情報処理学会論文誌 Vol.55 No.12 2537–2545 (Dec. 2014)

音響情報のベクトル量子化を用いた音声ドキュメントからの検索語検出

坂本 伊織[1,†1] 松永 徹[2,†2] 趙 國[3] 山下 洋一[3,a)]

受付日 2014年4月11日, 採録日 2014年9月12日

概要：音声を含むマルチメディアコンテンツを有効に利用するには，音声認識に基づいた情報検索が重要な技術となる．与えられた検索語を音声データから検出する音声中の検索語検出（STD：Spoken Term Detection）の研究が広く行われている．本論文では，検索対象の音声ドキュメントの表現手法として，音響情報をベクトル量子化（VQ）して得られる VQ コード列を用い，テキスト入力された検索語と照合する STD 手法を提案する．VQ コードと音素の関連度をあらかじめ話者ごとに学習しておくことによって，音声ドキュメントの VQ コード列と検索語の音素列の照合を行う．評価実験において，音声ドキュメントをサブワード列で表現する従来手法よりも高い検出性能が得られた．さらに，異なる音声認識結果で学習した関連度で照合を行った複数の検出結果を統合することによって検出性能が改善されることが示されている．

キーワード：音声中の検索語検出，ベクトル量子化，V-P スコア，テキストクエリ，連続 DP

Spoken Term Detection Using Vector Quantization of Spoken Documents

Iori Sakamoto[1,†1] Toru Matsunaga[2,†2] Kook Cho[3] Yoichi Yamashita[3,a)]

Received: April 11, 2014, Accepted: September 12, 2014

Abstract: The information retrieval based on speech recognition is an important technique to easily access large amount of multimedia contents including speech. The development of spoken term detection (STD) techniques, which detect a given word or phrase from spoken documents, is widely conducted. This paper proposes a new STD method based on matching between a text query and VQ (Vector Quantization) code sequences which represent spoken documents. The co-occurrence scores between a VQ code and a subword are a priori trained for each speaker. The continuous DP matching detects a subword sequence of the query term from VQ code sequences using the co-occurrence as a local score of matching. Evaluation experiments show that the proposed method improves the performance of STD. A fusion method of multiple detection results using the different cooccurrence scores gives more improvement of STD performance.

Keywords: STD (Spoken Term Detection), vector quantization, V-P score, text query, continuous DP

1. はじめに

データ蓄積容量の飛躍的な増大やインターネットの利用によって，音声を含むマルチメディアコンテンツが大規模化し，対象コンテンツへのアクセスが容易になっている．

[1] 立命館大学大学院情報理工学研究科
Graduate School of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga 525–8577, Japan
[2] 立命館大学大学院理工学研究科
Graduate School of Science and Engineering, Ritsumeikan University, Kusatsu, Shiga 525–8577, Japan
[3] 立命館大学情報理工学部
College of Information and Engineering, Ritsumeikan University, Kusatsu, Shiga 525–8577, Japan
[†1] 現在，村田機械株式会社
Presently with Murata Machinery, Ltd
[†2] 現在，日本電気株式会社
Presently with NEC Corporation
[a)] yama@media.ritsumei.ac.jp

これらのコンテンツを効率的に利用するために音声データに対する検索技術のニーズが高まっており，音声ドキュメントから与えられたキーワード（検索語）を検出する音声中の検索語検出（STD：Spoken Term Detection）の研究が活性化してきている [1], [2], [3]．講演や音声ブログなどの音声ドキュメントをあらかじめ大語彙連続音声認識技術によって文字テキスト化しておき，単語列として表現された音声ドキュメントと検索語との照合を行うことによってSTD が実現可能である．しかし，この手法には，音声認識誤りによって検出精度が劣化するという問題だけでなく，検出すべき検索語が音声認識の辞書に含まれていない場合には，検索語が原理的に検出できない，いわゆる未知語の問題がある．

