連続音節認識結果の距離つきトライグラムアレイ化による未知語音声の超高速検索

中川聖一[†1]　岩見圭祐[†1]
藤井康寿[†1]　山本一公[†1]

情報通信技術の発展，高速化に伴い，現在はウェブ上に音声や動画像などのマルチメディアデータが多く存在する．ニュースや新聞記事のようにテキスト情報を含むものであれば既存のテキスト検索エンジンを用いることで，欲しい情報を高速に検索することができる．しかし，現在のところ音声ドキュメントに対しての有効な検索手法は確立されていない．その理由として挙げられるのが，未知語や認識誤りといった音声ドキュメント特有の問題である．本稿では、これらの問題に対し、連続音節認識結果の距離付きトライグラムアレイ化に基づいて、未知語に頑健な音声ドキュメントの超高速検索法を提案する。

Out-of-Vocabulary Term Detection by Trigram Array with Distance from Continuous Syllable Recognition Results

Seiichi Nakagawa,[†1] Keisuke Iwami,[†1]
Yasuhisa Fujii[†1] and Kazumasa Yamamoto[†1]

Recently, in accordance with development of information and communication technology, we can find many multimedia data such as audio and video on the Web.We can find the information with an existing textual search engine if the target data consist of text information such as news and newspaper, however, any efficient spoken document retrieval (SDR) method has not been established, because spoken documents have specific problems such as recognition errors and out-of-vocabulary(OOV) terms. The aim of this experiment is to develop a robust SDR method by using trigram array with distance from continuous syllable recognition results.

1. はじめに

本稿では，音声ドキュメント内の未知語を高速で検索する手法を提案する．未知語を検索可能にするために音節単位での認識結果を用いる．認識した結果から音節ラティスのトライグラムを作成し，そのトライグラムを用いてインデックスを構築する．インデックスは辞書順にソートしておくことで 2 分探索による高速な検索が可能となる．認識誤りへの対策としては，今までにも複数候補を用いて置換誤りに対処する手法が提案されているが，本稿ではこれに加え，挿入誤りと脱落誤りに対して対処する手法を実装した．また，これらの認識誤り対策の結果，検出候補が増加するという問題に対して，高速に検索を行うために認識誤り対策をどの程度おこなったのかという情報を距離として導入し，この距離をトライグラムに付随させることで検出候補を削減した．

2. 音声ドキュメント検索手法

2.1 大語彙連続音声認識の利用

従来の音声ドキュメント検索手法として，最も簡単な方法は，大語彙連続音声認識の書き起こしの結果に対して単語単位のテキスト検索を行う方法である．しかし，この方法では，未知語や，音声認識誤りの問題に対処することができない．未知語とは，大語彙連続音声認識の辞書にない単語のことである．辞書にない単語は，認識結果に現れることがないため，単語単位のテキスト検索では，未知語を検出することは不可能である．また，認識誤りの問題もある．認識誤りには，主に 3 つあり，異なる単語や音節に置き換えられて認識されてしまう置換誤り，実際に発話していない単語や音節が認識結果に現れる挿入誤り，そして実際に発話したのに認識されない脱落誤りである．認識誤りによって，辞書に登録されている単語であっても認識結果に現れない場合があり，その場合はテキスト検索を使用しても検索することができなくなってしまう．そのため，大語彙連続音声認識システムの性能によってテキスト検索の性能も決まってしまう．

2.2 サブワード単位認識の利用

未知語に対しては，サブワード列として音節単位で認識した結果を使用する．ドイツ語に対しては，5000 個の音節を用いた音節同士の重み付きレーベンシュタイン距離に基づく検

†1 豊橋技術科学大学
Toyohashi University of Technology

索方法が提案されている（およそ半分の単語が１音節語）[1]．なお，中国語は音節数が 416 個と少ないため，検索の基本単位としてよく用いられる[2]．また，置換，挿入，脱落誤りを考慮した音節列同士のマッチングに基づく検索方法も試みられている[3]．音素の n-gram を用いた検索方法も種々提案されてはいるが，基本的には bag of words の使い方で，音素認識誤りは考慮されていない[4][5]．

