Vol.2011-SLP-87 No.6
2011/7/21

音声認識・検索のための未知語の扱い

中川聖一[†]　Welly NAPTALI[†]　岩見圭祐[†]

本稿では，大語彙連続音声認識システムにおける未知語認識および音声ドキュメント検索における未知語検索の問題について述べる．前者に対しては，認識辞書の語彙を増加させることが有効であること，および，いくら増加させても未知語の出現は避けられないので，未知語の認識辞書への登録とその言語モデルの構築法について述べる．後者に対しては，大語彙連続音声認識とサブワード単位系列の併用が有効であること，およびその併用法について述べる．

Unknown word Processing Method for Speech Recognition and Retrieval

Seiichi NAKAGAWA[†]　Welly NAPTLI[†]
Keisuke IWAMI[†]

In this paper, we point out some problems for out-of-vocabulary words on large vocabulary continuous speech recognition (LVCSR) and spoken term detection. For the former problem, we had better increase the vocabulary size as many as possible but we can not register all unknown words, therefore we describe a registration method and language modeling of unknown words. For the latter problem, we point out the effectiveness of the combination method of LVCSR and sub-word recognition, and describe an effective combination method.

1. はじめに

音声認識技術の進歩に伴い，認識対象の語彙は 10 万語～100 万語へと大語彙化している．これによって，ほとんどの語彙をカバーできるようになったが，依然として姓名，組織名などの固有名や新しい造語による未知語の問題が存在する．テキスト入力の場合の未知語は文字列としては正しく入力されているので，未知語部分の同定は比較的容易で，その未知語の品詞や意味の同定が問題となるが，検索には問題はない．一方，音声入力の場合の未知語は，そもそも発話音声中に未知語が存在するかどうかさえ不明である．

音声検索の応用を考えると，固有名詞などが検索語となることが多く，未知語対策が重要となる．検索語の入力がテキスト入力の場合，入力が未知語であることはわかるが，検索対象の音声ドキュメントには，対応する未知語が文字列として正しく認識されていない場合が多い．

本稿では，音声認識時の未知語対策と音声検索時の未知語対策について述べる．

2. パープレキシティと補正パープレキシティ

音声認識の困難さを表わす尺度としてパープレキシティがよく用いられる．N-gram 言語モデルによる単語列 $w = w_1, w_2, \cdots w_N$ のパープレキシティは次式で定義される．

$$PP = \left[\prod_{i=1}^{N} p(w_i \mid w_{i-n+1}, w_{i-n+2}, \cdots, w_{i-1} \right]^{-\frac{1}{N}}$$

これは単語列が情報源からの典型的な観測系列とすれば，この情報源のエントロピーを H とすると，n → ∞ のとき $pp = 2^H$ となる．

パープレキシティは情報理論的な意味での分岐数・出現候補数に対応する．ここで，パープレキシティを計算する時に問題となるのは，単語列中に未知語が存在する場合である．この場合，便宜上，次の二つの方法で計算する．一つは未知語をスキップして求める方法である．もう一つは，未知語の集合を一つの単語（UNK）とみなす方法である．最近は後者の方法が使用されている場合が多い．この場合，問題となるのは

* † 豊橋技術科学大学　情報・知能工学系
Toyohashi　University of Technology

認識語彙サイズが少ないと未知語率が大きくなり，結果的にパープレキシティが小さくなることである．そのために，次式で定義される補正パープレキシティが提案された〔1〕．

$$APP = \prod_{i=1}^{N} p(w_i \mid w_{i-n+1}, w_{i-n+2}, \cdots w_{i-1}) \cdot m^{-o})^{-\frac{1}{N}}$$

