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手法の検討

山本 直樹[1,a)]　甲斐 充彦[1,b)]

概要：近年, 音声や動画などのマルチメディアコンテンツの配信や利用が拡大しており，高い精度のマルチメディア検索技術が望まれている．音声ドキュメント検索に関しては，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 Term Detection : 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．本稿では, このようなサブワード列に基づく検索語との類似性の評価において，音響モデルの状態単位での構造的特徴を表す分布間距離ベクトルを用いて音響的な距離尺度を求め，検索語検出に利用する方法を提案する．評価実験では，NTCIR9 SpokeDoc CORE タスクにおいてサブワード単位の音響的距離尺度のみを用いる従来法と比較を行い，従来法と同じ音節単位 HMM を利用した場合においても新たな距離尺度および照合手法を導入することで，検索時間をほとんど増やさず検索性能を顕著に改善できることが示された.

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

Investigation of Spoken Term Detection Method Using an Acoustic Dissimilarity Measure Based on a Vector Representation of Distances between Two Distributions

Naoki Yamamoto[1,a)]　Atsuhiko Kai[1,b)]

Abstract: In recent years, demands for distributing or searching multimedia contents are rapidly increasing and more effective method for multimedia information retrieval is desirable. In the studies on spoken document retrieval systems, much research has been presented focusing on the task of spoken term detection (STD), which locates a given search term in a large set of spoken documents. One of the most popular approaches performs indexing based on the sequence of sub-word units which is converted from the recognition hypotheses from LVCSR decoder for considering recognition errors and OOV problems. In this paper, we propose a method for evaluating acoustic dissimilarity between two sub-word sequences based on a sequence of vector representation, which consists of all the distances between two possible combinations of distributions in a set of sub-word unit HMMs and represents a structural feature. The experimental result showed that the proposed method significantly outperforms a baseline me, thod which employs only acoustic dissimilarity measure between two sub-word units, by incorporating new acoustic dissimilarity measure and a verification method which relies on a common set of sub-word HMMs.

Keywords: spoken term detection, distance between two distributions, distance measure between two structures, acoustic similarity

[1] 静岡大学
Shizuoka University
[a)] yamamoto_nao@spa.sys.eng.shizuoka.ac.jp
[b)] kai@sys.eng.shizuoka.ac.jp

1. はじめに

近年，高速なインターネット環境や，パソコン，HDD ビ

デオレコーダなどの普及により，音声や動画などのマルチメディアコンテンツの増加・大容量化が進んでいる．また，スマートフォンの急速な普及により，このようなマルチメディアコンテンツに手軽にアクセスできる環境が広まっている．これに伴い，音声を含む大量のコンテンツデータから目的の情報を検索する技術が求められている．例えば，新聞記事や Web ページのようなテキストデータに対しては，既存のテキスト検索手法により，高度な検索を行うことができる．しかし, 音声ドキュメントに関しては話者や収録環境，発話様式の違いなど様々な変動を含むデータを対象とするため，安定した検索性能を実現する方法は十分に確立されていない．

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．また，類似したタスクとして，検索語が含まれる，または検索語に関連する音声ドキュメントを検索する，音声ドキュメント検索 (Spoken Document Retrieval : SDR) と呼ばれるタスクも存在するが，ここでも，STD の技術が応用される例がある [1]．2006 年には米国の NIST を中心に STD タスクが設定され，研究が盛んにおこなれている．日本では 2008 年からテストコレクションの構築が開始され，2010 年 5 月にはそのベースライン評価について報告が行われている [2]．また，2011 年に国立情報学研究所により開催された NTCIR-9 Workshop では，STD タスクが設定され，評価が行われた [3].

STD タスクへの最も単純なアプローチとして，音声データをあらかじめ大語彙音声認識システムによってテキスト化し, それに対してテキスト検索を行うという方法が挙げられる．しかし，大語彙音声認識システムの辞書に存在しない未知語を含む音声を認識した場合や既知語を誤認識した場合に正確にテキスト化がされないため, 検索語の検出性能は低下してしまう．そこで，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], [5], [6], [8], [10], [14] や，SDR タスクの検索手法 [4] 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, 認識誤りを許容した類似箇所を検出するもっとも基本的なアプローチとして, サブワードレベルでの連続 DP マッチングによるスポッティングが利用される．これにより，音声ドキュメントから自動抽出されたサブワード列の認識候補から, 検索語として入力されたサブワード列のパターンが含まれる可能性が高い箇所を抽出することができる．

我々は，サブワード間の非類似度尺度を前提とする最も単純な距離尺度に加えて，サブワード単位の音節モデル (HMM) から計算される状態単位の分布間距離を特徴量とする距離尺度を導入し，検索語と音声ドキュメントとの照合を行う方法を提案する．なお，提案する手法は従来の連続 DP マッチングに基づく音声検索語検出の手法に対して計算量を増加させないために，2 パスの照合手法を採用する．同様に多段的な照合を用いるアプローチは，一般に徐々により精密なモデルを適用することで，検索時間の増加を伴うことなく検索精度を向上させる方法として用いられる [8]．一方，本稿で述べる評価実験においては，一種類のサブワード単位音響モデル（音節 HMM）を，従来と同様なサブワード列の連続 DP マッチングに用いるサブワード対の音響的な非類似度の算出と，分布間距離ベクトルに基づくサブワード列の対の非類似度の算出の両方に利用する．すなわち，複数の音響・言語モデルやデコーダを用いるアプローチ [9] とは異なる．また，従来の DP マッチングに基づく方法と同様に音響モデルから距離テーブルをあらかじめ計算して用意することができるため，2 パスを通しての計算量の増加は極めて少なく実現できる．