STD における未知語の問題を解決するために，音声ドキュメントを音素や音節などのサブワード単位で認識して検索語と照合する手法が広く用いられている [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]．この手法では，検索対象の音声ドキュメントを，音素や音節などのサブワード列に音声認識によって事前に変換しておく．サブワード列への変換では，単語を単位とする大語彙連続音声認識によって得られる単語列をサブワードへ自動変換する手法やサブワードを単位とする音声認識を行う手法などが用いられる．テキストで与えられた検索語をサブワード列へ展開し，検索対象の音声ドキュメントのサブワード列に対し連続 DP マッチングなどの手法によって照合を行う．照合においては，検索語サブワード列と完全一致するサブワード列だけでなく類似した系列も検出することにより，大語彙連続音声認識における未知語でも検出できる可能性がある．また，音声認識誤りに対応するため，異なる音声認識システムでの認識結果を組み合わせて音声ドキュメントを表現することで，検出精度の改善が試みられている [9], [12], [13], [14]．しかし，音声ドキュメントを音素などのサブワード列やサブワードラティスとして表現するため，もとの音声ドキュメントの持つ音響情報をかなり粗く近似した記号列を対象に照合を行うことになる．サブワード間の照合では，音素の弁別素性や音声認識の音響モデルに基づいてサブワード間の距離を事前に定義しておくことによって，音声ドキュメント中のサブワードが正しく認識されなかった場合にも，たとえば /a/ と /i/ よりも /a/ と /o/ の方が類似性が高いなど，検索語との類似性に応じた評価をすることができる．この場合，たとえば /o/ が間違って /a/ と認識されたとき，認識された /a/ が典型的な /a/ なのか，/o/ に近い /a/ なのかといった音響的な特性は，照合時には無視される．音声ドキュメントを音素などのサブワード列として表現すると，もとの音声ドキュメントの持つ音響情報をかなり粗く近似した記号列を対象に照合を行うことになる．サブワードよりも詳細な記号列で音声ドキュメントを表現することによって，照合時に無視される音響的多様性を低減させることができる可能性があると考えられる．

大量の音声ドキュメントに対する高精度な検索を実現するには，どのような形式で音声ドキュメントを表現しておくかが 1 つの重要な問題となる．本論文では，サブワードよりも音響的特徴の多様性を表現できる形式として音響情報をベクトル量子化（VQ：Vector Quantization）によって離散化した VQ コードを考え，VQ コード列を音声ドキュメントの表現形式として用いる．文字テキストとして与えられた検索語をサブワードの 1 つである音素の系列に変換し，VQ コード列と照合する方式を提案する [16], [17]．この手法では，各 VQ コードと音素との関連度を共起関係に基づいてあらかじめ学習しておく．話者により音響的特徴の広がりが異なるため，VQ は話者ごとに行う．関連度を照合における局所スコアとし，連続 DP マッチング [18] によって検索語の候補区間を決定し，候補区間における音素の時間構造の不自然さも考慮することで検索語の検出を行う．さらに，異なる音声認識結果に基づいて VQ コードと音素との関連度を複数学習し，各関連度を用いた複数の検出結果を統合することによって STD の性能向上を試みる．

以下，2 章で提案する検索語検出の手法，3 章で評価実験について述べ，最後に，4 章で結論と今後の課題について述べる．

2. VQ コード列を用いた検索語検出

2.1 検索語検出手法

提案手法における処理の流れを**図 1** に示す．まず，検索対象となる音声ドキュメントの音響特徴ベクトルをクラスタリングによってベクトル量子化（VQ）し，音声ドキュメントを VQ コード列に変換する．音響的な特徴の広がりは話者によって異なるため，クラスタリングは話者ごとに行う．したがって，音声ドキュメントの話者が既知であることが前提となる．同時に，音声ドキュメントを大語彙連続

