TN を用いるかを発話ごとにランダムに決定する．
- **Normalized CM**：前節で用いた Normalized CM を用い，当該発話に含まれる正規化信頼度の最小値が閾値を超えていれば音節 TN を用いる．それ以外の場合には音素 TN を用いる．この手法は未知語区間推定結果ではなく，認識信頼度が高い場合にサブワード単位として音節を使っても検出精度への影響がないという仮説を検証するための手法である．
- **Proposed**：未知語区間推定に基づくサブワード単位選択法．
- **Proposed（Oracle）**：サブワード単位選択法において未知語区間推定の正解ラベルを与えた場合の性能．

前節と同様に音素 TN をベースとして，閾値を変えて少しずつインデックスサイズを削減していった．結果を**図 6** に示す．横軸と縦軸は図 5 と同様である．ランダムにサブワード単位を選択する手法（Random）は予想どおり性能が低く，特に未知語の検出において，インデックスサイズを小さくした際の検出精度劣化が顕著である．Normalized CM に基づくサブワード単位選択法は Random と比較するとインデックスを小さくした際の検出精度が高いがやはり未知語の検出において検出精度の劣化が確認された．特にインデックスの 20%以上を削減した段階（Index Size Ratio が 0.8 以下）では Random よりも悪い性能がみられており，本手法がインデックス削減に有効ではないことが確認された．一方，未知語区間推定に基づいてサブワード単位を選択する手法はインデックスサイズ比が 0.6 付近を除き，一貫してその他の手法よりも高い性能を示した．特に未知語区間推定が正しい結果を出力した場合には既知語 81.3%，未知語 70.8%と音素 TN の高い検出精度を維持したまま，31.2%のインデックスが削減された．これらの結果から未知語区間推定に基づいてインデックス単位を選択する提案法が有効であり，インデックス削減効果も大きいことが確認された．

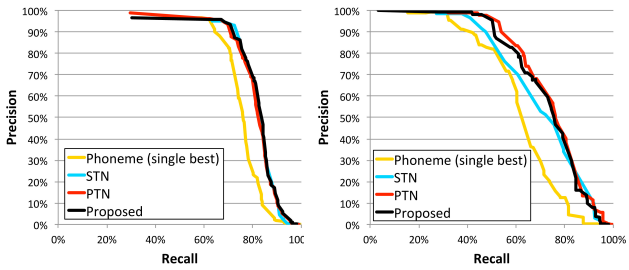
4.5.3 提案法を組み合わせた場合の評価

最後に，アーク選択法とサブワード単位選択法を組み合わせた場合の検出精度を**図 7** に示した．図のうちで “Phoneme（single best）” は単一の音声認識器を用いて音素単位のネットワークを作成したときに最も良いものを指す．具体的には，既知語では単語音声認識を用いた場合，未知語では単語–音節音声認識を用いた場合の性能を示している．また “PTN” および “STN” はそれぞれ 4 種類の音声認識器による認識結果を統合した音素 Transition Network と音節 Transition Network を示す．“Proposed” はアーク選択法（閾値 $N = 60\%$；図 5 においてアーク削減率 4.2%の点）とサブワード単位選択法（閾値 $\theta = 0.06$；図 6 にお



(a) 既知検索語 (b) 未知検索語

図 7 インデックス選択法を組み合わせた場合の評価
Fig. 7 Evaluation of mixed method.



(a) 既知検索語 (b) 未知検索語

図 8 各種法の再現率–適合率曲線
Fig. 8 Recall-Precision curve.

いてアーク削減率 18.9%の点）を組み合わせた場合の性能を示す．ここではまずサブワード単位選択法で音素 TN か音節 TN かを選択した後，それぞれの TN に上記の条件でアーク選択法を適用した．“Proposed（oracle）” は正しい未知語区間を与えた場合の提案法の性能を示す．図より，PTN は検出精度は未知語，既知語ともに高いがインデックスサイズは最も大きい．STN はインデックスサイズは小さいが，未知語の検出精度が低い．提案法（Proposed）の検出精度は既知語 79.8%，未知語 69.2%と，ともに PTN とほぼ同等の高い検出精度を保ったまま，インデックスサイズを 22.7%削減できた．本節では，まずサブワード単位選択法を適用した後にアーク選択法を適用しており，18.9%がサブワード単位選択法による削減効果，残りの 3.8% [*11] がアーク選択法による削減効果となる．このインデックスサイズは単一の音声認識器を用いた場合（Phoneme（single best））よりも小さく，かつ検出精度としてはそれぞれ 3.9 ポイントと 5.9 ポイントずつ高い．これはそれぞれ 16.3%と 16.0%の検出エラー削減に相当する．最後に，正しい未知語区間が与えられた場合（Proposed（oracle））には，PTN から検出精度劣化なく 37.6%のインデックスが削減され，提案法がインデックス削減手法として有望であることが確認された．

