様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 ９ 日現在

機関番号：１３９０４
研究種目：若手研究(B)
研究期間：2012～2014
課題番号：２４７００１６７

研究課題名（和文）Ｓｕｆｆｉｘ　Ａｒｒａｙを用いた音声検索における高速アルゴリズムの研究とその検証

研究課題名（英文）On quick search algorithm for spoken term detection using suffix array and its practical evaluation

研究代表者
桂田　浩一（Katsurada, Kouichi）

豊橋技術科学大学・国際交流センター・准教授

研究者番号：８０３２４４９０

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究ではコンパクトなデータ構造であるsuffix arrayを利用した高速で大規模メモリを要さない音声検索語検出システムを開発した．このシステムではsuffix arrayを木構造に見立て，ルートノードからDP(Dynamic Programming)マッチングを適用することによって，インデックスサイズの小さい高速音声検索語検出を実現している．本検索システムの性能を評価するため，N-gramインデックスを用いた基礎的な手法および他の高速検索法と，検索精度を比較した．その結果，本システムでは検索精度を損なうことなく他の手法と同程度の高速検索を実現できている事が確認できた．

研究成果の概要（英文）：We developed a fast spoken term detection (STD) system using a suffix array that does not spend a large memory space. This system realizes quick and memory-saving search by regarding a suffix array as a tree and applying DP matching to all paths from the root node. In addition, to overcome the problem that search time exponentially increases according to the length of search keyword, the search keyword is divided into shorter sub-keywords and they are searched for instead of the original keyword. We evaluate our system by comparing it with a general N-gram-index based approach and some other fast STD methods. The experimental results show that our method achieves accurate search without losing quickness of search.

研究分野：Speech Processing

キーワード：Spoken Term Detection　Suffix Array

１．研究開始当初の背景

コールセンターの応対記録や各種会議の議事録，e-learning コンテンツなど，膨大な量の音声・動画像データが様々な形で蓄積されている．こうしたデータを利用する一つの方法が音声検索語検出である．音声検索語検出（Spoken term detection: STD）とは与えられたキーワードの出現箇所を音声データ内から検索する技術を指す．音声検索語検出に関する研究は近年盛んに行われており¥cite{秋葉 10}，2006 年に NIST の主催でベンチマークテストが行われたのを始めとして，日本においても NICIR9 のタスクに組み込まれるなど，共通のタスクによる客観的な評価が行われ始めている．音声検索語検出の実用化により，音声・動画像コンテンツの更なる有効活用が期待できる．

上述のコールセンターを始めとする音声・動画像データの場合，総時間が数百時間～数万時間に及ぶ場合も多い．こうした大規模な音声データを対象にした場合，検索が高速であることやメモリ消費量が少ないこともユーザビリティやコストの面から非常に重要である．このうち検索の高速性に関しては，これまでにも幾つかの提案がなされてきたが，特にここ数年の間に，数十～数千時間の音声データを対象に数ミリ秒～数秒で結果を出力する非常に高速な検索法が提案され始めている．筆者らもこれまで suffix array を利用した高速音声検索語検出システムを開発してきた．

２．研究の目的

Suffix Array はテキスト中の全ての suffix（接尾辞）を辞書順にソートしたデータ構造であり，suffix に対する index のみを配列として保持する．したがって非常にコンパクトで二分探索による高速検索に適している．しかし，suffix array 上での探索では完全一致検索が想定されており，認識誤りを含む音声ドキュメントを扱えなかった．そこで筆者らは DP マッチングを suffix array に適用する手法を導入し，音響環境の変化に強い音素弁別特徴を距離尺度に用いることで，誤認識を含む音声ドキュメントから高速に検索結果を得ることを可能にした．さらに，キーワードが長くなることによる検索時間の指数的増加を防ぐために，Navarro らが文字列検索を対象に提案したキーワードの分割検索法と同様の分割検索法を音声検索に組み入れた．