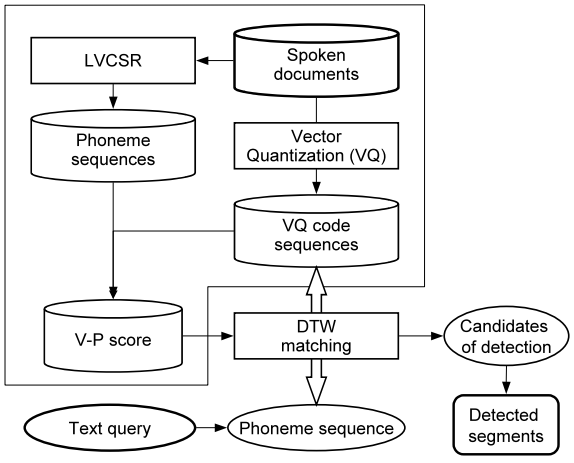


図 1 提案手法の処理手順
Fig. 1 Block diagram of the proposed method.

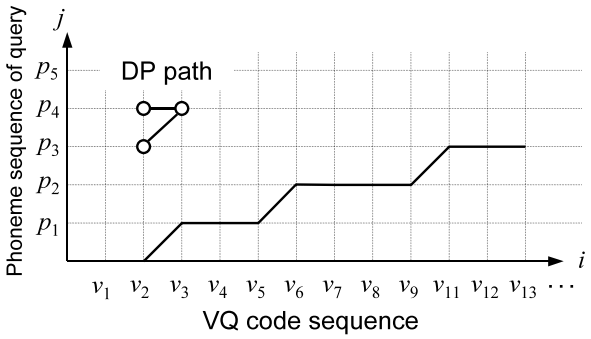


図 2 連続 DP マッチングの例
Fig. 2 A sample of DTW matching.

音声認識によって単語列に変換し，さらに音素列に変換する．話者依存の VQ コードブックごとに，VQ コードにおける音素の出現頻度に基づいて VQ コードと音素の関連度を V-P スコア（VQ-Phoneme score）としてあらかじめ学習しておく．テキスト入力された検索語を音素列に展開し，V-P スコアを局所スコアとする連続 DP マッチング [18] によって音素列と VQ コード列との照合を行う．この連続 DP マッチングの例を**図 2** に示す．図 2 に示した DP パスを用い，フレームごとの局所スコアとして 2.3 節で述べる V-P スコアを用いる．検索語の音素列を p_j（$1 \leq j \leq K$）（K は検索語中の音素数），音声ドキュメントの VQ コード列を v_i（$1 \leq i \leq L$）（L は音声ドキュメントのフレーム数），v_i と p_j 間の V-P スコアを $s(v_i, p_j)$ とするとき，i 番目のフレームにおける最大累積スコア $S_{i,K}$ は，

1) $S_{0,j} = 0 \quad (0 \leq j \leq K)$
2) $i = 1, 2, \cdots, L$ に対して 3)，4) を実行
3) $S_{i,0} = 0$
4) $$S_{i,j} = \begin{cases} S_{i-1,j-1} + s(v_i, p_j) & (\bar{S}_{i-1,j-1} > \bar{S}_{i-1,j}) \\ S_{i-1,j} + s(v_i, p_j) & (\bar{S}_{i-1,j-1} \leq \bar{S}_{i-1,j}) \end{cases}$$

によって算出する．ここで $\bar{S}_{i,j}$ は照合区間長で正規化した累積スコアで，i 番目のフレームと j 番目の音素が照合したときのその照合開始フレームを $start(i,j)$ とすると

$$\bar{S}_{i,j} = \frac{1}{i - start(i,j) + 1} S_{start(i,j),j} \tag{1}$$

で与えられる．

検索語の音素列と照合した区間およびスコアを連続 DP マッチングによってフレームごとに求め，照合区間における正規化スコアを $\bar{S}(i)$（$= \bar{S}_{i,K}$）とする．極大値を示した $\bar{S}(i)$ に対応した区間を検索語候補区間とし，2.4 節で述べる候補区間の再評価により，最終的に検索語の検出を決定する．