図 8 に再現率–適合率曲線を示す．横軸が再現率，縦軸が適合率である．図より，PTN と提案法（Proposed）が既知語，未知語ともにほぼ重なっており，提案法が PTN 法の検出精度の高さを再現率–適合率曲線の多くの点で保っていることが分かる．以上から，提案法によって PTN の検出精度の高さを維持したまま，インデックスサイズを削減できることが確認された．

4.6 インデックス削減効果に関する考察

本研究ではインデックス統合法によって得られた高い検出精度が劣化するのを 1.4 ポイントに抑えつつ，インデックスサイズを 22.7%削減した．また正しい未知語区間が与えられた場合の上限として 37.6%のインデックスが削減できることを確認した．本論文ではここまでインデックスサイズをアーク数で評価してきたが，本節では，実際のインデックスサイズを交えてインデックス削減効果について説明する．

インデックスサイズの削減は (1) インデックスのストレージに関わるコストの削減と，(2) 検出速度の高速化の 2 点の効果がある．このうち (1) についてはさらに，(a) インデックスを外部ストレージに保管することに要するコストと，(b) インデックスをサーバ内のメモリに展開する際に考慮すべきコストに分かれる．

まず (1)-(a) のインデックスを外部ストレージに保管するコストについて議論する．本研究の実装において，4 種類の音声認識器の出力をすべて統合したとき（従来法）のインデックスサイズは 1 時間あたり約 0.84 MB であった．インデックスサイズはアーク数に比例しており，提案法によってインデックスサイズが 1 時間あたり 0.65 MB に抑えられる．この場合従来法では，たとえば音声データ量が 120 万時間あったときにインデックスサイズが 1 TB となり，インデックスサイズが削減されると，このストレージコストが削減できることとなる．120 万時間の音声データ量というのは，大規模なコールセンタなど多数の音声チャンネルが同時に録音されるような環境において十分にありうるデータ量である．ただし近年は大容量ストレージの価格が安価になっていることを考えると，インデックスを外部ストレージに保管するコストが有用性を持つのは，120 万時間よりもさらに大規模な音声データが存在する場合に限るといえる．

一方で (1)-(b) の観点で，インデックスをサーバ内のメモリに展開する際に考慮すべきコストについて議論する．検出速度を考えた場合，インデックスをメモリに展開しておくことは重要である．本論文で用いた実装では，4 種類の音声認識器の出力をすべて統合したときに必要なメモリ量は 1 時間あたり 7.9 MB であった．これは，たとえば 64 GB のメモリを持ったサーバでは，1 台あたり 8,000 時間程度の音声に相当するインデックスをメモリに展開できる計算となる．これに対し，アーク選択法とサブワード単位選択

*11 サブワード単位選択法と組み合わせた場合，一部は音節 TN へのアーク選択法の適用となるため，音素 TN での削減率 4.2%とは若干異なる．

法の組合せ（4.5.3 項）において必要なメモリ量は 1 時間あたり 6.3 MB であり，おおむねアーク数に比例したメモリ削減効果（約 20%）が得られた[*12]，64 GB のメモリを持ったサーバを想定すると，たとえば検索対象の音声が 4 万時間ある場合に従来法では 5 台のサーバが必要となるが，メモリ使用量を 20%削減できた場合には 4 台のサーバで十分となり，サーバの台数を減らすことが可能となる．このように，サーバの台数が削減できるという点でインデックスサイズの削減効果は 20%であっても大きいといえる．

最後に (2) の観点で，検出速度の高速化について述べる．2 章で述べたように本研究で用いた編集距離の計算において，動的計画法の反復手続き中に実行される演算量はおおむねアーク数に比例している．したがって，アーク数を 20%削減できた場合，これは検出速度をおおむね 20%高速化できたことに相当する．なお検出速度の高速化に関しては，これまでも多くの研究（文献 [18], [29], [30] など）が行われている．実際の応用においてはこれらの高速化手法との組合せも重要となるが，この議論は本論文の範囲を超えるため上記の記述にとどめる．