2. 関連研究

サブワード単位の音響的な類似度を用いる関連研究としては，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 [5] や，音素弁別特徴に基づく距離尺度を利用する例 [6], 等がある．一般にサブワード単位の HMM は複数の状態から成りそれぞれの状態が出力分布を持つため, 分布間距離としての Bhattacharyya 距離は任意の状態間で定義できる. そのため，状態ペア毎に求まる分布間距離を集計してサブワード単位での非類似度を定義して用いることが多い．本稿で提案する方法は，あらかじめサブワード単位での非類似度を求める代わりに，任意の状態間で定義される分布間距離を要素として構成される分布間距離ベクトルに基づいて任意の音節・状態ペア毎の距離を定義し，サブワード列全体での非類似度を求める．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 [11], [12] 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，音声ドキュメント検索における関連研究として，Muscariello ら [7] は音声入力された検索語による音声ドキュメント中の類似部分を検出する方法として，音声セグメントを GMM や HMM の状態レベルの postriorgram 系列として表現し，セグメント内のあるフレームと他の任意のフレームの対での自己類似性行列として新たに表現された音声セグメント対（検索語と音声ドキュメント中の候補区間）の類似性を評価する方法を提案し，GMM 等の学習データと異なる言語（事前の音声言語資源の利用をほとんど仮定しない言語への適用の想定）に対する頑健性を示している．

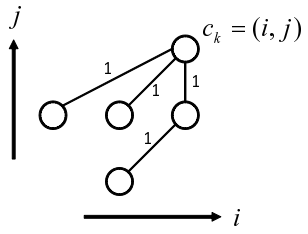

図 2 非対称 DP パス

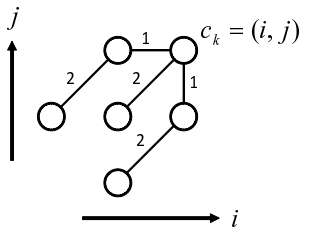

図 3 対称 DP パス

このように，本稿での提案手法はサブワード単位 HMM を用いて音響的な非類似度をあらかじめ計算する方法の一種であるが，誤りを含むサブワード列との照合においてより頑健な照合を実現することを意図している．

3. 連続 DP マッチングを用いた従来手法

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，認識誤りを考慮した検索ご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，まず，連続 DP マッチングを利用した STD システムの概要を述べる．そして，連続 DP マッチングの局所距離として使用される，分布間距離について述べる．

3.1 STD システム概要

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大語彙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おき蓄積する．

検索を行う際は，入力された検索語をサブワード系列に変換し，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．

3.2 サブワード対の音響的な非類似度

サブワード間の音響的な類似度を考慮するために，分布間距離を算出し，その距離を音響的な非類似度として使用する．音響モデルであるサブワード単位の HMM には一般に複数の状態が含まれ，それぞれの状態に出力分布が与えられる．任意の状態間に対して何らかの分布間距離を定義することができるが，分布間距離が近い状態対は，音響的に類似していると言うことができ，また逆に距離が遠い状態対は，音響的に類似していないと言える．

3.2.1 分布間距離の算出

分布間距離は，HMM のパラメータから Bhattacharyya 距離を利用して算出する．Bhattacharyya 距離は，確率分布の間の距離を計算する際に用いられ，2 つの確率分布が多次元正規分布であるとき次式で示される．

$$BD(P_a^{\{i\}}, P_b^{\{j\}}) = \frac{1}{8}(\mu_a^{\{i\}} - \mu_b^{\{j\}}) \left(\frac{\sum_a^{\{i\}} + \sum_b^{\{j\}}}{2} \right)^{-1} (\mu_a^{\{i\}} - \mu_b^{\{j\}})^t + \frac{1}{2} \log \left(\frac{|(\sum_a^{\{i\}} + \sum_b^{\{j\}})/2|}{|\sum_a^{\{i\}}|^{1/2} |\sum_b^{\{j\}}|^{1/2}} \right)$$

ここで，$P_a^{\{i\}}$ は音節 a の HMM の状態 i における出力分布 $N(\mu_a^{\{i\}}, \sum_a^{\{i\}})$，$\mu_a^{\{i\}}$ は音節 a の HMM の状態 i における平均ベクトル，$\sum_a^{\{i\}}$ は音節 a の HMM の状態 i における共分散行列である．

上記の Bhattacharyya 距離は単一正規分布の場合であるが，一般的には出力分布が混合正規分布 (GMM) の場合が多い．そこで，我々はある 2 つの混合分布間の距離として，任意の混合成分間の Bhattacharyya 距離の最小値を分布間距離とする．つまり，ある音節 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，音節 a の HMM 状態 i と音節 b の HMM の状態 j との距離を次式で定義する．

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

この $BD(P_a^{\{i\}}, P_b^{\{j\}})$ を分布間距離として使用する．

サブワードの状態系列間に対して，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 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．このマッチング距離を新たな局所距離として連続 DP マッチングで使用することにより，音響的類似度を考慮したサブワード単位のスポッティングを行う．なお，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 [8] も挙げられる．これを用いた場合，後に述べる評価実験においても良い性能を示したが，本稿では一貫して，分布間距離をもとに求めたサブワード間のマッチング距離を，サブワード単位のスポッティングの際の局所距離として利用する．

3.2.2 分布間距離の算出の際に利用する音響モデルの仕様

分布間の距離を算出する際に利用する音響モデルの仕様を，表 1 に示す．この音響モデルはモーラ単位の HMM であり，基本的には 7 状態 5 出力分布であるが，母音の/a/, /i/, /u/, /e/, /o/や, 無音の/N/, /q/, /sp/, /silB/, /silE/ は 5 状態 3 出力分布となっている．