2.2 音響特徴のベクトル量子化

音声ドキュメントをフレーム長 20 ms，シフト間隔 10 ms で分析し，12 次元の MFCC（Mel Frequency Cepstral Coefficient）パラメータを算出し，各フレームの特徴ベクトルとする．連続した数フレームの特徴ベクトルに対してベクトル量子化を行う，いわゆるセグメント量子化によって VQ コードブックを作成する．当該フレームの前後 2 フレーム，計 5 フレームの特徴ベクトルを連結した 60 次元の特徴ベクトルを用いる．予備実験により，12 次元の MFCC と 12 次元の △MFCC を合わせた 24 次元の特徴ベクトルよりも，セグメント量子化による 60 次元の特徴の方が性能が高くなることを確認している．ベクトル量子化の手法としては k-means 法に基づく LBG 法 [19] を用いた．

セグメント量子化は話者ごとに行い，音声ドキュメントを自話者のコードブックに基づいて量子化を行い，VQ コード列へ変換する．

2.3 V-P スコアの算出

VQ コード列で表現された音声ドキュメントとテキスト入力される検索語との照合を行うために，各 VQ コードと音素の関連度（V-P スコア）を話者ごとにあらかじめ学習しておく．検索対象となる音声ドキュメントを大語彙連続音声認識によって音素列に変換し，各フレームにおいて VQ コードと音素の対を求める．各 VQ コードにおける音素の出現頻度に基づいて，VQ コード v における音素 p に対する V-P スコア $s(v,p)$ を

$$\begin{aligned} s(v,p) &= \log\left(\frac{C_v(p)}{N_v}\right) - \log\left(\frac{C_v(p_{most})}{N_v}\right) \\ &= \log\left(\frac{C_v(p)}{C_v(p_{most})}\right) \end{aligned} \tag{2}$$

で定義する．ここで，$C_v(p)$ は VQ コード v に量子化されたフレームのうち音素 p に対応しているフレーム数，VQ コード v に量子化されたフレームのうち最も多く対応した音素を p_{most} とし，その対応フレームの数を $C_v(p_{most})$ とする．N_v は VQ コード v に量子化された総フレーム数である．$C_v(p)/N_v, C_v(p_{most})/N_v$ は，VQ コード v における音素の出現確率を表しており，$s(v,p)$ は最頻出音素の確率で正規化した対数尤度となっている．

2.4 検出候補の再評価

検索語の検出は，2.1 節で述べた正規化スコア $\bar{S}(i)$ に対する閾値処理だけでも可能であるが，実際に予備的実験で検出を行ってみると多くの湧き出し誤りが発生した．湧き出し誤りの傾向として，

(1) 検索語の 1 つの音素に対応するフレーム数が極端に小さい．

(2) 検索語中の少数の音素が照合区間の大部分を占める．

などが見られた．

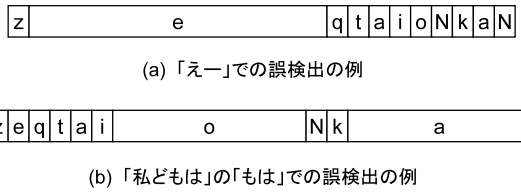


図 3 「絶対音感」に対する誤検出例
Fig. 3 Samples of false alarm for detecting /zeqtaioNkaN/.

図 3 は，検索語「絶対音感」に対する上記 (2) の誤検出例を模式的に示している．(a) では，発話内容「えー」の区間が，「絶対音感」の /e/ に対応し，/e/ の前後の音素列，すなわち /z/ および /qtaioNkaN/ がそれぞれ「えー」の前後の非常に短い区間と対応している．この例では，/z/ および /qtaioNkaN/ における V-P スコアは，非常に小さい値となるものの，そのフレーム数は少ない．一方，/e/ と対応した区間は V-P スコアが大きくフレーム数も多いため，正規化スコアでは比較的大きい値となり，誤検出となってしまう．(b) の例においても同様に，検索語中の /o/ および /a/ が，発話内容「私どもは」における「も」「は」の母音部と大きい V-P スコアで対応することによって誤検出となる．