5. まとめ

本研究では音声検索語検出において，未知語区間推定結果に基づいて複数の音声認識器から出力された認識結果を選択的に統合することで，検出精度の劣化を抑えつつインデックスサイズを削減する手法について提案した．評価の結果，提案法により，検出精度の劣化を 1.4 ポイントに抑えつつ音素 Transition Network から 22.7%のインデックスを削減できた．単一の音声認識結果から作成した音素単位のネットワークと比較した場合，提案法では，インデックスの統合による検出精度向上の効果（既知語で 16.3%，未知語で 16.0%の検出エラー削減）を保ちながら，単一の音声認識結果に基づくインデックスと同等以下の大きさまでインデックスサイズを抑えることができた．

参考文献

[1] Garofolo, J.S., Auzanne, C.G. and Voorhees, E.M.: The TREC spoken document retrieval track: A success story, *NIST SPECIAL PUBLICATION SP*, No.246, pp.107–130 (2000).

[2] Akiba, T., Nishizaki, H., Aikawa, K., et al.: Overview of the IR for spoken documents task in NTCIR-9 workshop, *Proc. NTCIR-9* (2011).

[3] Fiscus, J.G., Ajot, J., Garofolo, J.S., et al.: Results of the 2006 spoken term detection evaluation, *Proc. ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, pp.51–55 (2007).

[4] Yu, P. and Seide, F.: A hybrid word/phoneme-based approach for improved vocabulary-independent search in spontaneous speech, *Proc. ICLSP'04* (2004).

[5] Hori, T., Hetherington, I.L., Hazen, T.J., et al.: Open-vocabulary spoken utterance retrieval using confusion networks, *Proc. ICASSP*, Vol.4, pp.IV–73, IEEE (2007).

[6] Saraclar, M. and Sproat, R.: Lattice-based search for spoken utterance retrieval, *Proc. HLT-NAACL*, pp.129–136 (2004).

[7] Parada, C., Sethy, A. and Ramabhadran, B.: Balancing false alarms and hits in spoken term detection, *Proc. ICASSP*, pp.5286–5289, IEEE (2010).

[8] 神田直之，住吉貴志，小窪浩明ほか：多段リスコアリングに基づく大規模音声中の任意検索語検出（音声，聴覚），電子情報通信学会論文誌 D，情報・システム，Vol.95, No.4, pp.969–981 (2012).

[9] Kanda, N., Takeda, R. and Obuchi, Y.: Using rhythmic features for Japanese spoken term detection, *Proc. SLT*, pp.170–175, IEEE (2012).

[10] 伊藤慶明，岩田耕平，石亀昌明ほか：語彙制限のない音声文書検索における複数サブワードの統合—検索語彙に依存した検索性能推定指標の導入，情報処理学会論文誌，Vol.50, No.2, pp.524–533 (2009).

[11] Bufyko, I., Kimball, O., Siu, M., et al.: Detection of unseen words in conversational Mandarin, *Proc. ICASSP*, pp.5181–5184, IEEE (2012).

[12] Natori, S., Nishizaki, H. and Sekiguchi, Y.: Japanese spoken term detection using syllable transition network derived from multiple speech recognizers' outputs, *Proc. INTERSPEECH*, pp.618–684 (2010).

[13] Nishizaki, H., Furuya, H., Natori, S., et al.: Spoken Term Detection Using Multiple Speech Recognizers' Outputs at NTCIR-9 SpokenDoc STD subtask, *Proc. NTCIR-9* (2011).

[14] Lee, A., Shikano, K. and Kawahara, T.: Real-time word confidence scoring using local posterior probabilities on tree trellis search, *Proc. ICASSP*, Vol.1, pp.I–793, IEEE (2004).

[15] Yu, P., Shi, Y. and Seide, F.: Approximateword-lattice indexing with text indexers: Time-Anchored Lattice Expansion, *Proc. ICASSP*, pp.5248–5251, IEEE (2008).

[16] Gao, J., Zhao, Q., Yan, Y. and Shao, J.: Efficient system combination for syllable-confusion-network-based Chinese spoken term detection, *Proc. ISCSLP*, pp.1–4, IEEE (2008).