日本語の音節数は 100 余種類と比較的少なく扱いやすい．音節列として認識することによって，認識の際に単語辞書を使用しないので，文法の制約を無視でき，未知語の発音をそのまま認識できる可能性がある．そこで，音節単位で認識した音節ラティスをサブワード列として用意しておき，音節ラティスのトライグラムを用いる．手島らは音節認識結果をサフィックスアレイとしてテーブル化しておき，検索時に置換，挿入，脱落誤りを許しながらテーブルを探索する方法を提案している[6]．しかし，既知語に対しては，単語単位のテキスト検索よりも精度が低下することが知られている．そこで，本稿では，クエリが既知語の場合は，従来の大語彙連続音声認識の結果にテキスト検索を行い，クエリが未知語の場合のみ事前に用意しておいたサブワード（音節）の認識結果を用いて検索を行うこととする．

3. 提案する未知語検索手法

3.1 高速な未知語検索法[7]

本稿では，未知語に頑健な検索手法として，音節ラティスを使用して検索の際に認識誤りを考慮して検索を行う．本手法の概略を図 1 に示す．検索対象の音声ドキュメントに対して大語彙連続音声認識と連続音節認識を行い，インデックス化する．

未知語を検索可能にするため，サブワード列として音節ラティスの上位 3 ベストを使用する．そして，音節ラティスのデータを保持させておくデータ構造としてトライグラムアレイを定義する．トライグラムアレイとは，サフィックスアレイを元に，考えたデータ構造である．サフィックスアレイでは，音声ドキュメント内での出現位置情報のみを保持していたが，トライグラムアレイでは，出現位置と，そこで出現するトライグラムの情報も保持させておく (転置インデックス)．トライグラムアレイの作成方法の概要を図 2 に示す．トライグラムを辞書順に並べておけば，2 分探索で高速に検索できる．同じトライグラムが複数個連続してインデックス表に格納されることがあるが，これに対しては種々の改良法がある．また，トライグラムという固定長に限定しているため，トライグラムの種類とインデックスの位置が 1 対 1 に対応しているため，この関係を用いれば 2 分探索よりも高速に検索できる．4 音節長以上の検索後に対しては，図 3 のように 3 音節の複数の組に分割し，それぞれ

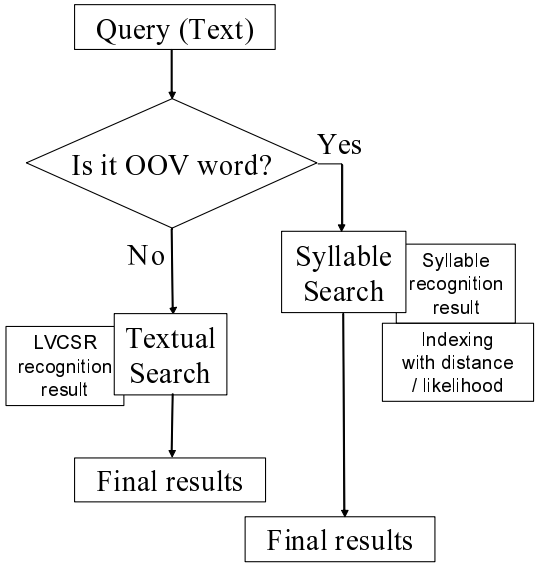


図 1　提案手法のフローチャート

で検索し，検索候補結果を連続性を考慮して，構成できるかどうかで，検索後の検索候補を出力する．

3.2 認識誤りに関する対策[7]

本稿では，それぞれの認識誤りに対処可能な方法を提案する．

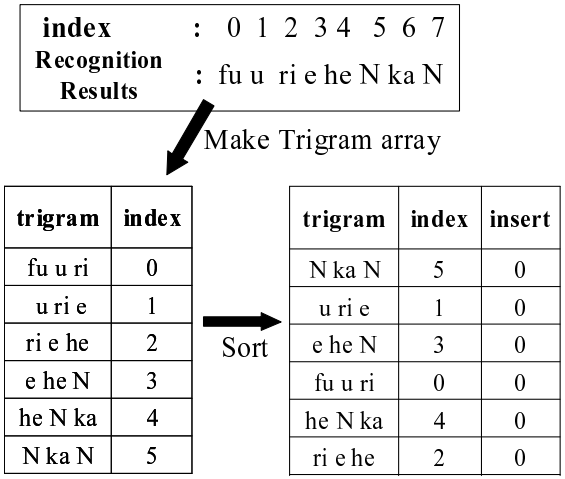
3.2.1 置換誤り対策

置換誤り対策としては，文献[7] に示されているように，音節ラティスの上位 3 ベストを用いる．本稿では索引を構築する際に音節ラティスの上位 3 ベストを組みあわせ，トライグラムを作成する．つまり，1 つのインデックスに対して $3 \times 3 \times 3 = 27$ 個のトライグラム情報を持たせる．