4. 提案手法

連続 DP マッチングによるスポッティングにより，与え

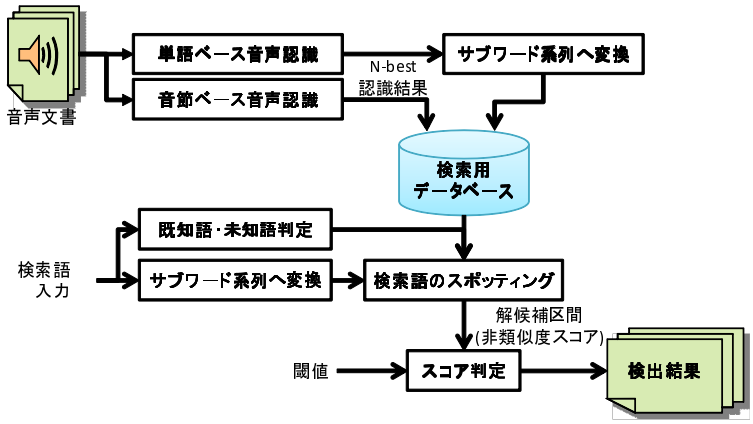

図 1　音声検索語検出のベースラインシステムの構造

表 1　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 ($MFCC + \Delta MFCC + \Delta\Delta MFCC$ $+\Delta Power + \Delta\Delta Power$)

られたクエリとしてのサブワード列に対し，検索対象の音声ドキュメント中のサブワード列としての類似性が高い候補区間を抽出することができる．しかし，サブワード単位の荒いスコア付けであり，音響的な非類似度の算出に用いた音響モデルによる非類似度の推定精度に左右される．そこで，この候補区間に対して，さらに詳細なスコア付け (リスコアリング) を行う．我々の提案手法では，候補区間のサブワード列を，それに対応する HMM の状態系列に拡張した時系列の表現に変換し，状態間の分布間距離に基づく距離尺度を用いる．特に，3.2 節で定義した分布間距離をもとに拡張した分布間距離ベクトルに基づく距離尺度を導入することで，検索性能の向上を図る．

4.1　分布間距離ベクトル

前節で定義した分布間距離 $BD(P_a^{\{i\}}, P_b^{\{j\}})$ は，HMM のある状態 a_i，b_j 間の 1 対 1 の距離の情報のみを表していた．我々は，ある音節の状態に対して 1 対 1 の距離を直接評価するだけではなく，他の状態との距離にも着目し，それぞれの距離を要素として含む特徴表現を利用することを考える．これにより，サブワード単位の認識誤りに対してより頑健な距離尺度とすることを意図している．

あるサブワード列に対応するサブワード単位の HMM の連結から成る状態系列を，

$$C = c_1, c_2, \cdots, c_m, \cdots, c_M \qquad (2)$$

と表わす．また，HMM の全サブワード単位の全状態から成る系列を

$$S = s_1, s_2, \cdots, s_l, \cdots, s_L$$

と表わす．ここで，M，L はそれぞれの HMM の状態系列の長さである．このとき，前節で定義した分布間距離を用いて次のようなベクトルを定義する．

$$\boldsymbol{\phi(c_m)} = (BD(s_1, c_m), BD(s_2, c_m), \cdots, BD(s_L, c_m))^{\mathrm{T}}$$

このベクトルは，HMM のある状態 c_m と全てのサブワード HMM 中の任意状態 s_l との距離を要素に持っていることから，**分布間距離ベクトル**と呼ぶ．この分布間距離ベクトルは，ある状態の相対的な距離特徴を表すことができるため，特徴量ベクトルとして用いることができる．そして，この分布間距離ベクトルを用いて，式 (2) で与えられた状態系列に対応して以下のようなベクトル列 (行列) を構成できる．

$$\Phi(C) = (\boldsymbol{\phi(c_1)}, \boldsymbol{\phi(c_2)}, \cdots, \boldsymbol{\phi(c_M)})$$
$$= \begin{bmatrix} BD(s_1,c_1) & \cdots & BD(s_1,c_m) & \cdots & BD(s_1,c_M) \\ \vdots & \ddots & & & \vdots \\ BD(s_l,c_1) & \cdots & BD(s_l,c_m) & \cdots & \vdots \\ \vdots & & & \ddots & \vdots \\ BD(s_L,c_1) & \cdots & \cdots & \cdots & BD(s_L,c_M) \end{bmatrix}$$

この行列は，各要素に分布間距離を含んでいることから，これ以降分布間距離行列 (Distribution Distance Matrix : DDM) と呼ぶことにする．この行列は C の状態系列に対応して分布間距離ベクトルを並べたものである．HMM の状態系列 C の分布間距離行列 $\Phi(C)$ の概念図を図 4 に示す．

4.2　分布間距離ベクトルを用いたスコア付け

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 状態の対に関して分布間距離ベクトル間を比較することで，状態間の (非) 類似性を求めることができる．そして，状態系列間の類似性を求める場合は，分布間距離行列の比較を行うことで，HMM の状態系列間の (非) 類似性を求めることができる．

比較を行う 2 つのサブワード列に対応する HMM の状態系列をそれぞれ，

$$A = a_1, a_2, \cdots, a_i, \cdots, a_I$$

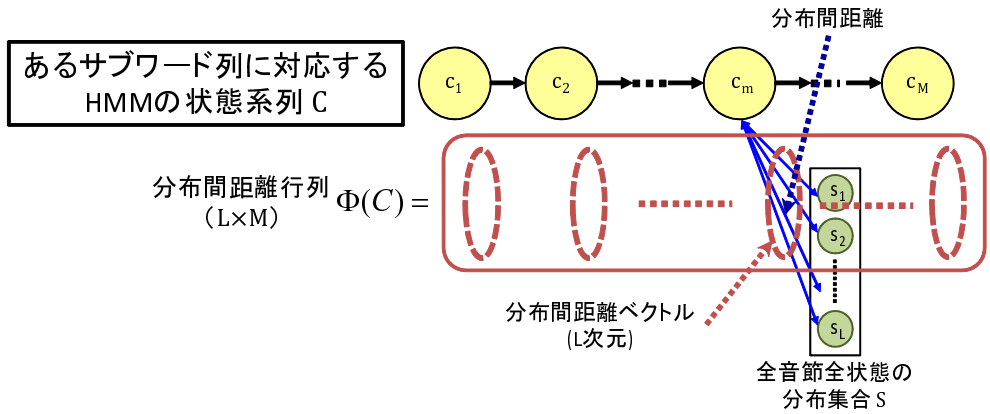

図 4　あるサブワード列に対して導出される分布間距離ベクトルと分布間距離行列の概念図

$$B = b_1, b_2, \cdots, b_j, \cdots, b_J$$

と表す．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正規化を行う．このときの局所距離としては式 (1) の Bhattacharyya 距離を用いる．DP マッチングにより得られた最適経路 F^* を

$$F^* = c_1, c_2, \cdots, c_k, \cdots, c_K$$

$$c_k = (a_{i_k}, b_{j_k})$$

と表し，この最適経路 F^* に沿って伸縮させた状態系列を A', B' とする．これにより，状態系列対が同じ長さ K に正規化される．この際に得られる DP マッチングスコアを $Score_{DP}$ とする．

