分布間距離ベクトル表現による音響的類似度を利用したテキスト及び音声クエリからの音声検索語検出の改善

牧野光晃[1,a)]　山本 直樹[2,b)]　甲斐 充彦[2,c)]

概要：近年, 音声ドキュメント検索技術に関連した研究として，与えられた検索語が発話されている箇所を音声ドキュメント中から特定する音声検索語検出 (Spoken TermDetection : STD) の研究が盛んに行われている．最も基本的なアプローチとして，音声認識結果を元にサブワード（音素や音節）列としてインデキシングを行い，誤認識や未知語の問題に対処するためサブワード単位の誤りを許容して検索語との類似性を評価する方法が用いられることが多い．以前に我々はサブワード単位音響モデルの分布間距離ベクトルという構造的特徴表現に基づく音響的類似度を検索語検出スコアの評価のために用いることを提案し，テキストクエリによる STD で検索性能を改善した．本稿では，提案した特徴表現による音響的距離尺度の定式化の違いによる影響や，サブワード n-gram 信頼度を用いた STD 手法との併用手法の違いによる影響を含め，検索精度改善のアプローチの比較分析結果について報告する．また，テキストクエリによる STD で検索性能を改善した分布間距離ベクトルという構造的特徴表現に基づく検出手法が，音声クエリの STD に対しても有効であるか検証を行った結果についても述べる．

キーワード：音声検索語検出，分布間距離，構造間距離尺度，音響的類似度，音声クエリ

1. はじめに

ユーザが入力した検索語 (クエリ) に対して，音声ドキュメント中から検索語が話されている箇所を特定することを，音声検索語検出 (Spoken Term Detection : STD) と呼ぶ．国立情報学研究所を中心に 2011 年から開催されている NTCIR Workshop では，SpokenDoc サブタスクの一部として STD タスクが設定され，2013 年まではテキストクエリを想定した STD タスクの評価が行われてきた [3], [4].

STD タスクに対する一般的なアプローチは，音声データを大語彙音声認識システムを用いて得られた複数候補 (N-best リスト，ラティス等) を使用して検索を行うというものである [5]．STD タスクでは認識された単語列をサブワード列に変換した結果を利用する等，サブワード列のインデキシングに基づく方法が有効と考えられ，従来提案されている多くの STD タスクの検索手法 [2], [6], [7], [8], [9] や，SDR タスクの検索手法 [10] 等でも検討されてきた．このような音声認識結果のサブワード列に対する検索では，認識誤りを許容して類似箇所を検出する方法が工夫されてきた．

我々の先行研究では，サブワード列間の照合に用いる音響的類似度評価の改善に焦点を当て，サブワード単位の音響モデルから導出できる，分布間距離ベクトルと呼ぶ構造的特徴表現を用いることを提案した．そして，サブワード系列間の照合を HMM 状態系列間の照合に拡張し，分布間距離ベクトルに基づいて定義される局所距離を用いた照合手法によって，テキストクエリによる STD で検索性能を改善した [11], [12]．また，DP マッチング法とは異なるもう一つの有望なアプローチで，自動音声認識システムから得られるサブワード n-gram の信頼度スコアを照合に利用する方法との併用によって，更に検索性能を改善できることを示した [13].

本稿では，まず先行研究で提案した HMM 状態系列のレベルでのリスコアリング手法に関して，系列レベルでの音響的距離尺度の定式化法の違いによる検索性能への影響や，検索閾値などの調整パラメータをオープンまたはクローズドの条件で最適化した場合の影響について，NTCIR SpokenDoc の一部の STD 用データセットを用いた評価実験結果に基づいて示す．また，分布間距離ベクトルによる構造的特徴表現に基づく検出手法を，新たに用意した音声

[1] 静岡大学工学部
Faculty of Engineering, Shizuoka University
[2] 静岡大学大学院・工学研究科
Graduate School of Engineering, Shizuoka University
[a)] makino@spa.sys.eng.shizuoka.ac.jp
[b)] yamamoto_nao@spa.sys.eng.shizuoka.ac.jp
[c)] kai@sys.eng.shizuoka.ac.jp

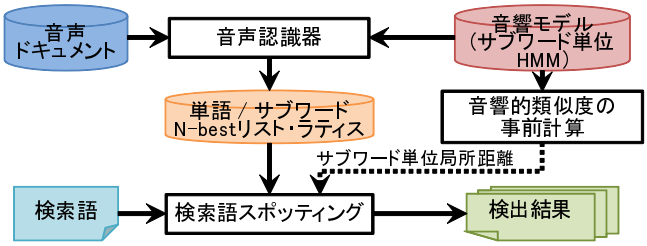

図 1　音声検索語検出のベースラインシステムの構造

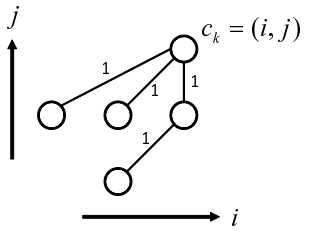

図 2　非対称 DP パス

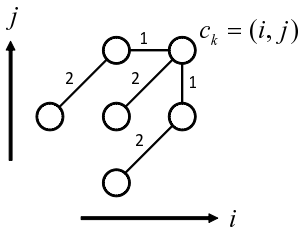

図 3　対称 DP パス

クエリを用いた STD タスクに適用した場合の評価実験結果を示し，提案手法の有効性について議論する．

2. サブワード間の音響的距離尺度に基づく STD 手法

STD に対する一般的なアプローチは，音声データを音声認識器に通してテキスト化し，サブワードレベルでの認識誤りを考慮した検索語との照合手法により検索を行うというものである．その手法の一つとして，連続 DP マッチングという手法が挙げられる．本節では，はじめに連続 DP マッチングを利用したベースライン STD システムと，そこで局所距離として使用する音響的な類似性を考慮したサブワード間距離尺度について述べる．

2.1 ベースライン STD システム

本研究で使用する音声検索語検出のベースラインシステムの構成を図 1 に示す．このシステムでは，あらかじめ音声データを自動音声認識システムによりテキスト化し，データベースに蓄積する．音声認識は，単語単位とサブワード単位の 2 種類の N-gram 言語モデルによる認識結果をそれぞれ求め，単語ベースの認識結果はさらにサブワード系列に変換して蓄積する．

検索を行う際は，テキストクエリの場合は入力された検索語をサブワード系列に変換する．音声クエリの場合は 6 節で詳述するが，自動音声認識システムによるクエリ音声の認識結果に基づいて，同様にサブワード系列に変換する．既知語から成る検索語の場合は単語認識データベース内のサブワード列，未知語から成る検索語の場合は音節認識データベース内のサブワード列と連続 DP マッチングを行う．そして，連続 DP マッチングの結果，非類似度スコア (マッチング距離) が閾値以下である解候補区間を検出結果として出力する．