本研究では我々が開発してきた検索システムの性能を評価するために，CSJ コーパス（CORE 講演 44 時間，ALL 講演 604 時間）を対象に，検索精度，検索時間，メモリ使用量に関して他の手法と比較する．まず検索精度に関しては，N-gram インデックスを用いた最も基礎的な手法，NTCIR9 で提供されているベースライン，および他の代表的な高速検索法と再現率，適合率，MAP(Mean Average Precision：キーワード毎の平均適合率の平均)スコア，および最大 F 値に関して比較する．検索時間に関しては NTCIR9 参加グループのうち検索速度が高速なグループとキーワードあたりの検索時間の比較を行う．さらにメモリ使用量については，本システムで必要とする記憶容量を詳細に分析すると共に，NTCIR9 参加グループを含む代表的な高速検索法と比較する．

３．研究の方法

音声検索語検出は一般に索引付けと検索の二つのフェーズから構成される．索引付けのフェーズでは音声データを LVCSR 等によって単語列あるいは音素列等の中間表現に変換し，その後，検索に適したデータ構造に格納する．本システムでは音声データを音素列に変換し，suffix array に格納する．

Suffix array（接尾辞配列）は，テキスト中の全ての音素に対する index を格納した配列を，suffix（接尾辞）の辞書順にソートしたものである．ソート済みのデータ構造であるため，検索キーワードを効率的に見つけ出すことができる．

Suffix array では完全一致検索を想定している．このため，誤認識を含む音声認識結果を対象とするには何らかの曖昧検索技術を導入する必要がある．そこで筆者は山下らによって提案された，suffix array に DP マッチングを適用する方法を利用している．このアルゴリズムは辞書類似検索を行う error-tolerant recognition アルゴリズムを suffix array を用いて全文曖昧検索に拡張したものである．山下らのアルゴリズムでは，suffix array を木構造に見立て，木構造の根から全てのパスに対して始端固定の DP マッチングを行う．本アルゴリズムが出力する検索結果は元の音素列に対して連続 DP マッチングを適用したものと同一である．

ただし，このアルゴリズムでは，DP マッチングの実行中に枝刈りの閾値内のすべてのパスが保持されるため，閾値が大きいと探索空間および処理時間が指数関数的に増加することが，山下らによって確認されている．閾値は検索キーワードの長さに比例して増加させる必要があるため，検索キーワード長に対して指数的に処理時間が増大する．そこで，この問題を解決するためにキーワードを分割し，分割キーワードを元のキーワードの代わりに検索する手法を導入している．

本システムではキーワードを分割した場合に分割しない場合と同一の検索結果が得られるよう，下記の方針で検索を行っている．

- 閾値 T のキーワードを分割して検索する際に，累積音素間距離 T 以内のキーワードを見落とすことがないよう，分割キーワードの閾値を必要十分に大きな値にして検索を行う．
- 分割キーワードの検索結果から音声デ

ータ内の近傍領域にあるものをフィルタリングし，その中からキーワード全体として累積音素間距離が T 以内のものを検索結果として出力する．

この方針に基づき，本システムでは 4 ステップで検索を行う．検索は (I)分割，(II)検索，(III)フィルタリング，(IV)検証の各ステップから構成される．まず (I)分割のステップでは閾値 T のキーワードを n 個に分割する．続く (II)検索のステップでは，suffix array 上での DP マッチングによって n 個の分割キーワードの出現位置が音声データベースから検索される．次の (III)フィルタリングステップでは，音声データベース中の近傍領域に m 個（種類）以上の分割キーワードが検索された箇所が検出される．最後に (IV)検証のステップでは(III)で検出された領域とキーワード全体との DP マッチングを元の音素列上で行うことにより，キーワード全体として累積音素間距離が T 以内の箇所が抽出される．

筆者らは各分割キーワードに与える閾値を等しくする場合，次の式(1)の通り設定すると(II)検索のステップにおいて累積音素間距離が T 以内のキーワードを見落とさないことを示している．すなわち，式(1)の T_s のように分割キーワードの閾値を設定することが，分割しない場合と同一の結果を得るための条件である．

$$T_s = T / (n - m + 1) \qquad (1)$$

なお，本研究と同様に suffix array 上でキーワードを分割し，DP マッチングを行う場合の理論的な分析が Navarro らによって行われている．Navarro らも本論文と同様に，キーワードを分割検索しても検索結果が変化しない条件を示している他，検索速度の最適化のための条件についても示している．しかし，Navarro らは分割キーワードと検索箇所が完全一致しない場合に関しては， m=1 のときの条件のみを示している．また，Navarro らは文字列の検索を主目的としているため，DP マッチングの距離尺度が編集距離であることを前提としている．これに対して，式(1)は曖昧検索において m 個の分割キーワードが検出されるための条件を与えている点，および音素間距離を実数値（例えば混同行列から算出される値）に拡張できる点が異なる．