正規化を行った状態系列対 A', B' をもとに，それぞれの分布間距離行列を求め，状態系列間のスコアを算出する．我々は，以下の 3 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．

$$Score_{DDM\_L1Max} = \frac{\max_{1 \le j \le K} \sum_{i=1}^{L} |\psi_{ij}|}{K \cdot L} \quad (3)$$

$$Score_{DDM\_L2} = \frac{1}{K} \sum_{j=1}^{K} \left\{ \frac{1}{L} \sum_{i=1}^{L} |\psi_{ij}|^2 \right\}^{1/2} \quad (4)$$

$$Score_{DDM\_L1} = \frac{\sum_{j=1}^{K} \sum_{i=1}^{L} |\psi_{ij}|}{K \cdot L} \quad (5)$$

ここで ψ_{ij} は $\Phi(A') - \Phi(B')$ の i 行 j 列の要素である．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．$Score_{DDM\_L2}$ は対応する分布間距離ベクトル間の L2 ノルムを時系列上で累積したスコア付けであり，$Score_{DDM\_L1}$ は L1 ノルムを時系列上で累積したスコア付けである．一方，$Score_{DDM\_L1Max}$ は状態系列上で L1 ノルムの最大値をとるスコア付けであり，非類似性を強調する狙いがある．

それぞれのスコア算出法による性能比較は 5.4 節で行う．

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．1 つ目は，HMM の状態系列間の長さを正規化する際の DP マッチングのマッチングスコア $Score_{DP}$ である．2 つ目は，状態系列長の正規化後に分布間距離行列を比較することにより得られるスコア $Score_{DDM}$ である．この 2 つのスコアの違いは，$Score_{DP}$ 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，$Score_{DDM}$ 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．また，結合スコアの算出過程の概念図を図 5 に示す．

$$Score_{fusion} = \alpha \cdot Score_{DP} + (1 - \alpha) \cdot \tau \cdot Score_{DDM} \quad (6)$$

ここで，α は，$0 \le \alpha \le 1$ の重み付け係数であり，1 に近づけるほど HMM の状態単位の DP マッチングスコア $Score_{DP}$ を重視し，0 に近づけるほど分布間距離ベクトルを用いたスコア $Score_{DDM}$ を重視する．また，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．

4.3　提案手法を用いた STD システムの概要

我々の提案する STD システムを図 6 に示す．提案システムは 2 パス手法になっており，1 パス目で区間検出と粗い絞り込みを行い，2 パス目で詳細なスコア付けを行う．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) 抽出された区間とキーワードそれぞれのサブワード列

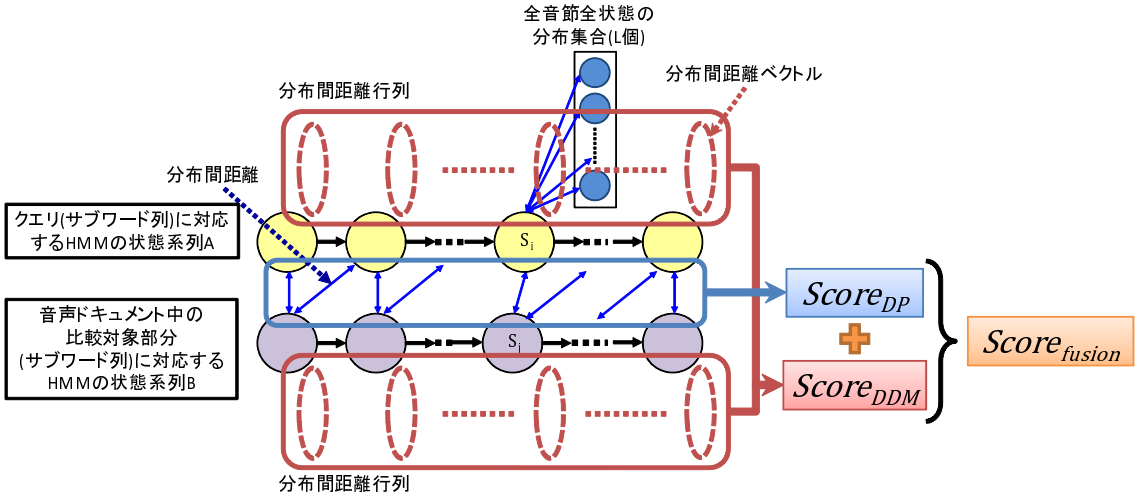

図 5　結合スコアの算出過程の概念図

に対応する音響モデル (HMM) の状態系列に対して，各状態間の分布距離を局所距離として DP マッチングを行い，スコア $Score_{DP}$ を求める．

(4) 手順 3 の DP マッチングによって得られた最適経路に沿って，HMM 状態系列を伸縮させて同一の長さに正規化し，スコア $Score_{DDM}$ を求める．

(5) $Score_{DP}$,$Score_{DDM}$ の重みつき結合スコアが，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，3〜5 の手順を 2 パス目とする 2 段階の処理から成る．このシステムのリスコアリング部 (2 パス目) 以外は 3 節に示した音声検索語検出システムと同じものである．図 6 に示すように，1 パス目及び 2 パス目ともに，全ての HMM 状態の組み合わせに対してあらかじめ距離テーブルを持っておくことで，計算量を大きく抑えることができ，式 (6) の 2 つの項の値の計算時間は同程度で済む．