連続 DP マッチングの際には図 2 に示す非対称の DP パスの制約を用いる．また，連続 DP マッチングで使用する局所距離には次節で述べるサブワード対の音響的な非類似度を使用する．なお，次節以降で用いる照合手法と区別するため，ここで述べる方法を **Baseline** 手法または **DP-SubwordBD** 手法と呼ぶことにする．

2.2 サブワード対の音響的な非類似度

サブワード単位の音響的な類似度としては，サブワード対の非類似度をサブワード単位 HMM の分布間距離 (Bhattacharyya 距離) に基づいて計算する例 [6] や，音素弁別特徴に基づく距離尺度を利用する例 [7] 等がある．我々は，前者の方法と同様に HMM のパラメータから求まる Bhattacharyya 距離に基づく分布間距離を利用してサブワード間の音響的な非類似度を算出する．一般的に，一つのサブワード HMM は複数の状態からなり，それぞれの状態の出力分布は混合ガウス分布 (GMM) でモデル化される．そこで，まず状態間の分布間距離を，混合成分間の Bhattacharyya 距離の最小値として定義する．つまり，あるサブワード a の HMM 状態 i において n 番目の混合成分の確率分布を $P_a^{\{i,n\}}$ と表わすと，サブワード a の HMM の状態 i とサブワード b の HMM の状態 j との距離を次式で定義する．

$$BD(P_a^{\{i\}}, P_b^{\{j\}}) = \min_{x,y} BD(P_a^{\{i,x\}}, P_b^{\{j,y\}}) \qquad (1)$$

そして，任意のサブワード対に対して，分布間距離を局所距離とし，図 3 に示す DP パスの制約を用いて状態系列間の DP マッチングを行うことにより，サブワード間のマッチング距離を求める．このようにして，あらかじめサブワード単位の音響モデルのみで求めておくことができるサブワード間非類似度を局所距離として使用し，2.1 節で述べたようにクエリと類似する区間の検出 (スポッティング) を行う．なお，NTCIR-9 及び 10 のベースライン評価ではサブワード単位の DP マッチングの局所距離として編集距離が用いられるが [3], [4]，それ以外はほぼ同等の手法といえる．

3. HMM 状態系列レベルでのリスコアリングに基づく STD 手法

我々は先行研究 [11], [12] において，前節で述べた方法で得られるクエリに音響的に類似した候補区間に対して，さらに詳細なスコア付け (リスコアリング) を行うことの有効性を示した．我々の提案手法では，クエリおよび候補区間のサブワード列を，それぞれに対応する HMM の状態系列に拡張した時系列の表現に変換して照合を行う．

この節では始めに，状態系列レベルのリスコアリングを併用する STD システムの概要について述べる．次に，

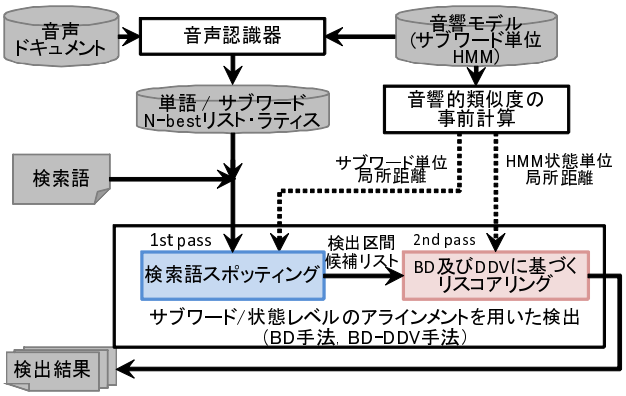

図 4　状態レベルのリスコアリングを用いる STD システム構成

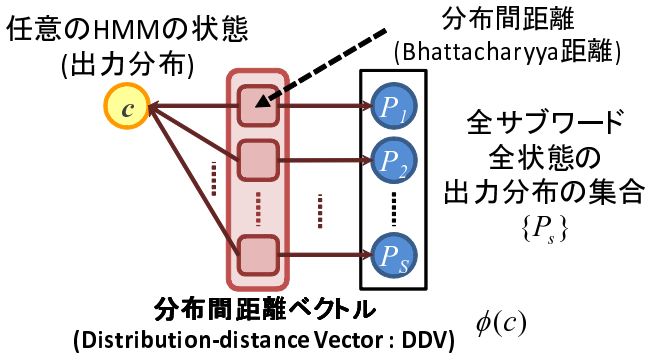

図 5　任意の HMM 状態に対して導出される分布間距離ベクトルの概念図

HMM 状態系列レベルでの照合方法として前節での方法を単純に状態レベルでの DP マッチングに拡張して適用する場合について述べる．最後に，分布間距離ベクトルと呼ぶ状態レベルの特徴表現 [11], [12], [13] と，それを用いた状態系列の距離尺度について比較する定式化法について述べる．

3.1　リスコアリングを併用する STD システムの概要

本節で述べる STD システムは，1 パス目は前節で述べたサブワード系列のスポッティング手法による候補区間検出を行い，2 パス目で HMM 状態レベルの詳細なスコア付けを行うためのリスコアリングを行う（図 4）．

検索語検出の手順は以下の通りである．

(1) 検索キーワードを音節列に変換し，事前に大語彙認識しておいた単語ベース認識結果 (音節列に変換しておく) もしくは音節認識結果に対して，連続 DP マッチングによるスポッティングを行う．

(2) スポッティングによるマッチング距離があらかじめ定めた閾値以内のサブワード列区間を抽出する．

(3) クエリとそれに対して抽出された区間のそれぞれのサブワード列に対応する音響モデル (HMM) の状態系列に対して，後述するリスコアリング手法によって新たな検出スコアを求める．

(4) 新たに求めた検出スコアが，あらかじめ定めた閾値より小さい解候補区間を検出結果として出力する．

以上の 1・2 の手順を 1 パス目，3〜4 の手順を 2 パス目とする 2 段階の処理から成る．このシステムの 1 パス目は 2 節に示した音声検索語検出システムと同じものである．

