Vol.2016-SLP-112 No.11
2016/7/29

音声クエリによる音声検索語検出のための認識結果および DNN ベースの特徴抽出と再照合手法の比較評価

大石　修司[1,a)]　松葉　達也[1,b)]　甲斐 充彦[1,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 Term Detection : STD) の研究が盛んに行われている．本稿では，音声によるクエリ入力を想定した STD の手法を対象として考える．従来の STD 手法は，音声認識を行い検索対象及び検索語 (クエリ) を元にサブワード (音素や音節) 列などの認識結果を用いて検索を行う．以前に我々はサブワード単位音響モデルのパラメータから求める様々な音響的類似度を用いる方法を提案し，STD で検索性能を改善した．しかし音声クエリを用いる場合，未知語 (OOV) や誤認識の影響はより大きくなり，検索性能を低下させてしまう．そのため本稿では音声認識結果を用いる従来の STD 手法によるスポッティングを行った後，話者や環境の違いに頑健な DNN に基づいた特徴量によって再照合を行う手法を提案する．さらに，認識結果より得られる信頼度やクエリの長さの特徴を素性として書き起こしを含む開発用データによって自動的に構築する検出事例から学習したスコアリングモデルにより，正規化されたスコアを得る．これらの方法の併用により，更なる STD 精度の改善が得られた．

Rescoring with ASR output-based and DNN-based features extraction for improved query-by-example spoken term detection

Shuji Oishi[1,a)]　Tatsuya Matsuba[1,b)]　Atsuhiko Kai[1,c)]

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．一般的には自動音声認識 (ASR) によって得られる認識結果を利用する手法が用いられている．しかし未知語 (OOV) の問題によって STD の性能を悪化させる．

OOV の問題に対応するために DNN に基づいた特徴量による音響的なマッチング手法が低リソースの STD タスクで提案されている [1]，[2]，[3]．しかし特徴量ベースの手法は時間がとてもかかり，言語情報が豊富な言語での ASR に基づいた手法に対しての改善はほとんど見込めない．また一方で，関連研究として ASR に基づいたサブワード単位の音響的類似度を用いる手法が提案されており未知語のクエリに対する検索精度を改善している [4]．我々の先行研究 [5]，[6] では音節の状態単位の音響的類似度を用いる手法を提案して検出精度を改善している．

本稿では，認識結果を用いる STD 手法と DNN に基づいた音声特徴量を用いる照合手法を併用する手法を提案する．同様なアプローチは，最近の Query-by-example STD の研究 [7] において効果が示されている．我々の方法では効率的な処理を実現するため，1 パス目の音節の状態単位のスポッティングを行い，2 パス目に DNN ベースの音声特徴量による再照合を行う．そして，認識結果より得られる信頼度やクエリの長さの特徴を素性として加えて，自動的に構築した検出事例データをもとにスコア統合モデルを学習し，スコアの正規化に利用する．

本稿の実験は，NTCIR-12 SpokenQuery&Doc-2[8] のタ

[1] 静岡大学大学院総合科学技術研究科
Graduate School of Integrated Science and Technology, Shizuoka University
a) oishi@spa.sys.eng.shizuoka.ac.jp
b) matsuba@spa.sys.eng.shizuoka.ac.jp
c) kai.atsuhiko@shizuoka.ac.jp

スク定義を基本として，我々の先行研究で提案されているサブワード単位と音節の状態単位の音響的類似度を用いた手法をベースラインとして，DNN に基づいた音声特徴量の再照合手法を比較する．さらに音声認識結果から得られる素性をスコアの統合にモデルの素性として加えることでの性能を比較評価する．

2. ベースライン STD システム

我々がこれまで構築してきた STD システム [10]，[6] をもとにした二つのベースライン STD システムを用いる．

2.1 サブワード間の音響的距離を利用した STD 手法

一つ目のベースライン STD システム (Baseline1) はクエリと検索対象ドキュメントのサブワード間の音響的な非類似度を利用してスポッティングを行う．

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 [11] がある．我々は，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 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状態系列間の DTW 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，クエリと類似する区間のスポッティングを行う．