5. 評価実験

5.1 実験条件

評価対象の音声ドキュメントとして，日本語話し言葉コーパス (CSJ)[13] のコア講演データ (177 講演，約 44 時間) を用いる．CSJ のデータは 200ms 以上のポーズで分割された転記基本単位 (Inter Pausal Unit : IPU) で書き起こしが行われている．この IPU を正解判定の基本単位とし，音声ドキュメント名と IPU の一致により，正解と定義する．

音声認識結果は，NTCIR9 SpokenDoc 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 を用いる．このリファレンス認識結果は表 2 の条件の下，以下の通り認識が行われた．

- 各講演音声に付与された ID が奇数か偶数かによって講演音声を 2 分割する (奇数セット・偶数セット)．
- 偶数セット，奇数セットそれぞれから音響モデル，言語モデルを学習する．
- 偶数セットの音声認識は奇数セットから学習したモデルを，奇数セットの音声認識には偶数セットから学習したモデルを利用して認識を行う．

表 3　音声ドキュメントの認識性能 [%]

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

リファレンスの認識性能を表 3 に示す．表 3 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．

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，CSJ コーパスのコア講演を除く全講演音声を用いる．NTCIR9 SpokenDoc タスクで使用された音響モデルの学習条件に従い，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．使用する分布間距離も条件に従い，ID が奇数のものには偶数で学習した音響モデルによる分布間距離，偶数のものには奇数で学習した音響モデルによる分布間距離を使用する．これにより，評価に用いた音声ドキュメントに対してオープンな評価条件とした．

検索語セットは，NTCIR9 SpokenDoc CORE タスク formal-run で用いられたクエリ 50 個を用いる．

5.2 評価指標

検索性能の評価指標として，Recall，Precision，F-measure(max)，Recall-Precision 曲線，MAP を用いる．F-measure(max) は閾値を動かした時の最大の F-measure の値である．

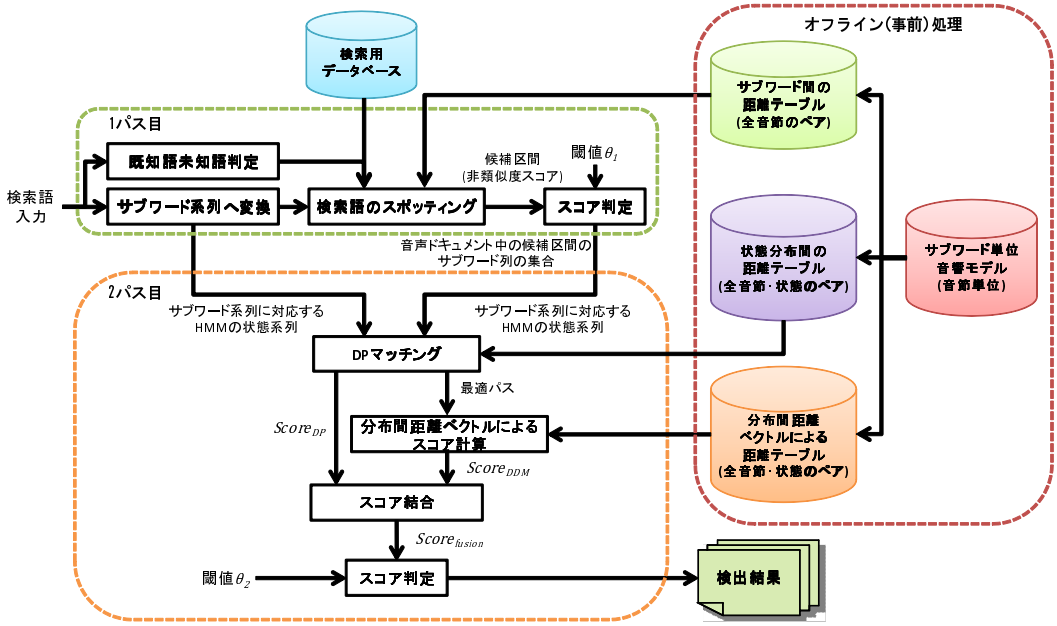

図 6　提案システムのイメージ

表 2　大語彙連続音声認識の条件

認識デコーダ	Julius 4.1.3 (standard)
サンプリング周波数	16kHz
音声認識辞書	約 27,000 語
言語モデル	奇数セット (343 万語)・偶数セット (351 万語) 毎に 学習された (コアを除く) 単語 trigram モデル
音響モデル	32 混合 38 次元 triphone モデル $(MFCC + \Delta MFCC + \Delta\Delta MFCC + \Delta Power + \Delta\Delta Power)$

5.3　従来手法との比較

1 パス目のみのサブワード単位の連続 DP マッチングによりベースライン手法での評価結果と，我々が提案する分布間距離ベクトルを用いたスコア付けによる 2 パス手法の評価結果を，表 4 に示す．また，Recall-Precision 曲線を図 7 に示す．1 パス目 (ベースライン手法) および 2 パス目ともに，表 1 に示した共通の音節単位 HMM を用いており，HMM の学習データがリファレンスの音声認識結果に用いられた triphone HMM と同じ条件となっている．2 パス目の $Score_{DDM}$ には式 (3) で定義した $Score_{DDM\_L1Max}$ を用いていおり，表 4 の 1 パス目のみの値は，1 パス目の閾値を既知語・未知語ごとに調整し，F-measure が最大となったときの値である．そして，2 パス手法の値は，1 パス目の閾値 θ_1，2 パス目のスコア結合の重み α，2 パス目の閾値 θ_2 を既知語・未知語ごとに調整し，2 パス目の出力において F-measure が最大となったときの値である．

表 4 より，1 パスのみ (連続 DP マッチング) より，2 パスを追加した提案手法の方が，全体的に良い性能になることが分かる．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．これにより，分布間距離ベクトルをスコア付けに用いることが有効であると言える．