ここで，mはトレーニングデータ中の未知語の種類数で，oは単語列中の未知語の出現数である．すなわち，未知語クラスから各未知語 unk_k の出現確率を $p(unk_k \mid UNK) = \frac{1}{m}$ の一様分布とみなすことに等しい．実際には，これにも問題がある．何故なら，トレーニングデータ量が多くなればなるほど，未知語率はあまり変わらなくても未知語の種類数が多くなり，APPの値が大きくなってくるからである．この問題点を防ぐために，筆者らは，未知語に対してはZipfの法則に従って出現すると仮定し，トレーニングデータ中の未知語数に依存しない方法を提案した（新補正パープレキシティ）〔2〕．しかし，一般にはある程度大規模なトレーニングデータ量であれば，言語モデルの比較（認識語彙サイズの変更も含めて）には，APPで十分役立つ．勿論，未知語部分を文字列の生成モデルで表現し，パープレキシティ（エントロピー）を求める方法もある〔3〕〔4〕．

以上の関係を図で示したのが図 1 である．この図からもわかるように，音声認識は，認識用単語辞書の語彙サイズを大きくすればするほど，言い換えれば未知語率を小さくすればするほど，認識性能は良くなる（基本的仮定：トレーニングデータの単語の出現傾向とテストデータの単語の出現傾向は同じ場合）．

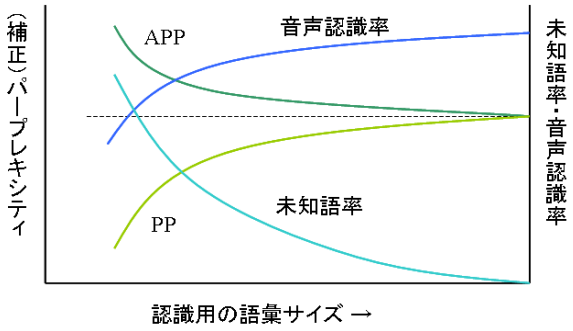


図 1　パープレキシティと補正パープレキシティの関係
Figure 1　Relationship between perplexity and adjusted perplexity

3.　未知語の登録と n-gram 確率の推定

前節の考察により，認識システムの計算量，メモリ量の許容範囲内で，トレーニングデータに依存する未知語（図 2 のOOV1）の未知語率を減らすために，認識用語彙サイズを大きくすれば良いことがわかったが，トレーニングデータに出現しないで，テストデータに出現する未知語（図 2 のOOV2）も存在する．この場合は，テストデータの認識に先立って，出現が予想される未知語を推定し，認識辞書に登録する必要がある．最近の一般的な手法は，直前の認識結果から，もしくは一旦認識対象文の音声認識を行ない，その結果から，話題やトピックを同定し（検索語を抽出し），Webから類似するコーパスを収集して，その中に出てくる未知語を登録することが試みられている〔5〕．この手法は，類似コーパスによる言語モデルの適応とみなすことができる〔6〕．