本手法における VQ コード列と音素列との照合では，検索語における 1 つの音素が複数の VQ コード（部分 VQ コード列）と照合することとなる．一般に，母音は長い部分 VQ コード列と，/r/ などの子音は短い部分 VQ コード列と照合するなど，1 つの音素が照合する VQ コードの数は大きく変化し，発話速度によってもさらに変化する．このような照合の特性から，DP パスの制御によって不自然な時間長を持つ候補区間を排除することは容易ではない．そこで，不自然な時間長を持つ区間を検出候補から排除するために，以下の処理によって候補区間を再評価し，最終的な検出区間を決定する．

2.4.1 検出区間フレーム長に関する条件

検出候補区間の総フレーム数に関して，フレーム数が極端に短い検出区間は候補から削除する．大語彙連続音声認識結果から得られる各音素の平均フレーム長を話者ごとに求め，予備実験の結果から，検索語を構成する音素の平均フレーム長の和の 0.48 倍を閾値とした．

2.4.2 音素列の時間構造に関する評価

検出候補区間内の各音素のフレーム数に関して，予測されるフレーム数との差を算出し，2.4.3 項で述べる最終的な検出を決定する評価値に組み入れる．フレーム数の差の平均値 $V_D(i)$ を

$$V_D(i) = \frac{1}{K} \cdot \sum_{j=1}^{K} \left(\frac{D_l(p_j)}{L_l} - \frac{D_d(p_j)}{L_d} \right)^2 \tag{3}$$

によって定義する．$D_l(p)$ は音素 p の予測される平均フレーム数，$D_d(p)$ は音素 p の候補区間におけるフレーム数である．また，L_l は検索語音素列の予測される総フレーム数で，

$$L_l = \sum_{j=1}^{K} D_l(p_j) \tag{4}$$

で与えられ，L_d は候補区間の総フレーム数で

$$L_d = \sum_{j=1}^{K} D_d(p_j) \tag{5}$$

で与えられる．ここで，音声認識結果から音素境界を決定し，音素区間内のフレーム数を求める．音声ドキュメントごとに音素の平均フレーム数を求め，各音素に対して予測されるフレーム数とした．

2.4.3 再評価スコアの算出

2.4.1 項で述べた条件を適用した後，正規化スコアと音素の時間構造に関する評価に基づいた再評価スコアを算出し，その値に対する閾値処理によって最終的な検出結果を得る．2.1 節で述べた正規化スコア $\bar{S}(i)$ を平均と標準偏差でさらに正規化した値を $P_{\bar{S}}(i)$ とする．すなわち，

$$P_{\bar{S}}(i) = \frac{\bar{S}(i) - \mu_{\bar{S}}}{\sigma_{\bar{S}}} \tag{6}$$

$$\mu_{\bar{S}} = \frac{1}{N} \sum_{i=1}^{N} \bar{S}(i) \tag{7}$$

$$\sigma_{\bar{S}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\bar{S}(i) - \mu_{\bar{S}} \right)^2} \tag{8}$$

とする．ここで，N は検索語候補区間の総数である．同様に，2.4.2 項で述べた音素の時間構造の評価値 $V_D(i)$ を平均と標準偏差で正規化した値を $P_{V_D}(i)$ とする．これらの値から算出した再評価スコア $P(i)$ を

$$P(i) = P_{\bar{S}}(i) - P_{V_D}(i) \tag{9}$$

で定義する．

2.5 検出結果の統合

一般に STD において複数の認識結果を用いる手法が有効であることが知られており，NTCIR10-SpokenDoc2 においても，複数の認識結果を利用したシステムの方が単独の認識結果を用いたシステムに比べて結果が良好であったことが報告されている [22]．そこで，異なる認識結果から何通りかの V-P スコアを学習し，各 V-P スコアを用いて照合した複数の結果を用いることで精度の改善を図る．本論文での検索語検出では，ポーズで分割された発話を単位として，各発話において検索語が発話されている程度を照合スコアとして算出し，発話中に検索語が含まれているか否かを判定する．STD において複数の認識結果を用いる手法が有効であるのは，複数の音声認識システムを使って多様な音声認識結果を生成し，そのどれかと類似していれ