ここで，キーワードの長さを l とし，一音素あたりの閾値を t(すなわち t=T/l)とする．t をキーワード長に依存しない共通の閾値として検索を実行した場合，t が同じ値であっても短いキーワードの検索結果が多く出力されるという問題がある．これはキーワードが長いほど，それにマッチする音素列中の該当箇所が少なくなるためである．

そこで短いキーワードの結果出力を抑え，長いキーワードの結果を多く出力するために，キーワードの長さ l を考慮に入れた式(2)の score を共通の閾値として検索を実施した．

$$\text{Score} = 1 / (t/l^{1/2}+1) \qquad (2)$$

４．研究成果

実験は Intel Core i7-2600 プロセッサ 3.4GHz，メインメモリ 8GB を搭載した PC で行った．実験で用いた音声ドキュメントは NTCIR9 の STD ワーキンググループにより提供された CSJ コーパス CORE 講演 44 時間，および ALL 講演 604 時間である．それぞれについて，syllable-based transcription（音節言語モデルを用いて認識した結果）と word-based transcription（単語言語モデルを用いて認識した結果）が提供されている．CORE 講演，および ALL 講演にはそれぞれ 50 語の検索キーワードが設定されている．本システムでは音節・単語を音素に変換し，suffix array に格納したものを検索対象としている．

一般に DP マッチングの距離尺度として混同行列が用いられることが多いが，音響環境が異なると劣化が大きい．それに対して音素弁別特徴は環境が異なる場合も性能劣化が少ないという結果が得られている．音素弁別特徴とは調音様式・調音位置によって音素を特徴付けしたもので，+または－を取る 15 次元の素性により音素を表す．本実験では，この素性から音素間のハミング距離を求め，局所音素間距離として用いることにした．
また，予備実験の結果から，高速検索に最適なパラメータとして m=1，分割キーワードの音素数は 6 音素を採用した．DP マッチングにおける脱落，挿入ペナルティは 3.0 と設定した．まず，本システムの基本的な性能を確認するために，NTCIR9 の CORE 講演の syllable-based transcription を用いて，(A)高速検索の基礎技術としてしばしば用いられる N-gram インデックスベースの検索法，(B)NTCIR9 のベースライン，(C)t=T/l を閾値として用いる提案手法，および(D)式の score を閾値として用いる提案手法を比較した．比較尺度は再現率，適合率，MAP スコア，最大 F 値である．(A)の N-gram インデックスベースの検索法では音素 3-gram を用いた．キーワードに含まれる音素 3-gram のうち，多くの 3-gram が検出できた箇所が高いスコアを得るようスコアリングしている．(B)の NTCIR9 のベースライン，および(C)と(D)の手法では DP マッチングを用いているが，(B)は編集距離を距離尺度としているのに対し，(C)と(D)は音素弁別特徴に基づく距離尺度を用いている点が異なる．また，(C)と(D)では suffix array を構築している．

再現率-適合率曲線を図 1 に，MAP スコアと最大 F 値を表 1 に示す．これらの結果から分かるように，(B)～(D)の DP マッチングを用いた手法は(A)の音素 3-gram インデックスを用いた手法の結果を上回っている事が分

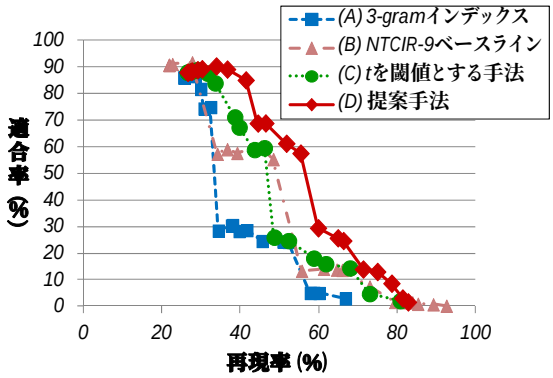

図 1：本システムの基本性能（再現率・適合率）

表 1：本システムの基本性能（最大 F 値等）

	MAP	最大 F 値
(A)音素 3-gram	47.5	45.2
(B)ベースライン	59.5	52.7
(C)提案手法(閾値:t)	65.4	52.0
(D)提案手法(score)	67.2	57.7