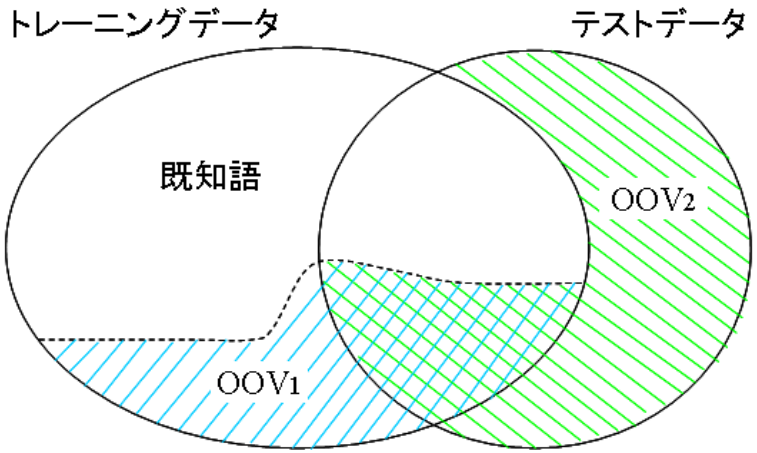


図 2　二つのタイプの未知語
Figure 2　Tow types for unknown words

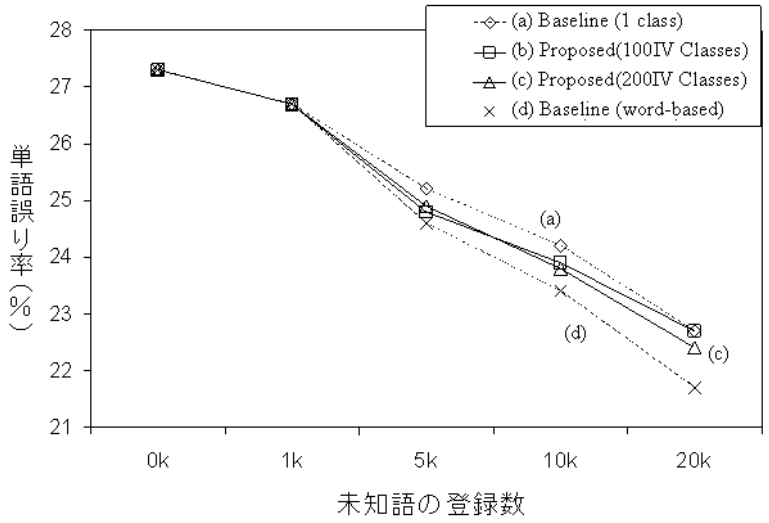


図 3　未知語の登録と言語モデルの相違による音声認識率
（ベースラインの語彙サイズは 20k）

Figure 3　Word accuracy based on various methods of unknown word registration and language model

何らかの方法によって未知語が認識辞書に登録されたとして，次の問題はn-gram確率をどのように推定するかである．未知語を一つのクラスとすれば，クラスから各未知語への出現確率の推定問題となるが，あまり頻出しない未知語の出現確率の推定は難しい. 文字列生成モデルから未知語の出現確率を推定することも可能であるが〔3〕，我々は観点を変え，未知語を意味的・構文的に多クラスに分類する手法を提案している〔7〕．たとえ多クラスへの分類がうまくいかず，ほぼランダムなクラス分類になっても補正パープレキシティは変化しなく悪影響はない．少しでもクラス分類がうまくできれば（言い換えれば，言語制約を反映したクラス分類）補正パープレキシティは小さくなり，音声認識率（登録した未知語の認識率）は向上する．

我々は既知語をLSAにより意味的，構文的に多クラスに分類しておき，未知語に対しては，Web上の未知語の出現するコンテキストと類似なコンテキストで出現する既知語を見つけ，その既知語のクラスと同一クラスに未知語を分類する手法を提案している．この手法に基づいて行った認識実験を図 3 に示す〔8〕．（あえて未知語を含む文をテスト文としたので認識率は相対的に悪い.）

本実験は，Wall Street Journal のトレーニングデータ（145,000 語の異なり単語数）から 20,000 語の言語モデルを作成し，OOV1 タイプの未知語を登録していった場合の結果である．トレーニングデータ中の未知語であるため，トレーニングデータから，この未知語を登録すれば単語トライグラムを求めることができるので，この場合が一番望ましい認識結果（上限値）である．テストデータのみに存在する OOV2 タイプの未知語に対しては，この未知語を登録しても単語トライグラムは求めることができないので，我々の提案法の有効性がわかる．

4.　サブワード単位の認識による未知語処理

(a) 大語彙連続音声認識

未知語をサブワードの系列で表現する方法は古くから行われてきた．例えば，大語彙連続音声認識システムで未知語クラスを予測した場合は，任意のサブワード系列と照合できるように拡張するのが一般的である〔9〕〔10〕．勿論，このときサブワード系列に対して言語モデルを組み込むことは有用である（クラス毎のサブワードのn-gramなど〔11〕）．日本語の場合，サブワードは“音素“や”音節・モーラ“がよく用いられる．新しい組織名などの入力には”基本単語“をサブワードとする方法もある〔12〕.