2.2 音節 HMM の状態間の音響的距離を利用した STD 手法

二つ目のベースライン STD システム (Baseline2) は前節で述べたサブワード単位の音響的距離を利用したスポッティングによる STD 手法をサブワード単位ではなく，より細かな音節 HMM の状態単位と設定した手法である．つまり，前節の方法で局所距離が音節単位であったのに対して，式 (1) による状態同士の分布間距離を用いる．先行研究 [9] ではこの Baseline2 の手法は Baseline1 に対して検出精度を改善することを示した．同様に紺野ら [12] は状態間やフレーム間照合で改善されることを示している．

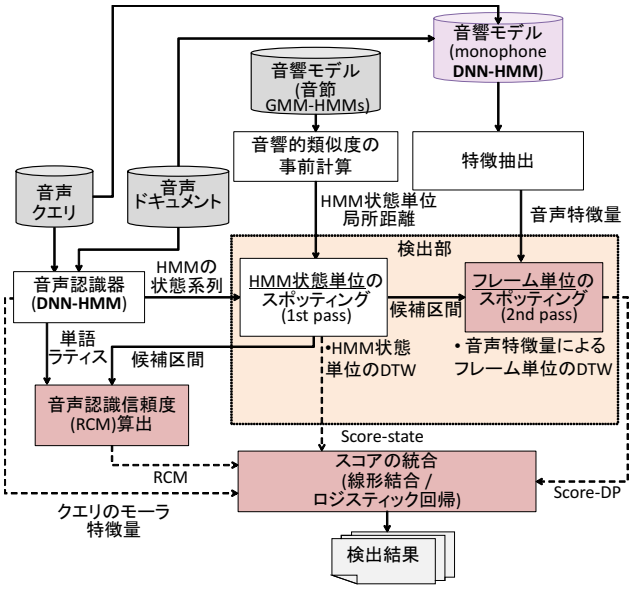

図 1 DNN ベースの特徴量による再照合を用いた STD システム構成

3. 提案手法

3.1 概要

図 1 に我々が提案するシステムの概念図を示す．このシステムは誤認識に対する STD の性能を効率的に改善するために 2 パスの構造をとる．まず 1 パス目は 2.2 節で述べた音節 HMM の状態単位のスポッティングを行う．次に 2 パス目は，1 パス目で検出した候補区間に対して DNN に基づいた特徴量を用いてフレーム単位の DTW によって再照合をする．本研究では，DNN に基づく特徴量としては 3.2 節で述べる事後確率特徴量を用いる．最後にロジスティック回帰を用いてスコアの統合を行う．その際に，ロジスティック回帰の素性として 3.3 節で述べる音声認識信頼度 (RCM) 及びクエリの長さに関する特徴 (素性) を加える．また，我々はスコアの結合方法を比較するために線形結合によるスコアの結合を行った．

$$Score_{Final} = \alpha Score_{DP} + (1-\alpha) Score_{state} \qquad (2)$$

1 パス目のスコアを $Score_{state}$ とし，2 パス目のスコアを $Score_{DP}$ とした．

3.2 DNN ベースの特徴量による音響的なマッチング

DTW に基づいた STD の手法では事後確率特徴量が局所距離を計算するための音声特徴ベクトルとしてよく用いられている．MFCC を事後確率特徴量に変換するために GMM または多層パーセプトロンに基づいた音響モデルが用いられる [13][14]．

我々は音素 (monophone) 状態単位を出力とするディープニューラルネットワーク (DNN) を用いて音素の事後確率特徴量を抽出した．事後確率特徴量の局所距離は二つの事後確率特徴ベクトルを $\boldsymbol{x}$ ，$\boldsymbol{y}$ とすると次式のように求

める．

$$d(\boldsymbol{x}, \boldsymbol{y}) = -log(\boldsymbol{x} \cdot \boldsymbol{y}) \quad (3)$$

DNN は事前学習として制限付きボルツマンマシン (RBM) で学習し，クロスエントロピー規準による識別学習を行う．

図 1 の 2 パス目は事後確率特徴量の局所距離を用いたフレーム単位の DTW を行う．この 2 パス目の DTW は 1 パス目の候補区間を用いるため，1 パス目の DTW での検出ミスに大きく影響を受ける．そのため我々は候補区間の始まりと終わりを一定の長さ (β フレーム) だけ拡張した区間に対して端点フリーの DTW を行うことで 1 パス目の誤りを許容する照合方式を用いる．