[17] Pinto, J., Szoke, I., Prasanna, S. and Hermansky, H.: Fast approximate spoken term detection from sequence of phonemes, *Proc. ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, pp.8–45 (2008).

[18] Kanda, N., Sagawa, H., Sumiyoshi, T., et al.: Open-vocabulary keyword detection from super-large scale speech database, *Proc. MMSP*, pp.939–944, IEEE (2008).

[19] Kanda, N., Itoyama, K. and Okuno, H.G.: Multiple index combination for Japanese spoken term detection with optimum index selection based on OOV-region classifier, *Proc. ICASSP*, pp.8540–8544, IEEE (2013).

[20] Ablimit, M., Kawahara, T. and Hamdulla, A.: Discriminative approach to lexical entry selection for automatic speech recognition of agglutinative language, *Proc. ICASSP*, pp.5009–5012, IEEE (2012).

[21] Rastrow, A., Sethy, A. and Ramabhadran, B.: A new method for OOV detection using hybrid word/fragment system, *Proc. ICASSP*, pp.3953–3956, IEEE (2009).

[22] Parada, C., Dredze, M., Filimonov, D., et al.: Contextual information improves OOV detection in speech, *Proc. NAACL-HLT*, Association for Computational Linguistics, pp.216–224 (2010).

[*12] メモリに展開する場合のメモリ量は，実際にはノード数の影響も受けるがアークを保管するメモリ量が主となった．

[23] Qin, L., Sun, M. and Rudnicky, A.: System combination for out-of-vocabulary word detection, *Proc. ICASSP*, pp.4817–4820, IEEE (2012).
[24] Murphy, K.P.: *Machine learning: A probabilistic perspective*, The MIT Press (2012).
[25] Okazaki, N.: CRFsuite: A fast implementation of Conditional Random Fields (CRFs) (2007).
[26] Maekawa, K.: Corpus of Spontaneous Japanese: Its design and evaluation, *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition* (2003).
[27] 西崎博光，胡　新輝，南條浩輝ほか：Spoken term detection のためのテストコレクション構築とベースライン評価，情報処理学会研究報告 SLP，音声言語情報処理，Vol.2010, No.13, pp.1–8 (2010).
[28] Lee, A., Kawahara, T. and Shikano, K.: Julius – An open source real-time large vocabulary recognition engine, *Proc. EUROSPEECH*, pp.1691–1694 (2001).
[29] Katsurada, K., Teshima, S. and Nitta, T.: Fast keyword detection using suffix array, *INTERSPEECH*, pp.2147–2150 (2009).
[30] Nakagawa, S., Iwami, K., Fujii, Y. and Yamamoto, K.: A robust/fast spoken term detection method based on a syllable n-gram index with a distance metric, *Speech Communication* (2012).

神田 直之　（正会員）

2004 年京都大学工学部情報学科卒業．2006 年同大学院情報学研究科修士課程修了．同年株式会社日立製作所入社．同社中央研究所に配属．2012 年京都大学情報学研究科博士課程入学．音声認識，音声文書検索，音声対話等の音声言語処理の研究に従事．日本音響学会会員．

糸山 克寿　（正会員）

2006 年京都大学工学部情報学科卒業．2008 年同大学院情報学研究科知能情報学専攻修了．2011 年同大学院博士課程修了．博士（情報学）．同年京都大学大学院情報学研究科助教．音源分離や音楽鑑賞インタフェース等の音楽情報処理の研究に従事．日本音響学会，IEEE 各会員．

奥乃 博　（正会員）

1972 年東京大学教養学部基礎科学科卒業．1996 年東京大学博士（工学）．1972 年日本電信電話公社入社．NTT 基礎研究所を 1998 年退職．科学技術振興事業団 ERATO，東京理科大学理工学部情報科学科を経て，2001 年より，京都大学大学院情報学研究科知能情報学専攻教授．プログラミング環境，人工知能研究を経て，音環境理解，音楽情報処理，ロボット聴覚の研究に従事．日本ソフトウエア科学会，人工知能学会，本学会各元理事，IEEE Fellow，人工知能学会フェロー，1990 年度人工知能学会論文賞，IROS-2010 NTF Award for Entertainment Robots and Systems，平成 25 年度科学技術分野の文部科学大臣表彰科学技術賞（研究部門）等受賞．