また，我々の提案手法は 2 パス手法であり，2 パス目の F-measure を最大にするよう調整した 1 パス目の閾値 θ_1 において，1 パス目で出力された時点の Recall が約 81%であり，Precision が約 3%と極端に低いが，2 パス目を通すことにより，Recall は落ちてしまったが，Precision を大幅に上げることができている．

なお現在は，4.3 節で述べたように距離テーブルを用いていること，そして並列処理をすること以外では計算量削減の工夫はあまり施していない．しかし，Xeon X5560 2.80GHz のマシンの仮想環境上で 2core を使用するという条件において，CSJ CORE 講演 (177 講演，約 44 時間) に対して 1 クエリあたりの検索時間は約 0.7 秒程度となっている．計算時間の大まかな内訳は，1 パス目が 0.68 秒，2 パス目が 0.02 秒となっており，1 パス目で充分に候補を絞っているため，2 パス目の計算量を削減することができている．

以降，1 パスのみ (連続 DP マッチングのみ) の手法をベースラインとする．

5.4　$Score_{DDM}$ 算出の定義式の比較

$Score_{DDM}$ の算出式の違いによる各性能評価を表 5 に

表 4　2 パスを追加したことによる性能比較 [%]

	Recall	Precision	F-measure	MAP
1 パスのみ (ベースライン)	50.559	80.088	61.986	0.632
1 パス+2 パス (提案手法)	58.101	85.246	69.103	0.635

* NTCIR9 SpokenDoc CORE タスク [3] で示されているベースラインの性能は F-measure:0.527，MAP:0.595 である．

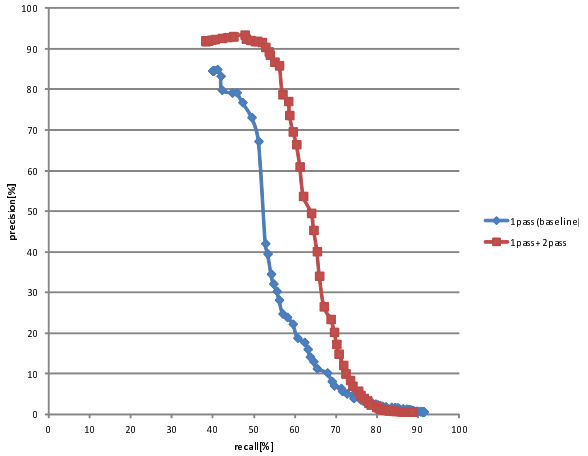

図 7　1 パスのみと 2 パスを追加した提案手法の検出性能の比較

表 5　$Score_{DDM}$ 算出の定義式の違いによる性能比較 [%]

	Recall	Precision	F-measure	MAP
$Score_{DDM\_L1Max}$	58.101	85.246	69.103	0.635
$Score_{DDM\_L2}$	58.101	77.323	66.348	0.627
$Score_{DDM\_L1}$	61.173	70.874	65.667	0.624

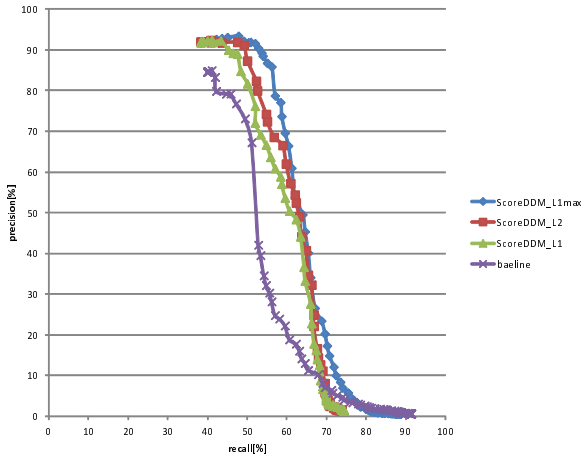

図 8　$Score_{DDM}$ 算出の各定義式による検出性能の比較

示す．また，Recall-Precision 曲線を図 8 に示す．表 5 より，$Score_{DDM\_L1Max}$ を用いたものが 1 番良い F-measure となった．これは，分布間距離ベクトルを特徴量として使用する際は，最もかけ離れているものに着目した方が，非類似性をうまく表わしているのではないかと思われる．そのためこれ以降は，$Score_{DDM}$ 算出式を式 (3) で定義した $Score_{DDM\_L1Max}$ に固定して評価を行う．

表 6　初期モデルの違いによる性能比較 [%]

HMM 学習データ	Recall	Precision	F-measure	MAP
CSJ のみ	58.101	85.246	69.103	0.635
ASJ+JNAS CSJ	56.983	90.667	69.983	0.618

5.5　分布間距離の算出に用いる音響モデルの違いの影響

分布間距離は音響モデルをもとに求める．この際の音響モデルの違いにより，音響的な距離の違いが生じる．そのため，距離算出の際に利用する音響モデルの学習データ及び仕様の違いによる性能評価への影響を調べ考察する．

5.5.1　初期モデル条件の違いによる影響の比較

まず，分布間距離算出に利用する音響モデルを作成する際の，初期モデルの条件の違いによる検索性能への影響を比較する．比較するモデル学習の条件として，以下の 2 つを挙げる．

- CSJ コーパスの学習データ (5.1 節参照) だけを用いてフラットスタート法により作成した音節モデル
- 研究用連続音声データベースの音素バランス文 (ASJ-PB) の全て (男性 30 名，女性 34 名 1，1 名あたり 3 セット) と新聞記事読み上げコーパス (ASJ-JNAS) のうち男女各 103 名分 (男性合計 15911 文，女性合計 15860 文) から作成した初期モデルから CSJ コーパスで追加学習して作成した音節モデル．

音響モデルの仕様は表 1 に示したものであり，これらに対して音声認識の際に利用した音響モデルの作成手順に沿って CSJ コーパスで学習を行った．なお，研究用連続音声データベースの音素バランス文 (ASJ-PB) と新聞記事読み上げコーパス (ASJ-JNAS) を合わせて ASJ+JNAS と表記することにする．この 2 種類の音響モデルをもとに性能評価を行った結果を表 6 に示す．また，Recall-Precision 曲線を図 9 に示す．