3.3 ロジスティック回帰によるスコアの統合

我々はロジスティック回帰モデルの学習に書き起こしがある大規模音声コーパスで自動的に検出事例を生成し，開発セットとして用いる．そして事前に開発用データを用いて認識結果や検出候補について得られるいくつかの特徴とマッチングから得られるスコアを合わせて，図 1 の 1 パス目によって検出された候補区間の正誤の教師データをロジスティック回帰モデルによって学習しておく．そして学習したロジスティック回帰モデルを用いて正解検出確率を推定し，統合かつ正規化された検出スコアとする．我々は切り出す際に長過ぎる (12 モーラ以上) クエリと短すぎる (3 モーラより小さい) クエリを除き，さらに tf-idf を算出し一つの講演に多数存在するものやほとんど存在しないものはクエリの候補から除いた．開発用の音声クエリは音響モデルを学習するために用いられるアライメントの結果より自動で抽出される．

我々は 1 パス目及び 2 パス目のマッチングのスコア以外に相補的な情報として音声認識の結果から得られる 3.3.1 節の音声認識信頼度 (RCM) と 3.3.2 節のクエリ長の特徴 (素性) をロジスティック回帰モデルの素性に追加する．

3.3.1 音声認識信頼度 (RCM)

音声認識システムが出力する各文仮説について，認識システムがどれだけの確信を持ってその文仮説を出力したかの尺度を信頼度という．我々は 1 パス目で検出した候補区間の信頼度として，音声認識信頼度 (RCM) を検出対象の音声ドキュメントのラティス上において，1 パス目の候補区間について最も高い尤度スコアの事後確率の平均によって算出する．スポッティングで得られた検出区間の状態系列を $B = \{B_1, \cdots, B_Y\}$ とし，対応する区間の音声を $X = \{X_1, \cdots, X_Y\}$ とすると、B に対する音声認識信頼度 $RCM(B)$ を以下のように定義する．

$$RCM(B) = \frac{\sum_{k=1}^{Y} P(B_k|X_k)}{Y} \quad (4)$$

ここで $P(B_k|X_k)$ はラティスの 1 ベスト上の状態 B_k の事後確率を表す．ただし，本研究では単純に以下の式で近似的に RCM を求める．

$$RCM(B) = RCM(T_B) = \frac{\sum_{t=i}^{j} \max_s \{P(s|t)\}}{j - i + 1} \quad (5)$$

$T_B = \{i, \cdots, j\}$ は B の音声フレームを表し，$P(s|t)$ はラティス上の t フレームにおける音素 s の事後確率を表す．

表 1 クエリの長さに関する特徴 (素性) の例

クエリ	モーラ数	バイナリ特徴量 L_k 4	6	8	10
ni ho n ji n (日本人)	5	0	1	1	1
shi zu o ka da i ga ku (静岡大学)	8	0	0	1	1
a ri ga to u go za i ma su (ありがとうございます)	10	0	0	0	1

3.3.2 クエリ長の特徴 (素性)

経験的に，前述のように求まる STD のスコアはクエリの長さによってスコアの分布に影響を与えると考えられる．そこで我々はクエリ長をモーラ数を利用して表現した特徴を用いる．本研究では音声クエリを想定するためクエリの長さ (モーラ数) は大語彙音声認識結果から推定する．具体的には，表 1 のような特定のモーラ数 (4，6，8，10 モーラ) よりも大きいかどうかを表現したバイナリの特徴を用いた．特定のモーラ数 L_k 以下ならば 0，大きければ 1 となる．