例えば,「フーリエ変換」の 1 ベストの認識結果が「フエキエヘンカン」になっていたとしても，3 ベスト中に正しい音節が含まれていれば検索することができる (図 4)．3 ベスト中に含まれない場合でも，次の挿入誤り対策と脱落誤り対策の併用で対処できる．

3.2.2 挿入誤り対策

挿入誤りに関しては，索引構築時に 1 音節飛ばしたトライグラムを作成することで対処



trigram	index
fu u ri	0
u ri e	1
ri e he	2
e he N	3
he N ka	4
N ka N	5

trigram	index	insert
N ka N	5	0
u ri e	1	0
e he N	3	0
fu u ri	0	0
he N ka	4	0
ri e he	2	0

図 2　トライグラムアレイ作成手順

図 3　検索後の分割例

する (3 連続音節に対して 1 箇所) ．先頭の音節を飛ばしたトライグラムは作成しない．したがって，1 つのインデックスに対し，2 つのトライグラムが索引に追加される (図 5).

3.2.3　脱落誤り対策

脱落誤りに関しては上記 2 つの対策とは違い，検索時にクエリを数音節脱落させたものを含めて検索することで対処する．したがって，1 つのクエリに対して複数回検索をおこなう．本実験でははは 4 音節以上のクエリに対しては 1 つの脱落，7 音節以上のクエリに対しては 2 つの脱落を許す．但し，脱落は連続する 3 音節以内に一ヶ所に制限する (図 6).

3.3　検出候補の削減

前述した認識誤り対策をおこなうことで認識誤りに頑健な検索をおこなうことができる．しかし，同時に湧き出し誤りが増加するという問題がある．そこで，DP マッチングによる後処理[7] 以外に，距離付きトライグラムアレイを提案する．

3.3.1　DP マッチングによる検出候補の削減[7]

DP マッチングを用いてクエリと検出結果との音節間距離を求め，その情報を元に検出候補を削減する．

図 7 の DP の制約条件 (1)，(2)，(3) は，

(1)　クエリと認識結果が一致した or 置換誤り

(2)　認識結果に挿入誤り

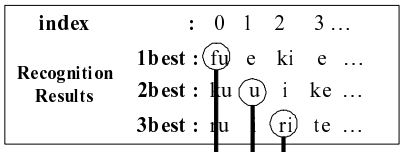
(3)　認識結果に脱落誤り

の場合のパスになっている．α と β はそれぞれ挿入コスト，脱落コストである．クエリと検索結果の距離を測るに当たり，音節間距離 d として Bhattacharyya 距離を使用した[7]．音節トライグラムによる検索結果として，クエリの出現位置候補を得ることができるので，その出現位置周辺を入力パターンとしてクエリとの距離を測る．そして，その距離がある閾値以下なら受理，より大きければ棄却することにする．

3.3.2　距離付きトライグラムによる検出候補の削減

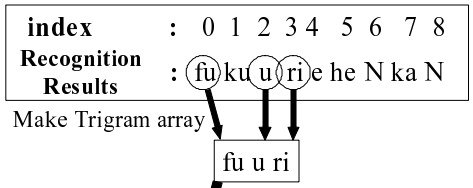
前述した DP マッチングによる検出候補の削減では，検出候補すべてに対して DP マッチングをおこなわなければいけないため，検索対象音声が長くなったり，検出候補が増えると処理時間が長くなるという問題がある．そこで，本実験では認識誤り対策をおこなう際にどれだけの誤りを許容したかという情報 (距離) を用いて，検索時に検出候補の削減をおこなう方法を提案する．距離を用いることで DP マッチングのように複雑な計算を行わずに，閾値との比較のみで検出候補の絞込みが高速におこなえる．

置換誤りの距離は 1 ベストからの音節間距離 (Bhattacharyya 距離) を使用した．1 ベストのみから生成されたトライグラムを規準とし，置換誤り対策によって生成されたトライグ



trigram	index	index	distance
fu e ki	0	0	0
:	:	:	:
fu u ri	0	0	d(e,u)+d(ki,ri)
ku e ki	0	0	d(fu,ku)
:	:	:	:

図 4　置換誤り対策例



trigram	index	insert
fu ku u	0	0
fu u ri	0	1
fu ku ri	0	1
ku u ri	1	0
:	:	:

図 5　挿入誤り対策例

ラムの各音節との距離の合計を置換誤りの距離とする．つまり，1 ベストの音節の距離は 0 となる．

次に，挿入誤りの距離は挿入誤り対策をおこなったかの有無を 0，1 の 2 値で表現する．

最後に脱落誤りの距離に関しては，検索時に脱落させた音節数を脱落誤りの距離とする．

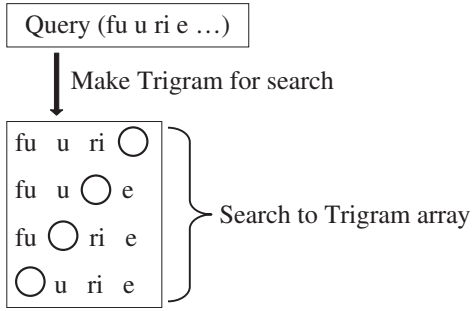
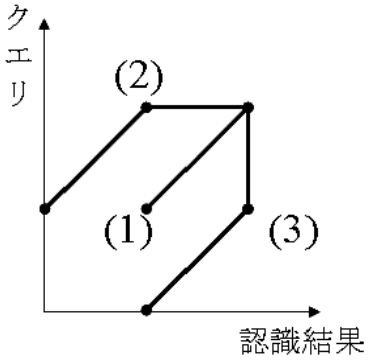


図 6　脱落誤り対策例 (○印は脱落)



$$g(i,1) = min \begin{cases} (1)\, g(i-1,j-1) \\ \quad +d(i,j) \\ (2) g(i-2,j-1) \\ \quad +(d(i-1,j) \\ \quad +d(i,j))/2+\alpha \\ (3) g(i-1,j-2) \\ \quad +d(i,j) \\ \quad +d(i,j-1)+\beta \end{cases}$$

図 7　DP の制約条件と漸化式

本実験では最大 2 音節の脱落を許すため (3 連続音節につき 1 箇所)，0，1，2 の 3 値となる．

4. 評 価 実 験

4.1 実験データ

実験データには CSJ(日本語話し言葉コーパス) のコアデータ 44 時間分を用い，本研究室で開発された SPOJUS による大語彙連続音声認識結果と連続音節認識結果を対象とし，検索，評価をおこなった[8]．LVCSR の辞書（20000 語）には，コア講演以外の CSJ2702 講演を学習データとし，カットオフを 4 とした．したがって，出現回数が 4 回以上あるものは未知語にはならない．今回検索語セットは伊藤らの報告[8] にあるものを使用した．音響モデ

表 1　認識結果 (%)

出力	Del	Ins	Subs	Corr	Acc
音節 1best	6.0	2.7	16.2	77.8	75.1
音節 3best	5.2	2.0	7.6	87.3	85.3
単語認識	5.4	4.6	22.7	71.9	67.3
音節に変換	3.6	15.3	13.5	83.0	67.6

ルは左コンテキスト依存音節モデル (928 個) で，コア講演を除いた CSJ の 2525 講演から学習した．連続音節認識には音節の 4 グラムの言語モデルを用いた．連続音節認識による音節認識率と大語彙連続音声認識 (LVCSR) の結果による単語認識率と音節列に変換後の音節認識率を表 1 に示す (単語正解率は 72%)．連続音節認識の第 1 候補の結果と LVCSR の結果から変換した音節認識結果は，ほぼ同程度である．連続音節認識結果の第 3 候補までを考慮すると音節認識結果の正解率は 87%とかなり高い．

4.2　既知語検索結果

既知語クエリ 39 種類 478 個を用いて性能評価をおこなった．それぞれの対策をおこなった際の検索結果を連続音節認識の結果を用いた場合を表 2，LVCSR の結果を用いた場合を表 3 に示す．①が置換誤り対策，②が挿入誤り対策，③が脱落誤り対策を示している．大語彙認識は 1best のみでしか実験をおこなっていないため，置換誤り対策の部分の結果は載せていない．表 2 の「対策なし」「②挿入」「③脱落」「①+②」は音節認識結果の第一候補のみを使う方法で，表 3 と直接比較できる．表 2 の検出候補結果を，DP マッチングで絞った結果を表 4 に，距離付きトライグラムによって検出した結果を表 5 に示す．DP マッチングと距離付きトライグラムによる絞込みに，性能差は見られなかった．最もよい再現率を示したのは LVCSR の結果を音節列に変換し，提案手法を用いた場合であった．テキスト検索では，湧き出し誤りが発生しにくく，今回の実験では適合率が 100% を示した． 再現率に関しても，最も良い条件と比べても 2% 程度の差しかない．再現率と適合率のバランスを考えると，既知語検索は LVCSR をかな漢字で検索した場合（テキスト検索）がよいと考えられる．