表 6 から，ASJ+JNAS で学習した HMM を初期モデルとして CSJ で学習し作成した音節 HMM を使用した場合は Precision が高いポイントにおいて，F-measure が最大となったことが分かる．これは，初期モデルに ASJ+NAS を加えることにより，音響モデルが上手く学習され分布間距離が厳密になったため，Precision が向上したと思われる．CSJ のみから作成した音節 HMM と比べると，F-measure においてはわずかに高く，MAP においてわずかに低い値となった．このことから，F-measure を高めるという目的においては，初期モデルにあまり影響はなかったと言える．

5.5.2　HMM 出力分布の混合数・次元数の比較

3.2.2 節で述べた音響モデルの仕様は 32 混合 38 次元 ($MFCC + \Delta MFCC + \Delta\Delta MFCC + \Delta Power + \Delta\Delta Power$) であった．そこで，分布間距離の算出に使用する音声特徴量 (出力分布) の次元数を，38 次元から 36 次元 ($MFCC + \Delta MFCC + \Delta\Delta MFCC$) に落とした際の影響を調べる．また，32 混合を 16 混合，1 混合に落とし

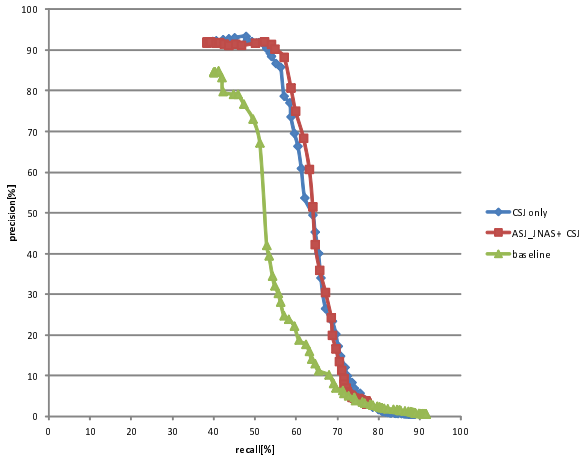

図 9　2 種類の音響モデルによる検出性能の比較

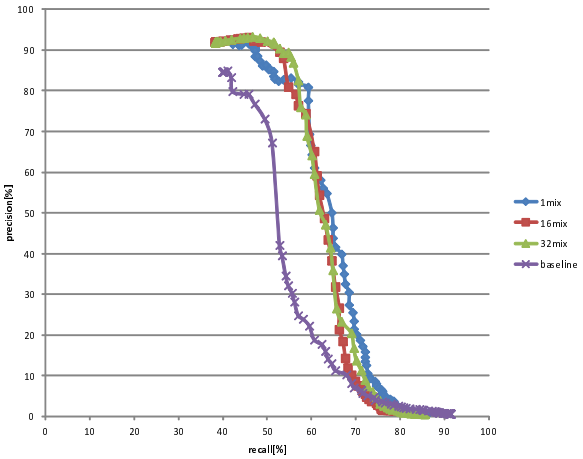

図 10　36 次元においての各混合数の検出性能の比較

表 7　36 次元においての混合数の違いによる性能比較 [%]

HMM 出力分布	Recall	Precision	F-measure	MAP
1 混合	58.939	81.467	68.395	0.670
16 混合	58.101	90.043	70.628	0.639
32 混合	56.983	89.474	69.625	0.613

表 8　38 次元においての混合数の違いによる性能比較 [%]

HMM 出力分布	Recall	Precision	F-measure	MAP
1 混合	57.542	84.774	68.552	0.584
16 混合	57.821	89.610	70.289	0.610
32 混合	58.101	85.246	69.103	0.635

た際の影響も調べる．

次元数を 36 次元とし，混合数を 1 混合，16 混合，32 混合にした場合の評価結果を表 7 に示す．また，Recall-Precision 曲線を図 10 に示す．同様に，次元数を 38 次元とし，混合数を 1 混合，16 混合，32 混合にした場合の評価結果を表 8 に，Recall-Precision 曲線を図 11 に示す．

表 7，表 8 より，36 次元および 38 次元ともに，16 混合で最も良い F-measure となった．式 (1) で定義されているように，各混合分布の要素成分間の Bhattacharyya 距離が最小となる値を分布間距離としている．そのため，混合成分を増やすほど，極端に重みの小さい混合成分でも Bhattacharyya 距離が最小となれば，その距離を分布間距離として採用してしまう可能性が高くなる．つまり，本質的に重要でない成分との距離をとってしまい，分布間距離が曖昧になってしまう恐れがある．そのことから，16 混合程度が妥当な混合数であったといえる．また，大語彙音声認識に用いる音響モデルと比べて正解精度が劣る 1 混合の HMM であっても 1 パス目の性能と比べて大きく改善しており，提案法の分布間距離ベクトルの効果が大きいことが分かる．

次元数の違いに着目してみると，パワーに関する特徴量を用いていない 36 次元と，用いている 38 次元の差は，各混合数においてまちまちであり，一概にどちらが良いということは言えない結果となった．