4. 音声クエリによる STD の評価実験

4.1 実験条件

本研究では，音声クエリによる音声検索語検出の評価実験を行った．我々は，NTCIR-12 SpokenQuery&Doc-2[8] のタスクに従って STD の評価を行った．開発セットと評価セットを用いることで提案手法の頑健性を検証した．開発セットに相当するものとして NTCIR12 では dryrun のクエリが定義されるが規模が小さいため我々は別途，CSJ のコア講演 (177 講演, 約 44 時間) を用いた. また 3.3 節で述べたスコア統合のモデル学習のため，3.2 節で述べた特徴量抽出用モデル (DNN-HMM) の学習に用いた CSJ の講演音声 (コア以外の 910 講演) から手動の書き起こしを用いて自動で切り出した 620 クエリ (既知語 457，未知語 163) を用いた. 我々は切り出す際に長過ぎる (12 モーラ以上) クエリと短すぎる (3 モーラより小さい) クエリを除いた．開発セットの音声認識には 3.2 節の特徴抽出用のモデルと同様のモデル (音響モデル 910 講演, 言語モデル 2503 講演) を用いた. 検索対象の CSJ のコア講演に対する単語正解率は 74.0%であった.

評価セットは NTCIR12[8] で用いられた formalrun を利用した．検索対象は音声ドキュメントワークショップ (104 講演) であり，クエリ数は 162(既知語 85, 未知語 77)

である. クエリは 10 人分の音声データが収録されている. NTCIR12 STD のデータセットには 1 つのクエリ内に複数の語句が定義されている場合がある. そのため我々は語句と語句の認識結果の間に 0.2 秒以上の無音区間が認識されている場合語句を分割する方法を行った. 評価セットの音声認識には NTCIR12 のオーガナイザによって提供された DNN-HMM(音響モデル 950 講演，言語モデル 2525 講演) を用いた.

我々は Kaldi ツールキット [16] を用いて，3.2 節の特徴抽出用のモデルとして monophone の DNN を学習した. DNN は 7 層 (入力層, 隠れ層 5 層, 出力層) から構成される. 出力層のユニット数は 145 である. 入力特徴量としては 39 次元の MFCC(MFCC+power+ΔMFCC+Δpower+ΔΔ MFCC+ΔΔ power) に平均分散正規化 (CMVN) と線形判別分析 (LDA) を適用したものを用いて学習を行った. また事後確率特徴量は monophone の状態数と同じ 145 次元の特徴である.

検索精度の評価指標としては, 正解の閾値を変化させたときに F 値 (F-measure) の最大値 (F(max)) と MAP 値を用いた.

提案した 2 パス構造の 1 パス目の検出閾値はクエリごとに 1000 個の候補区間となるように設定した.

4.2 比較する STD 手法

比較を行う STD 手法は以下の通りである.

Baseline1(syll_spot): 2 節で述べたサブワード間の音響的類似度による手法. (第 1 パスのみ)

Baseline2(state_spot): 2 節で述べたサブワードの状態間の音響的類似度による手法. (第 1 パスのみ)

+post: 3.2 節で述べた事後確率特徴量 (posteriorgram) を用いた照合スコアとの併用手法.

+RCM: 3.3.1 節の RCM との組み合わせ手法.

+mora: 3.3.2 節で述べたクエリ長の特徴 (素性) との組み合わせ手法.

LC: 線形結合によるスコアの統合.

LR: ロジスティック回帰によるスコアの統合.

4.3 開発用セットでの評価結果

1 パス構造の二つのベースラインと DNN ベースの特徴量による照合との組み合わせ手法の開発セットの結果を以下の表 2 に示す. 評価結果から DNN ベースの特徴量を用いた照合との併用手法が Baseline1，2 の性能を上回っており，特に事後確率特徴量とクエリ長の特徴 (素性) のロジスティック回帰による組み合わせ手法で F 値及び MAP 値において改善した. 一方で RCM をロジスティック回帰の素性として加えた手法は性能の改善が見られなかった.