(b) 音声ドキュメント検索のための未知語処理

通常，音声ドキュメント検索のために，前処理として大語彙連続音声認識システムでテキストに変換しておき，以後は，テキスト検索技術を用いるのが一般的である．しかし，未知語は既知語のいずれかに誤認識されて正しく認識できないので，前節で述べたように未知語相当区間はサブワード列に変換して認識・テキスト化することが考えられる．しかし，この方法でも既知語として認識されたところは誤認識で，実際は未知語である場合も多々生じる．そこで，図 4 に示すように既知語に対しては，大語彙連続音声認識結果（単語ラティス，単語混同グラフ）を，未知語に対しては，サブワードの言語モデルに基づくサブワード系列認識結果（コンフュージョンネットワークやラティス表現）を用いるのが有用である〔13，14〕．もちろん，コンフュージョンネットワークやラティス表現でカバーできないサブワードにも対処する必要がある〔14〕．しかし，出現頻度の少ない未知語に対して，全音声ドキュメント区間を未知語対象区間とするのは得策ではない．そのため，次の二つの方法が考えられる．一つは大語彙連続音声認識結果の信頼度の高い区間（正しく認識される区間，もしくは，既知語が既知語と認識される区間）以外の区間を未知語対象区間とする方法である．二つめは，前節の未知語のサブワード系列つきの大語彙連続音声認識結果で，サブワ

ード系列を多めに出力するようにする方法である．

3 節で述べた未知語を認識辞書に登録して，音声認識する方法の代わりに，未知語はサブワード列に認識変換し，このテスト文のトピックに関連するWeb上のコーパスとこのサブワード列から未知語を推定する方法もある〔15〕．

なお，大語彙連続音声認識結果とサブワード系列認識結果を併用する方法は，既知語が検索語の場合，サブワード系列でも検索することにより，既知語の認識誤りに頑健になる利点もある．

図 5 は，CSJのコア講演（44 時間分）に対する図 4 に基づく既知語と未知語の検索結果である〔16〕．特に未知語は 1 時間に 1 回以下の出現頻度であるが，既知語と遜色のない結果が得られている．我々が提案している音節トライグラムによるインデキシングと検索法は音節列同士のDPマッチングによる検索法と遜色ない検索精度で，数百倍の高速検索が可能であり，44 時間の音声ドキュメントに対して約 2.5m秒で検索できる（図 6 参照）〔16〕．

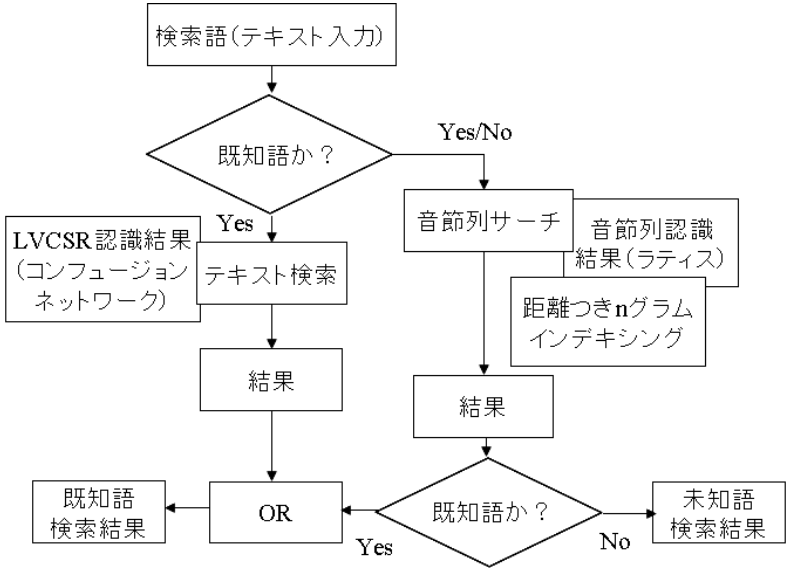
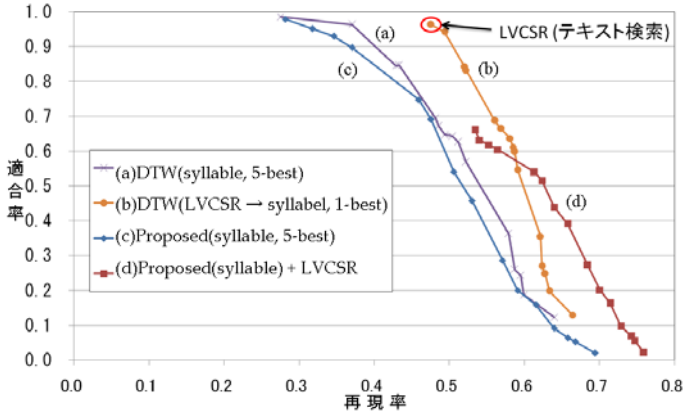
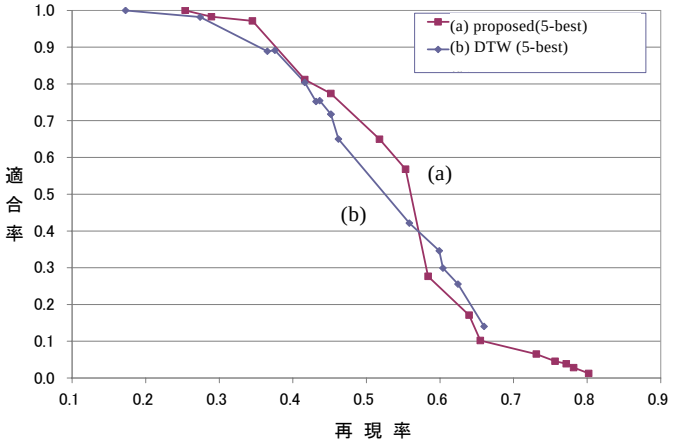