なお，本研究では音声クエリによる STD も取り扱うが，その際は上記の手順 1 において音声クエリを音節認識することで音節列に変換している．

3.2　状態単位の分布間距離に基づくリスコアリング

比較を行う 2 つのサブワード列を，それぞれ対応する HMM の状態系列 $A = \{a_1, \cdots, a_I\}$, $B = \{b_1, \cdots, b_J\}$ に展開する．ここで，実際には一方がクエリ，もう一方が音声ドキュメントとすると，連続 DP マッチング法としてこれ以降の手順をより一般化して扱うことができるが，比較を簡単化するため DP マッチングにより非線形時間伸縮を行い，2 つの HMM の状態系列の長さが等しくなるようにアライメントを行う．このときの局所距離としては式 (1) の Bhattacharyya 距離を用い，アライメントされた HMM の状態系列対を $F = \{c_1, \cdots, c_K\}$, $c_k = (a_{i_k}, b_{j_k})$ とする．この際に得られる DP マッチングスコアを $Score_{BD}$ とする．また，このリスコアリング手法のみを Baseline 手法 (DP-SubwordBD) と併用する場合を，**DP-StateBD** 手法と呼ぶことにする．

3.3　分布間距離ベクトル (DDV) 表現に基づくリスコアリング

2.2 節で定義した分布間距離 $BD(P_a^{\{i\}}, P_b^{\{j\}})$ は，2 つのサブワードに対応する HMM の状態間の該当分布間の距離のみを利用して求められた．分布間距離ベクトル表現は，2 つの異なる音節のある状態間の距離を定義するときに，他の音節の状態との距離にも着目し，それらの距離を要素として含む構造的な特徴表現である [12]．全てのサブワードに対応する HMM の全状態の出力分布の集合を $\boldsymbol{P} = \{P_s\}(s = 1, 2, \cdots, S)$ とすると，その中の任意の状態 c に対する特徴表現として以下のベクトル表現を用いる．

$$\boldsymbol{\phi}(c) = (D_{BD}(P_c, P_1), D_{BD}(P_c, P_2), \cdots, D_{BD}(P_c, P_S))^{\mathrm{T}} \quad (2)$$

このベクトルは，ある HMM の状態 c の出力分布と，自身を含む全ての状態の出力分布との分布間距離を要素に持っていることから，分布間距離ベクトル (Distribution-distance vector : DDV) と呼ぶ（図 5）．

このような構造的特徴を利用する考え方は，峯松らが提案している音声の構造的表象 [14], [15] を用いる考え方と関連しており，そこで指摘されているように伝達特性や話者固有の変動の要因に対する頑健性が期待できる．また，語彙を仮定しない音声ドキュメント中の類似部分検出方法としても同様な方法で有効性が示されている [16]．

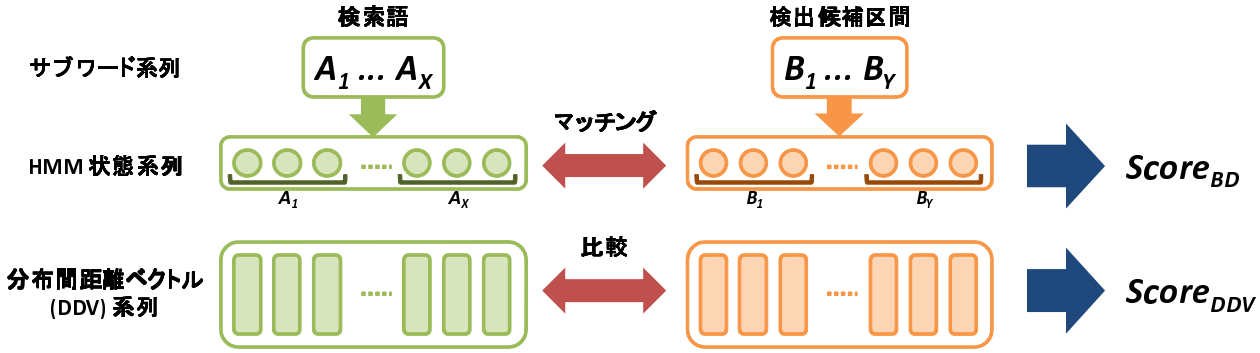

図 6　HMM 状態単位の距離尺度に基づくクエリの照合スコアの算出過程の概念図

分布間距離ベクトルを新たな特徴量表現と考えると，2 つの HMM 状態の対に対応する分布間距離ベクトルの対を比較することで，状態間の (非) 類似性を求めることができる．そこで，3.2 節で述べた方法でアライメントされた状態系列対に基づき，それに対応する分布間距離ベクトル系列対に直すことで，状態系列間の新たなスコアの算出を行う．我々は，以下の 4 つの式を定義し，それぞれの式をもとにスコアの算出を試みる．

$$Score_{DDV\_L1} = \frac{\sum_{k=1}^{K}\sum_{s=1}^{S}|\psi_s(c_k)|}{K \cdot S} \quad (3)$$

$$Score_{DDV\_L2} = \frac{\sum_{k=1}^{K}\left\{\sum_{s=1}^{S}|\psi_s(c_k)|^2\right\}^{1/2}}{K \cdot S} \quad (4)$$

$$Score_{DDV\_L1max} = \frac{\max_{1\le k\le K}\sum_{s=1}^{S}|\psi_s(c_k)|}{K \cdot S} \quad (5)$$

$$Score_{DDV\_L2max} = \frac{\max_{1\le k\le K}\left\{\sum_{s=1}^{S}|\psi_s(c_k)|^2\right\}^{1/2}}{K \cdot S} \quad (6)$$

ここで，$\psi_s(c_k)$ はベクトル $\phi(a_i) - \phi(b_j)$ の s 番目の要素である．いずれのスコアも状態系列 A と B の類似性が高いほど，0 に近い値を取るため，このスコアを非類似度スコアとして利用することができる．$Score_{DDV\_L1}$ は対応する分布間距離ベクトル間の L1 ノルムを時系列上で累積したスコア付けであり，$Score_{DDV\_L2}$ は L2 ノルムを時系列上で累積したスコア付けである．一方，$Score_{DDV\_L1max}$ と $Score_{DDV\_L2max}$ は状態系列上で L1 ノルムまたは L2 ノルムの最大値をとるスコア付けであり，非類似箇所の存在を強調する狙いがある．

また，上記の式 (3)〜(6) では 3.2 節で述べた方法で得られる状態のアラインメントを用いるが，分布間距離ベクトル間の L1 ノルムや L2 ノルムを局所距離として DP マッチング法によって時系列上の最小累積距離としてスコアを求めることもできる．上記の式 (3),(4) に対応してそのように求めるスコアをそれぞれ $Score_{DDV\_DPL1}$, $Score_{DDV\_DPL2}$ と呼び，上記の 4 種類と併せて比較する．