4.4 評価用セットでの評価結果

開発セットで調整を行った線形結合及びロジスティック回帰のパラメータを用いて評価セットで評価実験を行った結果を表 3 に示す. これらの評価結果から，開発セットの場合とは異なり，DNN に基づいた音声特徴量による照合手法を組み合わせることでの F 値の改善が見られなかった. しかし一方でクエリ長の特徴 (素性) の組み合わせ手法によって大きな F 値の改善が見られた. F 値と MAP 値を安定して改善しているのは，クエリ長の特徴 (素性) と事後確率特徴量の両者を組み合わせた手法であった. また図 2 に評価セットにおける Recall-Precison 曲線を示す. 図 2 よりクエリ長の特徴を組み合わせた手法は他の手法と比べ recall が小さいときに Precision の値が圧倒的に高くなる傾向がある. 逆に事後確率特徴量によるマッチングスコアを組み合わせる手法は recall が小さい時に Precision の値が低い. しかし，recall が高い場合はベースラインよりも高い Precision を保っている. また表 4 に OOV(未知語) クエリのみでの評価結果を示し図 3 に Recall-Precision 曲線を描く. . 評価結果より未知語の場合最も性能が高かったのは提案する全ての特徴を組み合わせた手法であった. DNN に基づいた事後確立特徴量の組み合わせ手法は必ず性能をベースラインの手法と比べて改善しており，開発セットにおいても同様の結果であった. 未知語クエリのみで評価した場合の Recall-Precision 曲線に対してもスコアの組み合わせは 2 と同様の傾向がみられた.

表 2 各手法における STD 性能評価 (開発セット)

手法	F(max)	MAP
Baseline1(syll_spot)	19.78	53.76
Baseline2(state_spot)	20.86	55.40
state_spot+post(LC)	25.58	57.69
state_spot+post(LR)	27.55	58.66
state_spot+RCM(LR)	20.83	53.91
state_spot+mora(LR)	21.75	54.08
state_spot+post+mora(LR)	**29.88**	58.46
state_spot+RCM+mora(LR)	21.72	53.92
state_spot+post+RCM(LR)	27.52	**58.74**
state_spot+post+RCM+mora(LR)	29.78	58.54

表 3 各手法における STD 性能評価 (評価セット)

手法	F(max)	MAP
Baseline1(syll_spot)	38.32	64.42
Baseline2(state_spot)	45.94	66.03
state_spot+post(LC)	42.75	72.27
state_spot+post(LR)	38.87	**72.85**
state_spot+RCM(LR)	45.10	66.08
state_spot+mora(LR)	**54.19**	65.27
state_spot+post+mora(LR)	51.62	72.74
state_spot+RCM+mora(LR)	53.86	64.87
state_spot+post+RCM(LR)	38.91	72.78
state_spot+post+RCM+mora(LR)	51.64	72.73

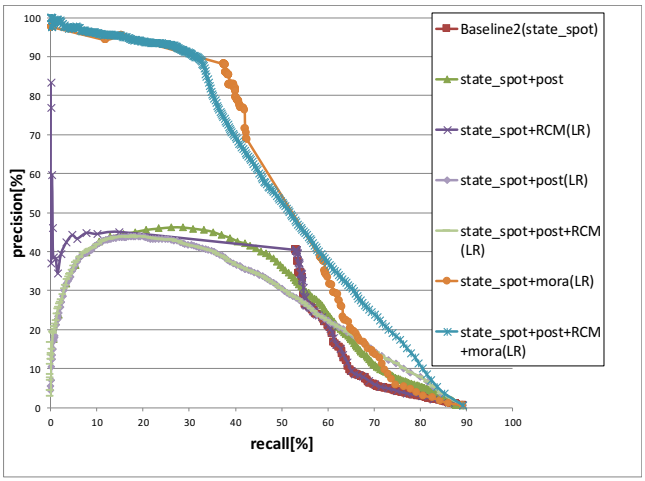

図 2 Recall-Precisionn 曲線 (評価セット)

表 4 各手法における OOV(未知語) クエリのみでの STD 性能評価 (評価セット)

手法	F(max)	MAP
Baseline2(state_spot)	19.71	51.20
state_spot+post(LR)	23.00	**60.55**
state_spot+RCM(LR)	19.55	50.46
state_spot+mora(LR)	23.56	49.68
state_spot+post+RCM(LR)	23.14	60.46
state_spot+post+RCM+mora(LR)	**31.79**	60.01