図 4　音声ドキュメント検索法

Figure 4　Spoken term detection method



（a）既知語検索



（b）未知語検索

図 5　音声ドキュメント検索結果

Figure 5　Retrieval results for spoken document

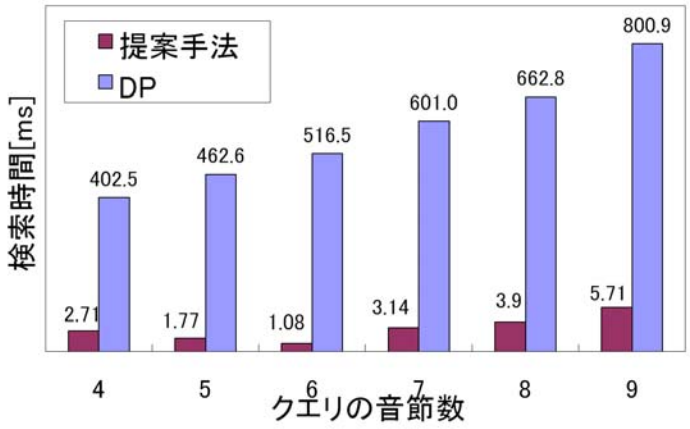


図 6　DTW と提案法の 1 ターム当りの検索時間
（検索対象音声ドキュメント：44 時間）
Figure 6　Retrieval time per term by DTW and proposed method

5.　システム全体から見た未知語処理の意義・役割

5.1　音声認識システム

姓名や施設名などの大語彙の単語入力の場合を除いて，講義や講演などの音声認識の場合，音声認識システムの認識用の単語数を十万語以上に大幅に増やせば，未知語率は多くの場合は 1%以下になると考えられる．一般に未知語があるとその周辺単語の認識も難しくなり，未知語率の 1.4 倍程度の誤りを生じることになる〔17〕．しかし，高精度な認識システムでも単語認識誤り率は 5〜10%は生じることを考えると，未知語の影響（1%程度の誤り率）は全体の誤り率と比べると比較的小さいと言える．テレビの字幕製作とか議事録作成とか，正確な書き起こしを要求される場合は別として，発話内容の把握には，80%程度の単語認識率があれば，十分という報告もある〔18〕．しかし，未知語は出現頻度が少なくても，内容のキーになる場合も多々ある（後述するように，検索語は未知語である場合が多い）．内容理解を助けるために，認識信頼度の低い区間には複数候補の表示も有用であろう〔19〕．また，信頼度の低い区間や未知語区間を音節列で認識できれば，たとえ音節認識率が 80%程度でも，読み手は正しい発話単語（未知語も含む）を推定できる可能性がある．ポッドキャストなどの音声認識の応用を考えると，流行に依存した新しい語や人名を逐次認識用辞書に登録していくことが望ましい場合もある〔20〕．