4.3　未知語検索結果

未知語クエリ 43 種 142 個を用いて性能評価をおこなった．それぞれの対策をおこなった際の検索結果を連続音節認識の結果を用いた場合を表 6，LVCSR の結果を用いた場合を表 7 に示す．最も再現率を増加させているのは認識誤り対策では置換誤り対策であることが表 6 から分かる．同時に適合率を低下させているのも置換誤り対策である．全ての認識誤り対策をおこなうことで約 50% の未知語を検出できた．また，連続音節認識の結果を用いた場合と，LVCSR の結果を用いた場合を比較すると，挿入誤り対策と脱落対策誤りを組み合わせた場合 (②+③)，再現率，適合率ともに連続音節認識の結果が良いという結果になった．このことから，未知語に関しては連続音節認識の結果を用いたほうがよいことが分かる．また，未知語の検索結果 (認識誤り対策有り) に対して DP マッチングで検出候補を絞った結果を表 8 に示す． DP マッチングにより，湧き出し誤りを約 90% 程度削減することができた．その際に正解部分も約 27% 減少した．未知語の検出結果を距離を用いて検出候補を

表 2　既知語の誤り対策別の検索結果 (連続音節認識)

既知語	音節認識							
	対策無し	①置換	②挿入	③脱落	①+②	①+③	②+③	①+②+③
検出数	50	155	59	255	232	2871	457	9811
正解数	41	105	48	94	118	211	117	260
再現率	0.09	0.22	0.10	0.20	0.25	0.44	0.24	0.54
適合率	0.82	0.68	0.81	0.37	0.51	0.07	0.26	0.03

表 3　既知語の誤り対策別の検索結果 (大語彙認識)

既知語	大語彙認識結果 → 音節列					テキスト検索
	対策無し	①置換	②挿入	③脱落	②+③	
検出数	246	-	248	557	850	233
正解数	192	-	192	245	250	233
再現率	0.40	-	0.40	0.51	0.52	0.49
適合率	0.78	-	0.77	0.44	0.29	1.00

表 4　既知語検索結果の DP マッチングによる絞込み (連続音節認識)

既知語	絞り込み無し	閾値①	閾値②	閾値③	閾値④	閾値⑤
検出数	9811	5858	4661	4009	2813	843
正解数	260	251	247	245	237	199
再現率	0.544	0.525	0.517	0.513	0.496	0.416
適合率	0.027	0.043	0.053	0.061	0.084	0.236

表 5　既知語検索結果の距離付きトライグラムによる絞込み（連続音節認識）

未知語	絞り込み無し	閾値①	閾値②	閾値③	閾値④	閾値⑤
検出数	9811	6017	4844	4082	2604	715
正解数	260	255	251	248	238	169
再現率	0.544	0.533	0.525	0.519	0.498	0.354
適合率	0.027	0.042	0.052	0.061	0.091	0.236

表 6　未知語の誤り対策別の検索結果 (連続音節認識)

未知語	音節認識							
	対策無し	①置換	②挿入	③脱落	①+②	①+③	②+③	①+②+③
検出数	61	964	101	323	1980	4521	609	10522
正解数	24	49	25	38	53	72	46	80
再現率	0.17	0.35	0.18	0.27	0.37	0.51	0.32	0.56
適合率	0.39	0.05	0.25	0.12	0.03	0.02	0.08	0.01

表 7　未知語の誤り対策別の検索結果 (大語彙認識)

未知語	大語彙認識結果 → 音節列				
	対策無し	①置換	②挿入	③脱落	②+③
検出数	26	-	45	255	574
正解数	17	-	17	30	35
再現率	0.12	-	0.12	0.21	0.25
適合率	0.65	-	0.38	0.12	0.06

表 8　未知語の DP マッチングによる絞込み
(連続音節認識)