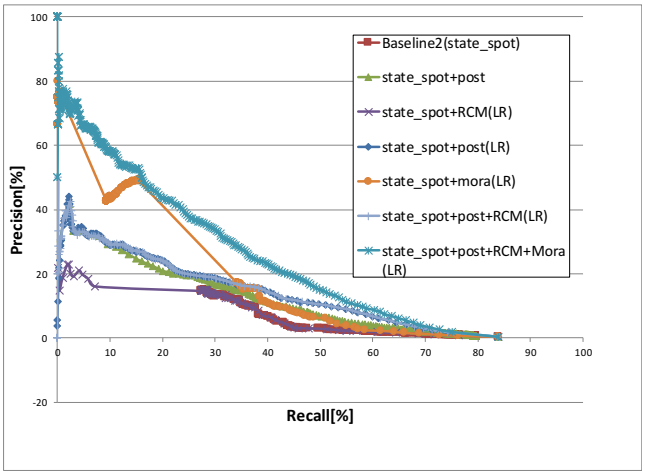

図 3 未知語クエリの Recall-Precisionn 曲線 (評価セット)

4.5 クエリ長の特徴 (素性) に対する考察

まず最初に開発セット (CSJ) と評価セット (formalrun) のクエリの分布を図 4 に示す.

評価用セットの F 値において大きな改善を示したクエリ長の特徴 (素性) の組み合わせについて考察をした. まず表 5 にクエリ長の特徴 (素性) をどの程度の長さまで入れるかを比較したものを示す. 表 5 よりクエリ長の特徴 (素性) は長さが 4 モーラかどうかという情報だけで改善することが示された.

表 5 の結果から 4 モーラのクエリ特徴量をロジスティック回帰の素性として与えた時に, 4 モーラ以下のクエリと 4 モーラより大きいクエリではどちらのクエリの検索性能を

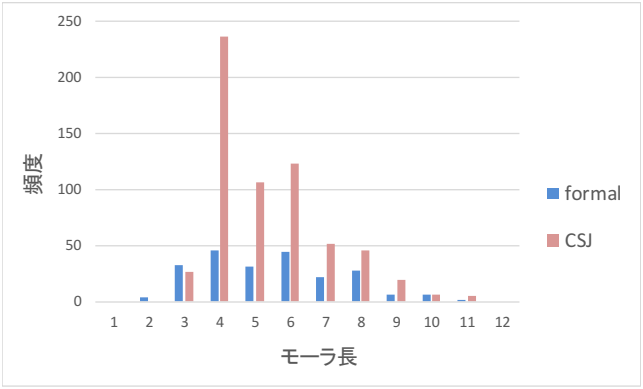

図 4 開発セット (CSJ) 及び評価セット (formalrun) のクエリの分布

改善しているか調べることにした. その結果を表 6 に示す. 表 6 より 4 モーラより大きい, つまりは 4 モーラのクエリよりも長いクエリで性能の改善が見られた. また図 5 に閾値 4 モーラのクエリ特徴量の有無によってロジスティック回帰のスコア (類似度) がどのように変化するか比較した結果を示す. クエリ長の特徴 (素性) を入れる前と後でスコアの差をみると, 4 モーラよりも長いクエリに関してはスコア (類似度) を高くする傾向があり, 4 モーラ以下のクエリに対してはスコアを低くする傾向があることが分かった. よってこのことから 4 モーラよりも長いクエリに関してはスコアが大きくなるようにロジスティック回帰モデルが学習され, クエリ長の特徴 (素性) を入れる前に比べて正解検出のスコアを大きくすることができていると考えられる.

5. おわりに

本稿では, 音声クエリによる STD において従来の方法で絞られた検出区間に対して, DNN に基づいた再照合を行った. さらに音声認識から得られる音声認識信頼度 (RCM) やクエリ長の特徴をスコア統合のモデルの素性として未知語クエリや誤認識に対して STD 精度の改善を図った. 評価結果より DNN に基づいた特徴量による再照合手法は有効であり, さらにクエリ長の特徴 (素性) と組み合わせるこ

表 5 クエリ長の特徴 (素性) 変更による性能比較 (評価セット)

手法	F(max)	MAP
Baseline2(state_spot)	45.94	66.03
state_spot+post+RCM+mora4	51.00	72.74
state_spot+post+RCM+mora4,6	51.38	72.73
state_spot+post+RCM+mora4,6,8	51.32	72.73
state_spot+post+RCM+mora4,6,8,10	51.64	72.73

表 6 クエリ長の特徴 (素性) の有無による性能比較

	4 モーラ以下		4 モーラより長い	
手法	F(max)	MAP	F(max)	MAP
state_spot+post+RCM	26.99	52.41	50.68	85.97
state_spot+post+RCM +mora4	26.39	52.50	68.62	85.86