未知語の登録の際，読み付与の問題がある．かな列が与えられれば，英語のように grapheme-to-phoneme のような問題は日本語ではないが，漢字列表記だと困難が生じる．これは，テキストからの音声合成法の技術を流用することが考えられるが，読み付与が難しい場合は Web の利用など新しい研究課題になりうる．

以上の考察から，未知語を如何に認識用辞書に登録するか，その言語モデルを如何に算出するか，未知語区間を如何にサブワード列として認識するか，が重要と考えられる．

5.2　音声ドキュメント検索

音声ドキュメント検索，特にターム検索では，検索語が固有名や新出語のように認識用辞書に含まれていない（音声ドキュメントの書き起こしに正しく変換されていない）場合が多く，検索語の 4〜5 割が未知語であるという報告がある〔21〕．このような場合の検索法は重要である．大量の音声ドキュメントから高速に検索するためには，一般には 4 節（b）で述べたように，音声ドキュメントをサブワード列に認識・変換しておき，サブワード列同士のマッチングにより検索する．これをさらに高精度化する方法としては，次の二つの方法が考えられる．一つは，高速化・省メモリ化のために述べた 4 節（b）の未知語候補区間だけで検索する方法〔22〕，二つ目は，検出された未知語区間を，音声パラメータレベルで再検証する方法〔22〕，である．勿論，この他にも，検索された未知語周辺のコンテキストの妥当性に基づく検索結果の検証方法〔23〕，複数認識システムの統合化〔22, 24〕などもありうる．

6.　むすび

本稿では，音声認識と音声ドキュメント検索で問題となる未知語処理について述べた．基本的には音声の認識辞書の登録語彙を多くして未知語率を減少させるのが良いが，絶えず新しい未知語が創出され，全て登録することは不可能である．この場合は，サブワード単位系列として未知語を認識し，大量コーパスである Web 情報を用いてサブワード系列を認識用登録辞書以外の単語に変換するか，未知語をサブワード系列で表現し，サブワード系列同士のマッチングによって検索するのが有用である．未知語処理を効率化するためには，大語彙連続音声認識法をサブワード系列認識法の密な統合法が重要である．