上述の手順で分布間距離ベクトルを用いたスコア付けを行うと，その過程で 2 種類のスコアが算出される（図 6）．この 2 つのスコアの違いは，$Score_{BD}$ は比較する HMM の状態間の分布間距離に基づいて算出されたスコアであるのに対して，$Score_{DDV}$ は HMM の全音節全状態との分布間距離を考慮して算出されたスコアである．この 2 つのスコアを次式に従い結合させ，1 つのスコア $Score_{fusion}$ として用いる．

$$Score_{fusion} = \alpha \cdot Score_{BD} + (1-\alpha) \cdot \tau \cdot Score_{DDV} \quad (7)$$

ここで，α は $0 \le \alpha \le 1$ の重み付け係数であり，τ は 2 つのスコアのレンジを調整するための係数である．このように $Score_{BD}$ と $Score_{DDV}$ の結合スコアを用いる方法を **BD-DDV** 手法と呼ぶことにする．

4. サブワード n-gram 信頼度に基づく STD 手法とその併用システム

前節で述べた STD 手法では，サブワードや状態レベルでの非線形な伸縮を考慮したアラインメントに基づいて検出スコアを求める．関連研究では，認識誤りの問題に対処するためサブワード n-gram 単位に注目した照合を行うことの有効性も示されている [6], [7], [9]．我々も先行研究において，関連研究 [17] と同様にサブワード n-gram 単位の信頼度を用いた STD 手法との併用手法を提案し，有効性を示した [13]．本節ではサブワード n-gram 信頼度を用いた STD 手法と，前節で述べた照合方法との併用手法について述べる．

4.1 サブワード n-gram 信頼度に基づく STD 手法

検索語のサブワード列を $Q = \{w_1, \cdots, w_M\}$ とし，検索語の部分 n-gram を $\{w_i, \cdots, w_{i+n-1}\}$ $(i = 1, \cdots, M-n+1)$ とする．そして，検索語 n-gram に対して信頼度を用いた検出スコアを以下の式で定義する．

$$R_{n-gram} = \sum_{i=1}^{M-n+1} \sum_{\acute{\boldsymbol{W}} \in \boldsymbol{W}(\boldsymbol{X})} CM(\acute{\boldsymbol{W}}) C(\acute{\boldsymbol{W}}, \{w_i, \cdots, w_{i+n-1}\}) \quad (8)$$

ここで，$C(\acute{\boldsymbol{W}}, \{w_i, \cdots, w_{i+n-1}\})$ は文仮説 $\acute{\boldsymbol{W}}$ 中に検索語

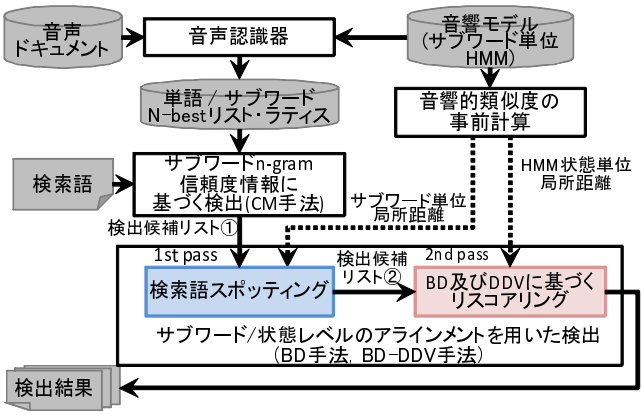

図 7 サブワード信頼度を用いた STD 手法を併用するシステム (サブワード/状態レベルのアラインメントを用いた検出の前段で併用する場合)

の部分 n-gram$\{w_i, \cdots, w_{i+n-1}\}$ が出現する数である．そして，次式のように 1-gram から N-gram まで重み a_n を与えて足し合わせ，最終的な検出スコアとして用いる．

$$Score_{CM} = \sum_{n=1}^{N} a_n R_{n-gram} \qquad (9)$$

重み a_n は，大きな n-gram ほど検出において重要であるため大きな値を割り当て，小さな n-gram に対しては湧き出し誤りを防ぐために小さな値を割り当てるように設定する．本研究では 5 節で述べる開発用データを用いて実験的に設定した．

なお，式 (8) 中では，検索語の部分 n-gram$\{w_i, \cdots, w_{i+n-1}\}$ が出現していない文仮説に対する信頼度はスコアに反映されないため，部分 n-gram が出現しているアーク (列) のみの信頼度が得られれば問題ない．そこで，forward-backward アルゴリズムによって文単位ではなく一部のアーク (列) の信頼度を推定することで効率的に計算を行う．このように，サブワード n-gram 信頼度を用いた方法を **CM** 手法と呼ぶことにする．

4.2 サブワード信頼度を用いた STD 手法を併用するシステム

3 節で述べた状態系列レベルのリスコアリング手法（BD-DDV 手法）と前述のサブワード信頼度に基づく検出（CM 手法）の併用による STD システムを図 7 に示す．後述の評価実験では，本節で述べたサブワード信頼度に基づく検出を，もう一方の組み合わせる検出手法の前段として行うか，後段として行うかで 2 通りの方法を考える．それぞれ，前段の検出法によって候補の絞り込みを行い，残った候補に対して次段の検出手法を適用する．このように，2 つの手法を段階的に適用し絞り込みを行うことで，検出精度を高める狙いがある．

表 1 CSJ コア講演 (開発用データ) の認識性能 [%]

書き起こし単位	W.Corr.	W.Acc.	S.Corr.	S.Acc.
単語ベース	76.7	71.9	86.5	83.0
音節ベース	-	-	81.8	77.4

5. テキストクエリによる STD の評価実験

5.1 実験条件

テキストクエリによる STD の評価実験では，開発用セットと評価用セットを用いることで提案手法の頑健性を検証する．開発用セットとして，STD のためのテストコレクション [2] の既知語クエリセット (50 個)・未知語クエリセット (50 個) を用いる．検索対象の音声ドキュメントは，日本語話し言葉コーパス (CSJ)[18] のコア講演データ (177 講演，約 44 時間) である．検索の評価は NTCIR SpokenDoc の CSJ データでの評価方法 [3], [4] と同様に，ポーズで分割された転記基本単位 (Inter Pausal Unit : IPU) での検出を正解判定の基本単位とする．そして，音声認識結果として，テストコレクションと共に配布された単語ベースと音節ベースの 2 種類のリファレンス認識結果 (各 10-best，ラティス) を用いる．リファレンスの認識性能を表 1 に示す．表 1 中の「W.Corr.」は単語正解率，「W.Acc.」は単語正解精度，「S.Corr.」は音節正解率，「S.Acc.」は音節正解精度を表わしている．