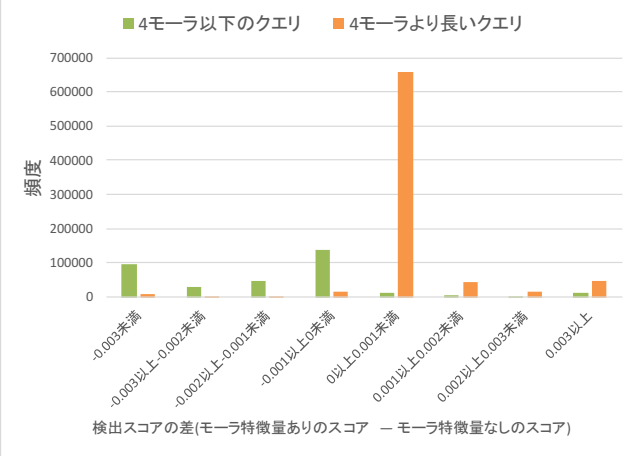

図 5　閾値 4 モーラのクエリ長特徴量の有無によるスコアの変動

とで更なる改善が見られた．

今後の課題 [10] として，クエリの今回 RCM は検索対象のドキュメントの音声認識結果のみから求めておりクエリとドキュメントの類似度を扱っていなかったため，RCM の組み合わせであまり性能の改善が見られなかった．更なる改善を実現するためにクエリとの類似性を考慮した音声認識信頼度を求める必要がある．

参考文献

[1] G. Mantena, and K. Prahallad : “Use of articulatory bottle-neck features for query- by-example spoken term detection in low resource scenarios,” Proc. of ICASSP, (2014).
[2] J. Tejedor, I. Szoke, and M. Fapso : “Novel methods for query selection and query combination in query-by-example spoken term detection,” Proc. of SSCS, (2010).
[3] H. Wang, T. Lee, C. Leung, B. Ma, and H. Li : “Acoustic Segment Modeling with Spectral Clustering Methods, ” IEEE/ACM Transaction on Audio, Speech,
[4] S. Nakagawa, K. Iwami, Y. Fujii, and K. Ymamoto : “A robust/fast spoken term detection method based on a syllable n-gram index with a distance metric,” Speech Communication, Vol.55, pp.470-485, (2013).
[5] N. Yamamoto, and A. Kai : “Using acoustic dissimilarity measures based on state-level distance vector representation for improved spoken term detection,” Proc. of APSIPA ASC, (2013).
[6] M. Makino, N. Yamamoto, and A. Kai : “Utilizing State-level Distance Vector Representation for Improved Spoken Term Detection by Text and Spoken Queries” Proc. of INTERSPEECH, (2014)
[7] J. Tejedor D. Toledano, P. Lopez, L. Docio C. Garcia “Comparison of ALBAYZIN query-by-example spoken term detection 2012 and 2014 evaluations ” Eurasip Journal on Audio, Speech, and Music Processing, 2016:1 (2016)
[8] Tomoyosi Akiba, Hiromitsu Nishizaki, Hiroaki Nanjo, Gareth J. F. Jones: “Overview of the NTCIR-12 SpokenQuery&Doc-2 task,” Proc. of the NTCIR-12 Conference, Tokyo, Japan (2016).
[9] 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).
[10] 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).
[11] 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).
[12] 伊藤慶明、紺野良太、小嶋和徳、李時旭、田中和世、”音声中の検索後検出におけるフレームレベル状態系列間照合方式”、電子情報通信学会信学技報、vol 115、no. 146、SP2015-37、pp. 7-12、2015 年 7 月
[13] T. J. Hazen, W. Shen, and C. White: “Query-by-example spoken term detection using phonetic posteriorgram templates,” Proc. ASRU, pp. 421-426, (2009).
[14] Y. Zhang and J. Glass : “Towards multi-speaker unsupervised speech
[15] S. Oishi T. Matsuba M .Mitsuaki K. Atsuhiko“Combining state-level spotting and posterior-based acoustic match
for improved query-by-example spoken term detection ” Interspeech 2016, to appear, (2016)
[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely: “The Kaldi Speech Recognition Toolkit,” Proc. of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, (2011).