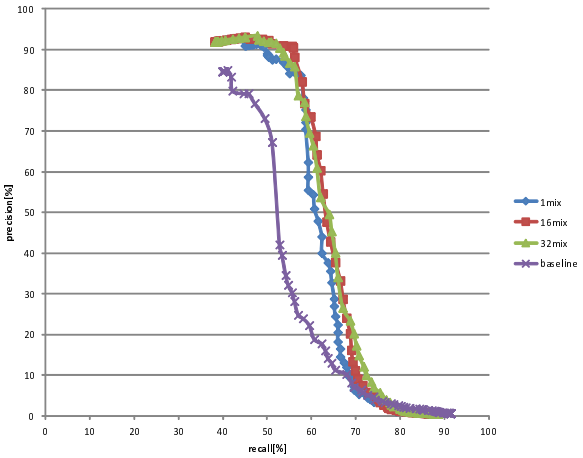

図 11　38 次元においての各混合数の検出性能の比較

5.6　スコア結合の重みの影響

これまでに示してきた評価値は，調整可能なパラメータである，1 パス目の閾値 θ_1，2 パス目のスコア結合の重み α，2 パス目の閾値 θ_2 を既知語・未知語ごとに調整し，最も F-measure が高くなったパラメータでの値である．そこで，我々の提案する分布間距離ベクトルを用いたスコア付けに大きく影響する，2 パス目のスコア結合の重み α を変化させたときの検索性能を解析する．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 12 に示す．なお，既知語クエリに対しては単語ベース認識結果，未知語クエリに対しては音節ベース認識結果を使用している．

スコア結合重み α は，1 に近いほど HMM の状態単位の DP マッチングスコア $Score_{DP}$ を重視し，0 に近いほど分布間距離ベクトルを用いたスコア $Score_{DDM}$ を重視する．よって，$\alpha = 1$ のときは $Score_{DP}$ のみを 2 パス目スコアとして使用し，$\alpha = 0$ のときは $Score_{DDM}$ のみを 2 パス目スコアとして使用することになる．

図 12 から，既知語 (IV) においては重みの影響が小さいことが分かる．しかし，未知語 (OOV) においては 0.2 から 0.6 あたりを頂点とする山なりになっており，既知語・

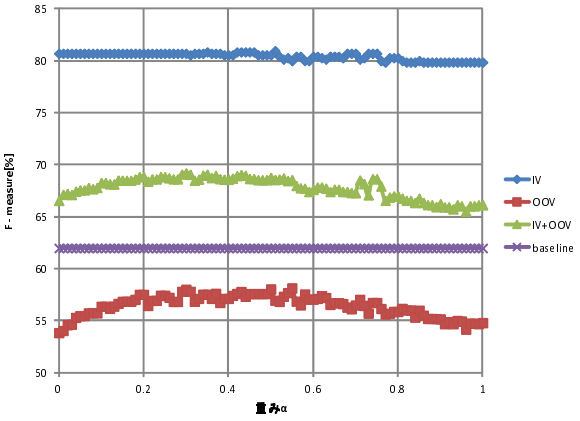

図 12 スコア結合重み α を変化させた際の F 値の推移

未知語を合わせた結果も同様である．そのため，$Score_{DP}$ と $Score_{DDM}$ を組み合わせることは有効であると言える．また，$\alpha = 1$ のポイントや $\alpha = 0$ のポイント，つまり，$Score_{DP}$ のみや $Score_{DDM}$ のみを 2 パス目スコアとして使用した場合も，ベースラインよりも良い性能を示している．つまり，同じ音響モデルの分布間距離を用いる場合でも，あらかじめサブワード単位で求めて局所距離として用いるよりも，サブワード列に対応する状態列として非類似度スコアを DP マッチング法で求めることで検索性能が改善されることを示している．

6. おわりに

本稿では，HMM の状態において多くの相対的な距離情報を要素に持つ分布間距離ベクトルを定義し，この分布間距離ベクトルを特徴量として用いるスコアリング手法を提案した．そして，この手法を 2 パス手法に組み込むことにより，性能の改善を試みた．また，分布間距離を算出するための音響モデルの違いによる検索性能の比較を行った．

NTCIR9 SpokenDoc CORE タスク formal-run のテストコレクションにおいて，連続 DP マッチングのみを用いるベースライン手法での F-measure が 0.61 だったのに対して，提案手法は，0.69 と顕著に改善された．このことから，分布間距離ベクトルを特徴量として利用することが有効であると思われる．

今後の課題として，計算量の削減が挙げられる．中川らは，音節単位で認識した音節ラティスの n-gram を検索用インデックスとすることで未知語に頑健で高速に動作する検索手法を提案している [14]．このようなインデキシングによる高速化手法との併用も可能と今後考えられる．

その他に，閾値の自動推定といったことも課題として挙げられる．我々の閾値の決定法は，与えられた検索語セットに対して F-measure が最大となるように調整したものであるため，未知の検索語に対して頑健であるとは言い切れない．Zhang ら [15] は，既知語の検索語に対してのアプローチではあるが，事前に検出スコアと誤検出確率の関係性を学習し，検索語に対しての検出スコアから誤検出確率を推定するといった手法を提案している．これはスコアの正規化手法の一つであり，未知の検索語に対しても一貫したスコアを与えることができる．我々は，最近，識別モデルの条件付き確率場 (CRF) による検出判定を検討している．これにより音声ドキュメントやクエリの内容，または非類似度尺度の性質などを総合的に評価した検出・非検出の判定を行うことが可能となると考えている．

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”, Proceedings of NTCIR-9 Workshop Meeting, pp.223-235 (2011.12.6-9).

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

[5] 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).

[6] 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”, Vol.2012-SLP-94, No.15, pp.1-6 (2012).

[7] A.Muscariello, et al. ：“Zero-resource audio-only spoken term detection based on a combination of template matching techniques”, Proceedings of INTERSPEECH 2011, pp.921-924 (2011).

[8] 神田直之, 他：“任意語彙音声発話検索のための多段階リスコアリング手法の性能評価”, 音声ドキュメント処理ワークショップ講演論文集 2, pp.73-78 (2008.2).

[9] Hiromitsu Nishizaki, et al. :“Spoken Term Detection Using Multiple Speech Recognizers’ Outputs at NTCIR-9 SpokenDoc STD subtask”, Proceedings of NTCIR-9 Workshop Meeting, pp.236-241 (2011.12.6-9).

[10] 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).

[11] 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).

[12] 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).

[13] 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).

[14] 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).

[15] Zhang.B , et al. ：“White Listing and Score Normalization for Keyword Spotting of Noisy Speech”, Proceedings of InterSpeech 2012, (2012).