ば検出する処理がうまくいっていると考えられる．そこで本論文では，同一発話に対する複数の照合スコアのうち最大のスコアを採用することで統合を行う．i 番目の発話の統合後のスコア $P_{new}(i)$ は

$$P_{new}(i) = max(P_1(i), P_2(i), \cdots, P_M(i)) \qquad (10)$$

と定義される．ここで $P_j(i)$ は j 番目の照合結果の i 番目の発話に対するスコアを表しており，M は用いる照合結果の数である．

3. 評価実験

3.1 実験条件

検索対象となる音声ドキュメントとして，CSJ 日本語話し言葉コーパスのコア講演（177 講演）を用いて評価実験を行った．CSJ のデータは，200 ms のポーズによって IPU（Inter Pausal Unit）と呼ばれる単位に分割されている．本評価実験では，IPU を発話と考え，検索語を含んでいる発話を正しく検出できたか否かで検出の正否を判定する．検索語としては，音声ドキュメント処理 WG が STD 評価として選択している未知語セット 50 語を用いた [21]．

V-P スコアの学習には，NTCIR9 SpokenDoc タスクオーガナイザから提供されている音声認識結果を利用した．この音声認識結果には，10-best の連続単語認識と 10-best の連続音節認識の結果が含まれており，いずれも音素列に変換して V-P スコアの学習に用いた．音声認識は以下の条件で行われている．

- 学習データは CSJ 講演音声を用いる．
- 単語ベースの認識に用いる辞書は CSJ の人手書き起こしテキストを Chasen with UniDic-1.3.9 によって定義された形態素によって形態素解析して得られる約 27,000 語である．
- 音節ベースの認識には日本語全音節を用いる．
- CSJ 講演音声には固有の ID 番号が付与されており，下 1 桁の番号が偶数か奇数かによって偶数セット，奇数セットに分割する．
- 偶数セット，奇数セットそれぞれで triphone 音響モデル，単語 3-gram 言語モデル，音節 3-gram 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で学習したモデル，奇数セットの音声認識は偶数セットで学習したモデルを用いる．
- 音声認識エンジンは Julius である．

利用した音声認識結果における音素認識率を**表 1** にまとめておく．(a) が連続音節認識，(b) が連続単語認識での結果である．10-best での認識だけでなく，1-best だけを用いた場合の認識率もあわせて示してる．10-best の評価では，正解音素列とアラインメントをとったうえで，各音素区間において，10 個の結果に 1 つでも正解が含まれていれば正解と判断した．ここで，音素正解率は正解音素数を総音素数で割った値であり，音素正解精度は正解音素数から挿入誤り音素数を引いた数を総音素数で割った値である．連続音節認識，連続単語認識のどちらにおいても，10-best の結果を用いることで 1-best よりも音素の正解率が 5%程度向上することが分かる．

表 1 音声認識結果における音素認識率
Table 1 Phoneme Recognition Score of Speech Recognition Results.

(a) 連続音節認識

データ	音素正解率（%）	音素正解精度（%）
1-best	87.8	81.5
10-best	90.7	86.6

(b) 連続単語認識

データ	音素正解率（%）	音素正解精度（%）
1-best	90.5	85.8
10-best	93.3	90.5

評価は再現率，適合率，F 値，平均的な適合率を与える MAP（Mean Average Precision）[7] で行う．検索語ごとに性能を評価し，それらを平均することで評価を行った．

従来手法としては，音声ドキュメントに対して大語彙連続音声認識を行い，1-best の認識結果を音素列に展開して検索語の音素列と照合する手法を用いる．照合には編集距離（edit distance）を局所距離とした連続 DP マッチングを用いた．各音声ドキュメントの話者が既知でれば，話者適応によって音響モデルを適応することもできる．しかし一般に STD では，検索対象となる音声ドキュメントの書き起こしテキストは得られない．話者が既知の場合でも，音声認識結果を使った音響モデルの話者適応により認性能改を善することは容易ではないと考え，本論文では，音響モデルとしては話者非依存のモデルを用いている．