未知語	絞り込み無し	閾値①	閾値②	閾値③	閾値④	閾値⑤
検出数	10522	6503	5353	4202	2518	819
正解数	80	79	78	76	73	58
再現率	0.563	0.556	0.549	0.535	0.514	0.408
適合率	0.008	0.012	0.015	0.018	0.029	0.071

表 9　未知語検索結果の距離付きトライグラムに
よる絞込み（連続音節認識）

未知語	絞り込み無し	閾値①	閾値②	閾値③	閾値④	閾値⑤
検出数	10522	6074	5255	4439	2464	768
正解数	80	78	78	77	72	64
再現率	0.563	0.549	0.549	0.542	0.507	0.451
適合率	0.008	0.013	0.015	0.017	0.029	0.083

絞った結果を表 9 に示す．また，DP マッチングと検出数をほぼ同等にした場合に比較した結果を図 9 に示す．DP マッチングと距離付きトライグラムによる絞込みに，性能差はほとんど見られなかった．

4.4　検 索 時 間

基本的にクエリの音節数が少ないほど湧き出し誤りが多くなる傾向があり，検出候補数が増えると処理時間も増加する．また，クエリが 7 音節以上の場合，脱落を 2 つまで許すため，検索回数が増え，処理時間が増加する．44 時間の音声データに対して，検索後の音節

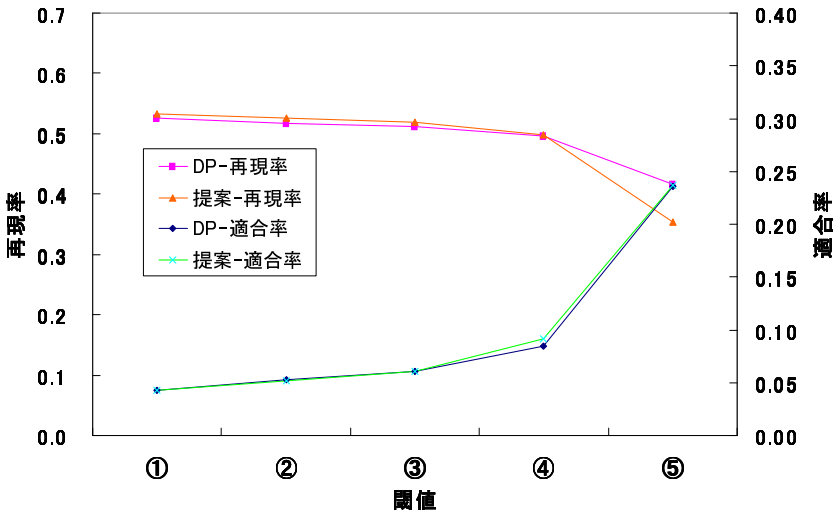


図 8　既知語検索結果の絞込み比較

数毎のトライグラムアレイによる検出候補数と従来法 (DP マッチングによる絞込み) と提案手法の検索時間の比較を図 10 に示す．検出候補数が多い程，従来法と提案手法の検索時間の差は大きい．DP マッチングの平均検索時間が 15[ms] なのに対し，距離を用いた場合の平均検索時間は 2.5[ms] となっている．この結果から，DP マッチングは検出候補数の影響を大きく受け，検出候補が多いと時間がかかることがわかる．一方，提案手法である距離を用いると，検出候補数が増えた場合においても高速に検索が可能である．この検索時間の差は，検索対象音声の時間が大きくなると顕著になる（前者は線形に増加，後者は対数スケールで増加）．

5.　お わ り に

本稿では，音節認識率 77% (3 ベストで 87%) の音声ドキュメントから約 50% の未知語を検出することに成功した．湧き出し誤りに関しても距離を用いて検出候補を絞り込むことで高速に抑えることができた．今後の課題として，検索精度と絞込みの精度の改善が挙げら