かる．音素 3-gram インデックスを用いた手法では，連続した三音素が正しく認識されなければ検出できず，1 箇所の誤りが前後 3 つの 3-gram の非検出につながる．これに対し，DP マッチングでは 1 音素の誤認識が大きな影響を与えない．このため，DP マッチングを用いた(B)～(D)の手法が良好な結果を残したと考えられる．次に，(B)と(C)を比較すると，(C)は最大 F 値で若干下回ったものの，MAP スコアでは大きく上回っていることが分かる．したがって音素弁別特徴の利用は検索精度の向上に一定の効果があったといえる．最後に，(C)と(D)を比較すると，(D)の score を用いた手法は(C)の t を用いた手法を，MAP スコアと最大 F 値で共に上回っていることが分かる．このことから，式(2)のキーワード長による補正の有効性を確認できた．

(A)の結果が示すように，N-gram インデックスのみでは十分な精度が得られない．そこで N-gram インデックスを用いた高速検索では精度向上のための様々な手法を組み入れている．例えば Iwami らは単語言語モデルと言語モデルを用いて得られた音声認識結果の 2 種類を用いている他，挿入，脱落，置換誤りに対応するよう N-gram インデックスを拡張している．一方，神田らはキーワード前後の文脈を考慮したリスコアリングにより精度を向上している．本節では本手法と同程度の高速検索を実現している Iwami らの手法(E)，Kaneko らの手法(F)，NTCIR-9 で最も良い検索性能を示した Nishizaki らの手法(G)，および高速検索を特徴の一つとして挙げている神田らの手法(H)と前節の提案手法(D)を検索精度に関して比較する．

実験の結果を図 2 と表 2 の(D)～(H)に示す結果から分かるように，(F)を除く全ての手法で最大 F 値が提案手法(D)の最大 F 値を上回っている．このうち(E)と(G)の手法が本システムの結果を上回った主な理由は，Iwami らの手法は 2 種類の異なるタイプの音声認識結果を，Nishizaki らも 5 種類の言語モデルと 2 種類の音響モデルから得られた 10 種類の音声認識結果を用いているためである．そこで本手法でも syllable-based transcription と word-based transcription の 2 種類を用いて suffix array を構築し，精度の向上を試みた．結果を図 2 および表 2 の(I)に示す．結果が示す通り，2 種類の認識結果を用いることにより，(G)，(H)の精度には及ばないものの最大 F 値を 10%以上向上させることができた．

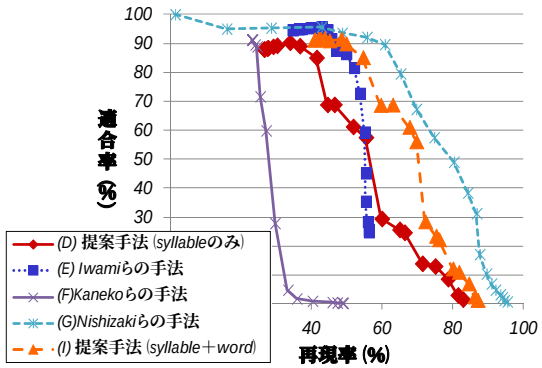

図 2：他の手法との比較（再現率・適合率）

表 2：他の手法との比較（最大 F 値等）

	MAP	最大 F 値
(D)提案手法(音節のみ)	67.2	57.7
(E)Iwami らの手法	49.1	64.5
(F)Kaneko らの手法	27.2	38.5
(G)Nishizaki らの手法	83.7	72.5
(H)神田らの手法	-	73.3
(D)提案手法(音節+単語)	74.0	68.1

5．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　7 件）

[1] Seng Kheang, Kouichi Katsurada, Yurie Iribe and Tsuneo Nitta: "Solving the phoneme conflict in Grapheme-To-Phoneme Conversion using a Two-Stage Neural Network-based approach", IEICE Transaction on Information and System, Vol.E97-D, No.4, pp.901-910 (2014-4).