評価用セットは，NTCIR10 SpokenDoc-2 formal-run SDPWS(moderate-size) タスクのクエリセット (100 個)[4] を用いる．検索対象の音声ドキュメントは，音声ドキュメント処理ワークショップ (Spoken Document Processing Workshop : SDPWS) の講演音声 (104 講演，約 29 時間) である．音声認識結果として，NTCIR10 SpokenDoc-2 の際に配布された単語ベースと音節ベースの 2 種類のリファレンス認識結果 (各 10-best，ラティス) を用いる．認識性能を表 2 に示す．

分布間距離の算出の際に利用する音響モデルの学習には，モーラ単位の HMM を用いた．開発セットの際と同様に，音声認識に使用された音響モデルと同条件で学習した音響モデルを使用することで，オープンな評価条件とした．分布間距離の算出の際に利用する音響モデルの仕様を表 3 に示す．

本稿で述べる評価実験においては，音声認識用の音響モデルとは別に，分布間距離の算出用の音響モデルを使用するが，サブワードレベルの認識結果を得る過程では一種類の Triphone モデルしか使用しておらず (リファレンス認識結果のみ利用)，複数の音響・言語モデルやデコーダを用いる一般的なアプローチ [19] とは異なる．

検索性能の評価指標として，Recall，Precision，F-measure，Recall-Precision 曲線を用いる．

表 2 SDPWS 講演 (評価用データ) の認識性能 [%]

書き起こし単位	W.Corr.	W.Acc.	S.Corr.	S.Acc.
単語ベース	68.4	63.1	79.7	75.3
音節ベース	-	-	72.7	67.7

表 3 分布間距離の算出の際に利用する HMM の仕様

カテゴリ/単位	133 音節 (モーラ)
状態数	7 または 5
出力分布数	5 または 3
出力分布	32 混合の多次元正規分布 (対角共分散行列)
特徴パラメータ	38 次元 ($MFCC + \Delta MFCC + \Delta\Delta MFCC + \Delta Power + \Delta\Delta Power$)

5.2 比較する STD 手法

比較を行う STD 手法は以下の通りである．

Baseline(DP-SubwordBD): 2 節で述べた方法 (NTCIR9，10 でのベースライン手法とほぼ同様だがサブワード間距離の定義が異なる)

BD(DP-StateBD): 3 節で述べた方法の内，$Score_{BD}$ のみを用いた（式 (7) において $\alpha = 1$ とした）方法

BD-DDV: 3.3 節で述べた方法

CM: サブワード n-gram の信頼度情報に基づく検出手法 (4 節の方法単独)

BD-DDV+CM: BD-DDV の出力結果上位 K 個の検出候補に対して CM を適用

CM+BD-DDV: CM の出力結果上位 K 個の検出候補に対して BD-DDV を適用

5.3 開発用セットでの評価結果

3.3 節で分布間距離ベクトルに基づくスコア $Score_{DDV}$ として，6 種類のスコアを定義した．評価実験ではまず，BD-DDV における 6 種類のスコア性能比較を行った後，$Score_{DDV}$ が検索性能に与える影響を解析する．その後，各手法における性能評価を行う．

5.3.1 BD-DDV における $Score_{DDV}$ 算出の定義式の比較

各定義式による評価結果を表 4 に，未知語クエリの Recall-Precision 曲線を図 8 に示す．表 4 中の値は，閾値を変化させて F-measure が最大となったときの値である．また，図 8 は BD-DDV における 1 パス目の閾値と，2 パス目のスコア結合重み α を調整し，F-measure が最大となったときのパラメータを用いてプロットしている．評価結果から，既知語クエリに対しては $Score_{DDV\_L2}$ を用いた手法が他の定義式と比べて僅かであるが高い性能を示し，未知語クエリに対しては $Score_{DDV\_L2max}$ が高い性能を示した．そして，BD-DDV における 1 パスのみ (サブワードレベルの連続 DP マッチング) の手法である Baseline と比較して，DDV に基づくスコアリングを行うことで大きく性能を改善することができていることがわかる．

表 4 $Score_{DDV}$ 算出の各定義式による性能比較 (開発用セット：CSJ コア講演データ)[%]

		Recall	Precision	F-measure
Baseline	IV	61.05	89.88	72.71
	OOV	37.61	77.88	50.72
$Score_{BD}$ のみ ($\alpha = 1$)	IV	64.29	91.73	75.59
	OOV	46.15	78.26	58.07
$Score_{DDV\_L1}$	IV	64.15	92.61	75.80
	OOV	49.15	72.33	58.52
$Score_{DDV\_L2}$	IV	64.15	92.97	**75.92**
	OOV	48.72	75.00	59.07
$Score_{DDV\_DPL1}$	IV	64.42	91.57	75.63
	OOV	47.01	82.71	59.95
$Score_{DDV\_DPL2}$	IV	64.15	92.97	75.68
	OOV	47.86	81.16	60.22
$Score_{DDV\_L1max}$	IV	63.75	92.93	75.62
	OOV	48.72	82.01	61.13
$Score_{DDV\_L2max}$	IV	64.02	92.59	75.70
	OOV	50.00	81.25	**61.91**

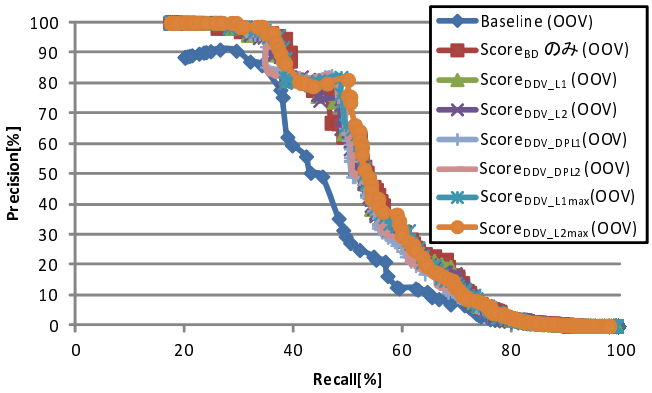

図 8 $Score_{DDV}$ 算出の各定義式による未知語クエリの Recall-Precision 曲線 (開発用セット：CSJ コア講演データ)

また，$Score_{DDV\_DPL1}$ や $Score_{DDV\_DPL2}$ については，DP-StateBD の段階（3.2 節）で得られた状態アラインメントを用いる $Score_{DDV\_L1}$ や $Score_{DDV\_L2}$ と近い性能が得られている．従って，結合スコアを用いる場合に 2 段階に分けずに両者を併せた局所距離でアラインメントをとっても（または，$Score_{DDV\_L1max}$, $Score_{DDV\_L2max}$ で事前のアラインメントを用いない方法でも）同様な結果が得られるものと予想される．