なお，本稿で紹介した各手法の関連する文献は紙面の都合上，代表的なもの一つに限った．

参考文献

1) J. Ueberla: Analyzing a simple language model – some general conclusion for language models for speech recognition, Computer Speech and Language, Vol.8, No.2, pp.153-176 (1994)
2) 中川聖一, 赤松裕隆: 未知語を含む文集合のパープレキシティの算出法－補正パープレキシティー, 日本音響学会講演論文集, 2-1-13 (1998.9)
3) 森信介, 山地浩: 日本語の情報量の上限の推定, 情報処理学会論文誌, Vol.38, No.11, pp.2191-2199 (1997)
4) 森信介, 宅間大介, 倉田岳人: 確率的単語分割コーパスからの単語 N-gram 確率の計算, 情報処理学会論文誌, Vol.48, No.2, pp.1-8 (2007)
5) T. Ngyz, M. Ostendorf, M. Y. Hwangy, M. Siuz, I. Bulykoy, X. Leiy: Web-data augmented language models for Mandarin conversational speech recognition, Proc. ICASSP., pp.589-592 (2005)
6) 徳田翔, 西崎博光, 関口芳廣: 講義音声認識のための Web 文書を用いた言語モデルの適応化と語彙選択, 第 2 回音声ドキュメント処理ワークショップ講演論文集, pp.97-104 (2008)
7) W. Naptali, M. Tsuchiya, S. Nakagawa: Modeling out-of-vocabulary words using multi class-based n-gram language models for automatic speech recognition, 第5回音声ドキュメント処理ワークショップ講演論文集 (2011) ttp://www.cl.ics.tut.ac.jp/~sdpwg/sdpws2011_proceedings_pre/
8) W. Naptali: Study on n-gram language models for topic and out-of-vocabulary words, 豊橋技術科学大学博士論文 (2011) http://www.slp.ics.tut.ac.jp/
9) K. Kita, T. Ehara, T. Morimoto: Processing unknown words in continuous speech recognition, IEICE Trans., Vol.E74, No.7, pp.1811-1816 (1991)
10) 甲斐充彦, 廣瀬良久, 中川聖一: 単語 N-gram 言語モデルを用いた音声認識システムにおける未知語・冗長語の処理, 情報処理学会論文誌, Vol.40, No.4, pp.1385-1394 (1999)
11) 山本博史, 小窪浩明, 菊井玄一郎, 小川良彦, 匂坂芳典: 複数のマルコフモデルを用いた階層化言語モデルによる未知語認識, 電子情報通信学会論文誌, Vol.J87-D Ⅱ, No.12, pp.2104-2111 (2004)
12) 押川洋徳, 北岡教英, 中川聖一: 高頻度組織名と基本単語を用いた任意組織名入力インターフェース, 日本音響学会講演論文集, 2-5-10 (2005.3)
13) 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, 電子情報通信学会, 音声技法, SP2007-93 (2007)
14) 高橋将史, 藤井康寿, 山本一公, 中川聖一: 音声ドキュメントに対する未知語に頑健な検索手法の検討, 日本音響学会講演論文集, 3-Q-35 (2009.3)
15) C. Parada, A. Sethy, M. Predze, F. Telinek: A spoken term detection framework for recovering out-of-vocabulary words using the Web, Proc. Interspeech, pp.1269-1272 (2010)
16) 岩見圭祐, 藤井康寿, 山本一公, 中川聖一: 音節 n-gram インデックスによる未知語の音声検索法の改善, 第 5 回音声ドキュメントワークショップ講演論文集 (2011)
17) J.-G. Gauvain, L. Lamel: Large vocabulary continuous speech recognition: from laboratory systems towards ralworld applications, IEICE Trans. Vol.J79-DII, No.12, pp.2005-2021 (1996)
18) C.Munteanu, G. Penn, R. Baecker, E. Toms, D. James: Measuring the acceptable word error rate of machine-generated webcast transcripts, Proc. Interspeech, pp.157-160 (2007)
19) 後藤真孝, 緒方淳, 江渡浩一郎: PodCastle: ユーザ貢献により性能が向上する音声情報検索システム, 人工知能学会論文誌, Vol.25, No.1, pp.104-113 (2011.1)
20) 松原勇介, 緒方淳, 後藤真孝: ポッドキャスト音声認識の性能向上手法:集合知によって更新される Web キーワードを活用した言語モデリング, 情報処理学会, 音声言語情報処理, SLP-71(6) (2008)
21) 岩田耕平, 伊藤慶明, 小嶋和徳, 田中和世, 李時旭: 語彙フリー音声文書検索手法における新しいサブワード音響距離の有効性の検証, 情報処理学会論文誌, Vol.48, No.5, pp.1990-1999 (2007)
22) 西崎博光, 中川聖一: 音声認識誤りと未知語に頑健な音声文書検索手法, 電子情報通信学会論文誌, Vol.86-D Ⅱ, No.10, pp.1369-1381 (2003)
23) D. Schneider, T. Mertens, M. Larson, J. Kohler,: Contextual verification for open vocabulary spoken term detection, Proc. Interspeech, pp.697-700 (2010)
24) 名取賢, 西崎博光, 関口芳廣: 複数認識システムを用いた音声中の検索語検出の検討, 第 4 回音声ドキュメント処理ワークショップ講演論文集 (2010)