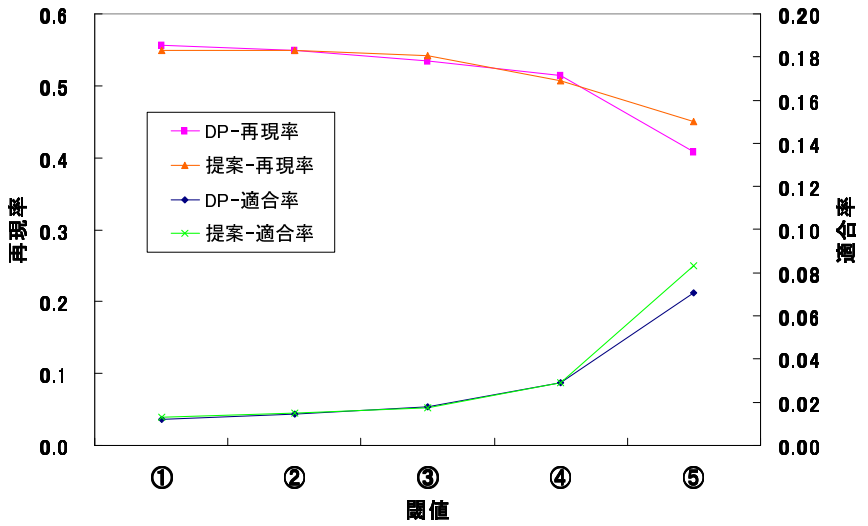


図 9　未知語検索結果の絞込み比較

れる．検索精度に関しては大語彙連続音声認識の単語認識結果の信頼度の低い区間だけを未知語候補区間とする方法が考えられる[9]．距離付きトライグラムに代わる検索結果の絞込みに関しては，認識の際に出力される尤度を用いて尤度つきトライグラムで絞込みをおこなうなどの方法が考えられる．

参　考　文　献

1) Martha Larson, Stefan Eickeler, “Using syllable-based indexing features and language models to improve German spoken Document Retrieval“ proc. EuroSpeech, pp. 1217 – 1220(2003)
2) Hsin-min Wang,“Experiments in syllable-based retrieval of broadcast news speech in Mandarin Chinese “ Speech Communication,Vol 32,pp. 49 – 60(2000)
3) Martin Wechsler, Eugen Munteanu,Peter Schauble,“New techniques for open-vocabulary spoken document retrieval“ Proc.SIGIR ,pp. 20 – 27(1998)
4) Corinna Ng,Ross Wilkinson,Justin Zobel, “Experiments in spoken document retrieval using phoneme n-grams “ Speech Communication,Vol 32,pp. 61 – 77(2000)
5) S Dharanipragada, S Roukos,“A multistage algorithm for spotting new words in speech“ IEEE Transactions on Speech and Audio Processing, vol.10, No.8, pp. 542 – 550,(2002)
6) 手島茂樹, 桂田浩一, 新田恒雄, “Suffix Array を用いた音声文書の高速検索 “ 第 3 回音声ドキュメント処理ワークショップ，pp. 29 – 32(2009.2)
7) 中川聖一, 高橋将史, 他, “未知語に頑健な音声ドキュメント検索手法の検討 “ 第 3 回音声ドキュメント処理ワークショップ論文集, pp. 7 – 14(2009.2)
8) 伊藤慶明, 西崎博光, 他.“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, SLP-78 No.4(2009)
9) 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-DII, No.10, pp. 1369 – 1381(2003)

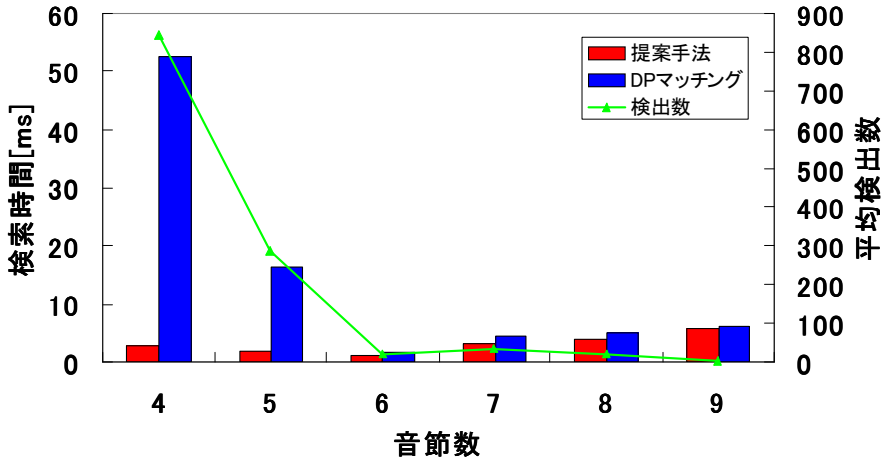


図 10　DP マッチングと提案手法の検索時間