以降の BD-DDV 手法では，既知語検索語・未知語検索語ともに $Score_{DDV}$ の算出式を $Score_{DDV\_L2max}$ に固定して評価を行う．

5.3.2 BD-DDV における $Score_{DDV}$ の影響

BD-DDV では，式 (7) で表わされるように HMM の状態単位の DP マッチングスコア $Score_{BD}$ と，分布間距離ベクトルに基づくスコア $Score_{DDV}$ をスコア結合係数 α により結合し，最終的な検出スコア (非類似度スコア) としている．そこで，このスコア結合係数 α を変化させた際の検索

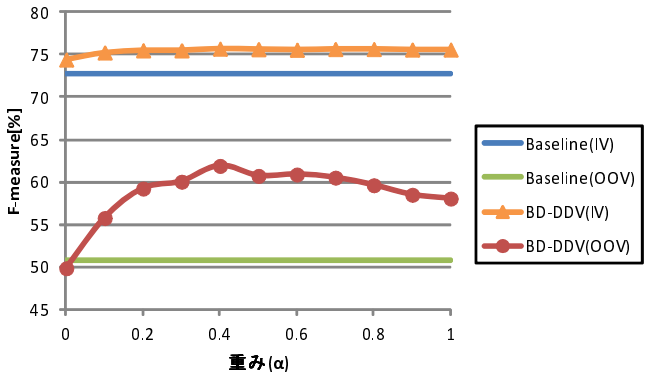

図 9 スコア結合重み α を変化させた際の F 値の推移 (開発用セット：CSJ コア講演データ)

性能の影響を解析する．スコア結合重み α を 0 から 1 まで変化させた際の，F-measure の推移を図 9 に示す．

図 9 より，既知語クエリ (IV) においては重みの影響が小さいことがわかる．しかし，未知語クエリ (OOV) においては 0.4 あたりを頂点とする山なりになっており，Baseline と比較して大きく改善されている．これは，分布間距離ベクトルが多くの相対的な距離情報を持つため，誤りを含むサブワード対の非類似性を評価する特徴量としてうまく働き，検出性能が向上したためと思われる．なお，$\alpha = 1.0$ の場合においても Baseline よりも改善していることから，サブワード系列を状態系列に展開した後でのアライメントは有効であることがわかる．

5.3.3 各手法における性能評価

5.2 節で挙げた比較する STD 手法における性能比較結果を表 5 に，未知語クエリに対しての検索性能のみを評価した際の Recall-Precision 曲線を図 10 に示す．表 5 より，CM は Baseline よりも優れた性能を示し，未知語クエリのみに対しても高い性能を示していることが図 10 からも読み取れる．この開発用セットは，評価セットと比べてリファレンスの音声認識性能が高いため，信頼度情報に基づく検出を行う CM が良い性能を示したと思われる．

併用手法である CM+BD-DDV は既知語クエリ・未知語クエリ共に最も良い性能を示している．単体の手法と比較しても，特に未知語クエリにおいて大きな改善が見られる．CM は低い Recall において高い Precision を示している．つまり，検出結果の上位が正解である割合が高いといえる．そして，CM+BD-DDV ではその検出結果の上位候補に対して更に BD-DDV でスコア付けすることで，より優れた性能を示すことができた．

5.4 評価用セットでの評価結果

開発用セットで調整を行ったパラメータを用いて評価用セットで評価実験を行った結果を表 6 に，未知語クエリに対しての検索性能のみを評価した際の Recall-Precision 曲線を図 11 に示す．

表 5 各手法における性能評価 (開発用セット：CSJ コア講演データ)[%]

		Recall	Precision	F-measure
Baseline	IV	61.05	89.88	72.71
	OOV	37.61	77.88	50.72
BD(DP-StateBD) ($\alpha = 1$)	IV	64.29	91.73	75.59
	OOV	46.15	78.26	58.07
CM	IV	62.80	93.76	75.22
	OOV	49.57	87.88	63.39
BD-DDV	IV	64.02	92.59	75.70
	OOV	50.00	81.25	61.91
BD-DDV+CM	IV	61.86	96.03	75.25
	OOV	47.86	96.55	64.00
CM+BD-DDV	IV	66.58	90.81	**76.83**
	OOV	62.39	84.88	**71.92**

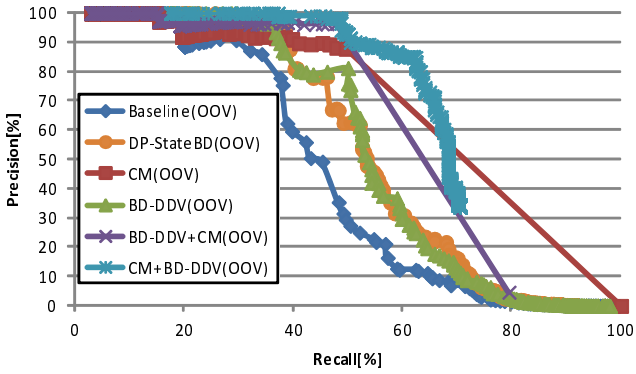

図 10 各システムにおける未知語クエリの Recall-Precision 曲線 (開発用セット：CSJ コア講演データ)

これらの評価結果から，開発用セットの場合とは異なり Baseline や CM よりも BD-DDV が優れた性能を示した．これは，開発用セットよりも評価用セットのリファレンスの音声認識性能が悪いため，調整パラメータが開発セットと評価用セットの違いに影響を受け易くなり，CM の性能が大きく低下したと思われる．この影響は，後段に CM を適用する併用手法である BD-DDV+CM においても現れている．一方，BD-DDV は CM と比較して音声認識性能の低下による影響が小さく，Baseline からの改善が大きいことがわかる．

図 11 から，CM は未知語クエリに対して Recall を延ばした際の Precision の低下が大きいことがわかる．しかし，開発用セットで示された検出結果の上位が正解である割合が高いという性質は，評価用セットにおいても現れていることが読み取れる．そして，併用手法である CM+BD-DDV は，評価用セットに対しても単体の手法よりも良い性能を示した．つまり，音声認識性能が低い条件においても提案手法は有効であるといえる．

さらに，スコア結合重み α を 0 から 1 まで変化させた際

表 6 各手法における性能評価 (評価用セット：SDPWS 講演データ)[%]