[2] Narpendyah W. Ariwardhani, Yurie Iribe, Kouichi Katsurada and Tsuneo Nitta: "Mapping Articulatory Features to Vocal-Tract Parameters for Voice Conversion", IEICE Transaction on Information and System, Vol.E97-D, No.4, pp.911-918 (2014-4).

[3] 桂田 浩一，勝浦 広大，入部 百合絵，新田 恒雄：“Suffix Array を用いた高速音声検索語検出システムの性能評価”，電子情報通信学会論文誌，Vol.J96-D，No.10，

pp.2540-2548 (2013-10).
[4] **木村　優志，入部 百合絵，桂田 浩一，新田 恒雄**: “**調音特徴–声道音響パラメータ変換を用いた調音特徴運動** HMM **音声合成**”, **電子情報通信学会論文誌,** Vol.J96-D, No.5, pp.1356-1364 (2013-5).
[5] Narpendyah W. Ariwardhani, Masashi Kimura, Yurie Iribe, Kouichi Katsurada and Tsuneo Nitta, "Phoneme Recognition based on AF-HMMs with Optimal Parameter Set", Journal of Signal Processing, Vol.16, No.6, pp.571-579 (2012-11).
[6] **木村 優志，澤田 心大，入部 百合絵，桂田 浩一，新田 恒雄**：“**音声と画像シーンを用いた潜在意味解析に基づくタスク推定**”, **電気学会論文誌** C, Vol.132 No.9 pp.1473-1480 (2012-9).
[7] Yurie Iribe, Takuro Mori, Kouichi Katsurada and Tsuneo Nitta, "Generation of CG Animation Based on Articulatory Features for Pronunciation Training", The Journal of Information and Systems in Education, Vol.11, No.1, pp.1-13 (2012-5).

〔学会発表〕（計　7 件）

[8] Narpendyah W. Ariwardhani, Yurie Iribe, Kouichi Katsurada and Tsuneo Nitta: "Voice Conversion For Arbitrary Speakers Using Articulatory-Movement To Vocal-Tract Parameter Mapping", Proc. of MLSP2013, pp.1-6 (2013-9).
[9] Kouichi Katsurada, Seiichi Miura, Kheang Seng, Yurie Iribe and Tsuneo Nitta: "Acceleration of Spoken Term Detection Using a Suffix Array by Assigning Optimal Threshold Values to Sub-Keywords", Proc. of InterSpeech 2013, pp.11-14 (2013-8).
[10] Yurie Iribe, Silasak Manosavanh, Kouichi Katsurada, Ryoko Hayashi and Chunyue Zhu and Tsuneo Nitta, "Introducing Articulatory Ancho-point to ANN Training for Corrective Learning of Pronunciation", Proc of. ICASSP2013, pp.3716-3720 (2013-5).
[11] Silasak Manosavanh, Yurie Iribe, Kouichi Katsurada, Ryoko Hayashi, Chunyue Zhu and Tsuneo Nitta: "Articulatory Movements from Speech for Pronunciation Training", Proc. of ICCE2012, pp.499-507 (2012-11).
[12] Yurie Iribe, Takurou Mori, Kouichi Katsurada, Goh Kawai and Tsuneo Nitta: "Real-time Visualization of English Pronunciation on an IPA Chart Based on Articulatory Feature Extraction", Proc. of InterSpeech2012, Tue.P5d.01 (2012-9).
[13] Tsuneo Nitta, Silasak Manosavan, Yurie Iribe, Kouichi Katsurada, Ryoko Hayashi and Chunyue Zhu: "Pronunciation Training by Extracting Articulatory-Movement from Speech", Proc. of IS ADEPT, pp.75-78 (2012-6).
[14] Yurie Iribe, Silasak Manosavan, Kouichi Katsurada and Tsuneo Nitta: "Animated Pronunciation Generated from Speech for Pronunciation Training", Proc. of KES IIMSS 2012, pp.73-82 (2012-5).

〔その他〕
http://www.mmi.cs.tut.ac.jp/

６．研究組織

(1)**研究代表者**
桂田　浩一（KATSURADA , Kouichi）
豊橋技術科学大学・国際交流センター
研究者番号：80324490

(3)**連携研究者**
新田　恒雄（NITTA , Tsuneo）
早稲田大学・グリーン・コンピューティング・システム研究機構
研究者番号：70314101