3.2 コードブックサイズの違いによる性能評価

ベクトル量子化におけるコードブックサイズを 1,024，2,048，4,096 と変えて，STD 性能の比較を行った．どの話者に対しても同じコードブックサイズを用いている．V-P スコアの学習に用いる音声認識結果は，音節 1-best の結果である．コードブックサイズを変えたときの再現率–適合率曲線を**図 4** に，最大 F 値，MAP を**表 2** に示す．コードブックサイズが 2,048 のとき，最も高い性能を示しており，従来手法に比べ最大 F 値で 3.9% MAP で 18.4%の性能が改善されていることが分かる．検定を行ったところ，F 値の改善は有意差はない一方で，MAP の改善は 1%の危険率で有意差が示された．

図 5 は 2.4 節で述べた再評価を行った場合と行わなかった場合の再現率–適合率曲線を比較している．コードブックサイズは 2,048 である．再評価を行わないと，多数の湧き出し誤りが発生し，性能が著しく劣化している．スコア

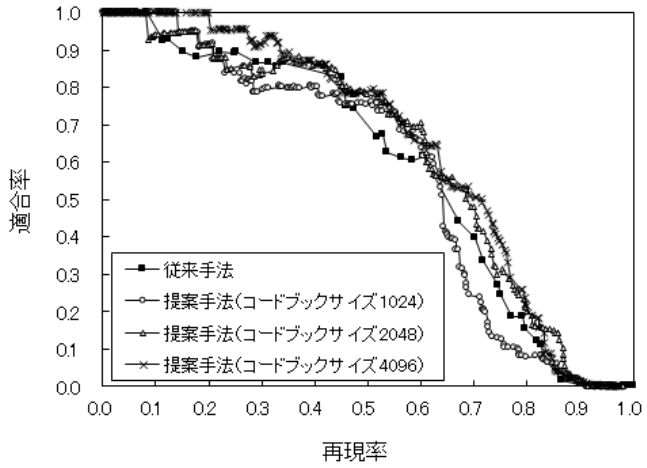


図 4 コードブックサイズを変えたときの再現率–適合率曲線
Fig. 4 Precision-Recall curve for various codebook sizes.

表 2 コードブックサイズを変えたときの最大 F 値と MAP
Table 2 Max F-measure and MAP for various codebook sizes.

手法	コードブックサイズ	最大 F 値 (%)	MAP (%)
従来手法	−	60.9	50.0
提案手法	1,024	63.1	63.0
	2,048	64.8	68.4
	4,096	63.8	66.9

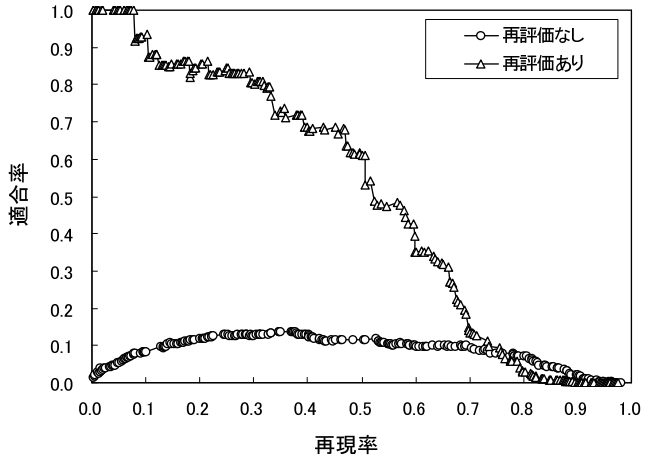


図 5 再評価の有無による検出性能の比較
Fig. 5 Comparison of STD performance with and without rescoring.