		Recall	Precision	F-measure
Baseline	IV	46.95	45.92	46.43
	OOV	14.60	28.63	19.34
BD(DP-StateBD) ($\alpha = 1$)	IV	49.21	58.45	53.43
	OOV	23.31	20.98	22.09
CM	IV	40.41	77.49	53.12
	OOV	19.61	22.96	21.15
BD-DDV	IV	48.76	59.02	53.40
	OOV	18.95	40.47	25.82
BD-DDV+CM	IV	46.95	75.91	**58.02**
	OOV	20.04	36.80	25.95
CM+BD-DDV	IV	49.89	68.85	57.85
	OOV	24.62	47.28	**32.38**

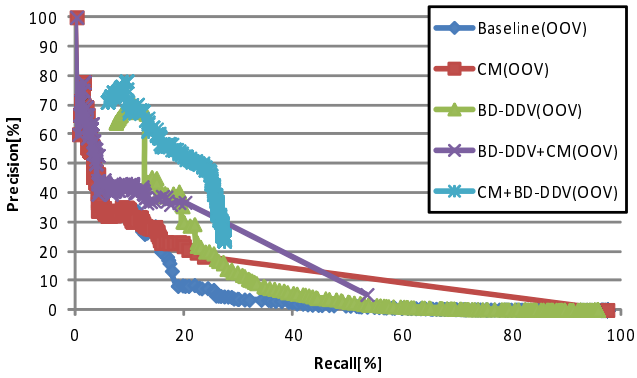

図 11 各システムにおける未知語クエリの Recall-Precision 曲線 (評価用セット：SDPWS 講演データ)

の，BD-DDV 手法に対する F-measure の推移を図 12 に示す．この結果をみると，未知語クエリにおいて $Score_{DDV}$ 単独（$\alpha = 0$）の場合の方が $Score_{BD}$ 単独（$\alpha = 1$）の場合よりも高い性能を得ている．このことは，開発用セットでの傾向と逆であるが，これも検索対象の音声認識精度が低い場合に $Score_{DDV}$ の方が影響を受けにくいことを示しているものと考えられる．

6. 音声クエリによる STD の評価実験

6.1 実験条件

音声クエリによる STD の評価実験では，開発用セットを用いず，評価用セットで性能が最大となったパラメータを利用して評価を行った．評価用セットは，NTCIR9 SpokenDoc Dry-run CORE タスクの未知語クエリセット（50 個）の内，CSJ のコア講演以外の講演で発話されているもの（21 個）を用いた．CSJ のコア講演以外の講演音声から，その発話部分を切り出したものを音声クエリとしている．検索対象の音声ドキュメントは，CSJ のコア講演データである．音声認識結果として，NTCIR9 SpokenDoc の際に配布された音節ベースのリファレンス認識結果 (10-best)

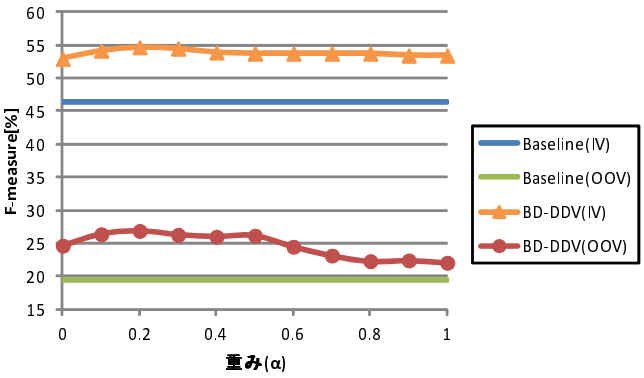

図 12 スコア結合重み α を変化させた際の F 値の推移 (評価用セット：SDPWS 講演データ)

表 7 音声クエリの音節認識性能比較 [%]

	音節正解率	音節正解精度
NTCIR9 言語モデル	60.8	58.4
名詞句言語モデル	65.6	63.2

を用いる．

検索性能の評価指標として，Recall，Precision，F-measure，Recall-Precision 曲線を用いる．

6.2 音声クエリの認識性能

本研究では，3 節で述べたように，音声クエリを音節認識することで音節列に変換している．その際に，一般的にクエリには名詞句が多いことを踏まえて，名詞句のみで学習した音節 trigram 言語モデル (名詞句言語モデル) を用いることを検討した．名詞句言語モデルの学習データには CSJ のコア講演以外の書き起こしの名詞句を用いた．学習条件は，文献 [2] で述べられているリファレンスの言語モデルの作成手順に従い，各講演音声に付与された ID が奇数か偶数かによって 2 分割し，それぞれで学習を行った．音節認識時には，切り出し元の ID が奇数のクエリには偶数で学習した言語モデル，偶数のクエリには奇数で学習した言語モデルを使用する．

音声クエリによる評価実験ではまず，音声クエリの認識時に検索対象の音節認識に使われている NTCIR9 SpokenDoc の際に配布された音節 trigram 言語モデル (NTCIR9 言語モデル) を用いた場合と名詞句言語モデルを用いた場合での音節認識性能の比較を行った．その結果を表 7 に示す．表 7 より，名詞句モデルを用いるほうが音声クエリの認識性能が高いことがわかる．よって，以降の実験では音声クエリの認識に名詞句モデルを用いた．

6.3 比較する STD 手法

比較を行う STD 手法は以下の通りである．

Baseline(DP-SubwordBD): 2 節で述べた方法 (NTCIR9，10 でのベースライン手法とほぼ同様だがサブ

表 8 各手法の音声クエリにおける検索性能 (CSJ コア講演データ)[%]

		Recall	Precision	F-measure
音声クエリ	Baseline	20.83	45.46	28.57
	BD(DP-StateBD)	39.17	29.19	33.45
	BD-DDV	34.17	50.00	40.59
テキストクエリ	Baseline	38.33	77.97	51.40
	BD(DP-StateBD)	39.17	97.92	55.95
	BD-DDV	45.83	85.94	59.78

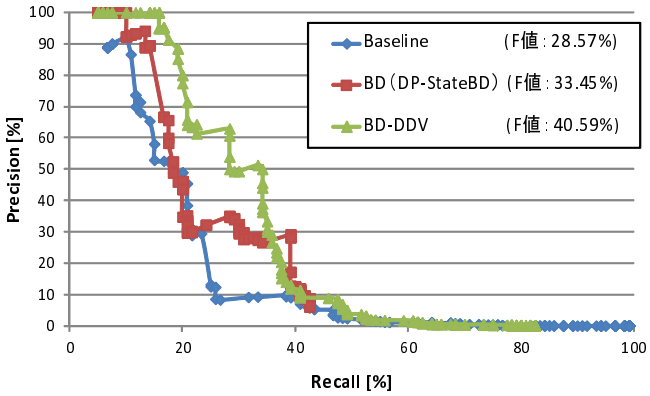

図 13 Recall-Precision 曲線 (音声クエリ，CSJ コア講演データ)

ワード間距離の定義が異なる)

BD(DP-StateBD): 3 節で述べた方法の内，$Score_{BD}$ のみを用いた（式 (7) において $\alpha = 1$ とした）方法

BD-DDV: 3.3 節で述べた方法

BD-DDV における $Score_{DDV}$ 算出の定義式に関しては，テキストクエリにおいて高い性能を示した $Score_{DDV\_L2max}$ を用いた．また，各手法に関して，音声クエリを正確な音節列に直したもの（テキストクエリ）に対しても実験を行った．

6.4 評価結果

閾値などのパラメータを変化させて実験を行い，各手法で F-measure が最大となったパラメータでの結果を表 8 に，Recall-Precision 曲線を図 13,14 に示す．

これらの評価結果から，音声クエリによる STD においても HMM 状態レベルでのリスコアリング手法を用いることで検索性能を改善している．そして，BD-DDV 手法は BD(DP-StateBD) 手法と比べても，Recall が 10%を超える辺りからの Precision を更に改善できていることがわかる．このことから，音声クエリによる STD においても BD-DDV 手法が有効であると考えられる．

7. おわりに

本稿では，テキストクエリによる STD において，先行研究で提案した分布間距離ベクトルという構造的特徴表現

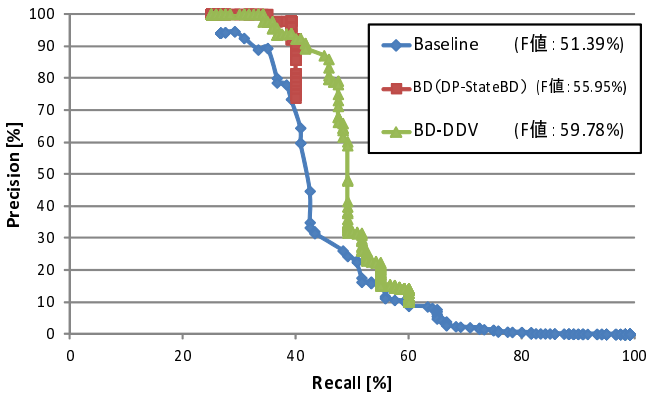

図 14 Recall-Precision 曲線 (テキストクエリ, CSJ コア講演データ)

に基づく検出手法と，信頼度情報に基づく検出手法を併用することで性能の改善を行った．また，音声クエリによる STD に対する，分布間距離ベクトルという構造的特徴表現に基づく検出手法の有効性を検証した．

今後の課題として，検索の高速化やシステムパラメータの自動推定が挙げられる．前者については，これまでに音節ラティスの n-gram を検索用インデックスとする等による高速化手法が提案されており [9]，このようなインデキシングによる高速化手法との併用は容易に実現できると考えている．

参考文献

[1] 滝上, 他 : “音声検索語検出を前処理に用いた未知語や認識誤りに頑健な音声ドキュメント検索,” 情報処理学会論文誌, Vol.54, No.2, pp.1-12, (2013).

[2] 西崎博光, 他 : “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, Vol.2010-SLP-81, No.13, (2010).

[3] Tomoyosi Akiba, et al. : “Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop,” Proc. of 9th NTCIR Workshop Meeting, pp.223-235 ,(2011.12.6-9).

[4] Tomoyosi Akiba, et al. : “Overview of the NTCIR-10 SpokenDoc-2 Task,” Proc. of the 10th NTCIR Workshop Meeting, (2013).

[5] 秋葉友良 : “音声ドキュメント検索の現状と課題,” 情報処理学会研究報告, Vol.2010-SLP-82, No.10, (2010).

[6] 石見, 他 : “音声ドキュメント検索のための音節ラティスの拡張と n-gram 索引の削減手法,” 情報処理学会研究報告, Vol.2011-SLP-89, No.5, (2011).

[7] 勝浦, 他 : “複数認識結果を用いて構築した Suffix Array に対する音声検索語検出,” 情報処理学会研究報告, Vol.2012-SLP-94, No.15, pp.1-6, (2012).

[8] 岩見圭介, 他 : “距離付き n-gram インデックスによる認識誤りと未知語に頑健な高速検索法,” 情報処理学会研究報告, Vol.2010-SLP-83, No.3, (2010).

[9] 中川聖一, 他 : “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,” 音声ドキュメント処理ワークショップ講演論文集 3, pp.7-14, (2009.2.27).

[10] M.Wechsler, et al. : “New Techniques for Open-Vocabulary Spoken Document Retrieval,” Proc. of the 21st Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, pp.20-27, (1998).

[11] 山本直樹 , 甲斐充彦 : “分布間距離ベクトルを特徴量とする距離尺度を用いた音声検索語検出の検討,” 第 7 回音声ドキュメント処理ワークショップ講演論文集, No.4, (2013.3).
[12] Naoki Yamamoto, Atsuhiko Kai : “Using Acoustic Dissimilarity Measures Based on State-level Distance Vector Representation for Improved Spoken Term Detection,” Proc. of APSIPA ASC 2013, (2013.10).
[13] 山本直樹 , 甲斐充彦 : “分布間距離に基づく音響的類似度とサブワード事後確率の併用による音声検索語検出の改善,” 情報処理学会研究報告, Vol.2013-SLP-99, No.1, (2013.12).
[14] 朝川 智 , 峯松信明 , 広瀬啓吉 : “音声の構造的表象に基づく英語学習者発音の音響的分析,” 電子情報通信学会論文誌, Vol.J90-D, No.5, pp.1249-1262, (2007).
[15] 村上隆夫 , 峯松信明 , 広瀬啓吉 : “音声の構造的表象に基づく日本語孤立母音系列を対象とした音声認識,” 電子情報通信学会論文誌. Vol.J91-A, No.2, pp.181-191, (2008).
[16] A.Muscariello, et al. : “Zero-resource audio-only spoken term detection based on a combination of template matching techniques,” Proc. of INTERSPEECH 2011, pp.921-924, (2011).
[17] Lee. H, et al.: “Open-Vocabulary Retrieval of Spoken Content with Shorter/Longer Queries Considering Word/Subword-based Acoustic Feature Similarity,” Proc. of Interspeech, (2012).
[18] 国立国語研究所 : “日本語話し言葉コーパス”, http://www.kokken.go.jp/katsudo/seika/corpus/, (2004).
[19] Hiromitsu Nishizaki, et al. :“Spoken Term Detection Using Multiple Speech Recognizers’ Outputs at NTCIR-9 SpokenDoc STD subtask,” Proc. of NTCIR-9 Workshop Meeting, pp.236-241, (2011.12.6-9).