に対する閾値をゆるくすると再現率は上がるものの，適合率は低い．閾値を厳しく設定した場合でも適合率が非常に低いことから，スコア上位の候補区間がほとんど湧き出し誤りで占められていることが分かる．時間構造の情報を用いて再評価を行うことにより，湧き出し誤りをうまく抑制できていることが分かる．一方で，スコアに対する閾値をゆるくし再現率が高くなった場合を比較すると，再評価を行わない場合の方が適合率がやや高くなっており，再評価によって一部の正解区間が排除されていることがうかがえる．

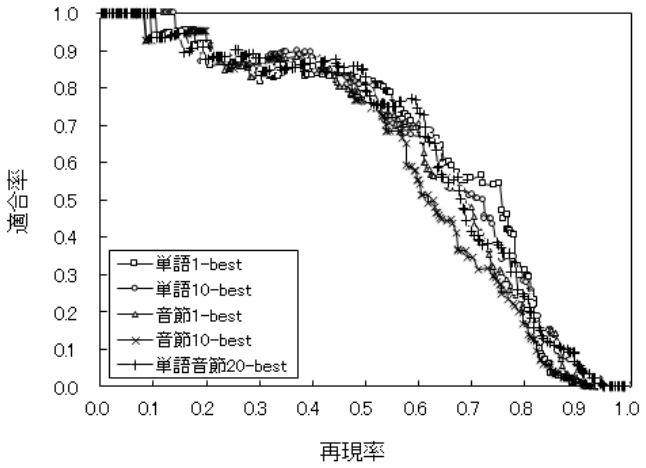


図 6 様々な V-P スコア学習による再現率–適合率曲線
Fig. 6 Precision-Recall curve for various V-P score training.

表 3 様々な V-P スコア学習による最大 F 値と MAP
Table 3 Max F-measure and MAP for various V-P score training.

	最大 F 値（%）	MAP（%）
単語 1-best	65.0	69.2
単語 10-best	65.4	69.1
音節 1-best	64.8	68.4
音節 10-best	62.0	63.7
単語音節 20-best	67.1	68.0

3.3 V-P スコア学習に用いる音声認識結果の違いによる性能評価

V-P スコアの学習に用いる音声認識結果を変えた場合の STD 性能の比較を行った．比較に用いた音声認識結果は，単語 1-best，単語 10-best，音節 1-best，音節 10-best，単語 10-best と音節 10-best の両者を用いる単語音節 20-best の 5 通りである．コードブックサイズは 2,048 である．

再現率–適合率曲線を**図 6** に，最大 F 値と MAP を**表 3** に示す．結果を比較すると，音節認識よりも単語認識の方が性能がやや高い．これは，表 1 から分かるとおり，音素認識の性能が単語認識の方が良いため，安定した V-P スコアの学習が行われていると思われる．音節認識と単語認識の結果を併用すると，音素正解率もさらに高くなると考えられ，STD 性能も向上している．一方で，1-best を 10-best にしても性能が改善するわけではなく，音節 10-best では性能がかなり劣化してる．どれか 1 つが正解であれば正解と判断する音素正解率でみると 10-best の方が性能が高い反面，10-best に含まれる認識誤りも増加し，それらが V-P スコアの学習にも使われてしまうため，必ずしも STD 性能の向上にはつながっていない．

3.4 複数の検出結果の統合

3.3 節で比較した 5 通りの検出結果のうちから，いくつかの結果を統合して STD 性能の比較を行った．統合の手

表 4 複数の検出結果を統合したときの最大 F 値と MAP
Table 4 Max F-measure and MAP using fusion of several detection scores.

統合に用いる検出結果					最大 F 値（%）	MAP（%）
単語 1-best	単語 10-best	音節 1-best	音節 10-best	単語音節 20-best		
○	○	○	○		70.5(5.1)	72.5(3.3)
○	○	○		○	70.3(3.2)	73.6(4.4)
○	○		○	○	69.1(2.0)	73.2(4.0)
○		○	○	○	69.7(2.6)	72.8(3.6)
	○	○	○	○	69.9(2.8)	72.8(3.5)
○	○	○	○	○	70.6(3.5)	73.3(4.1)

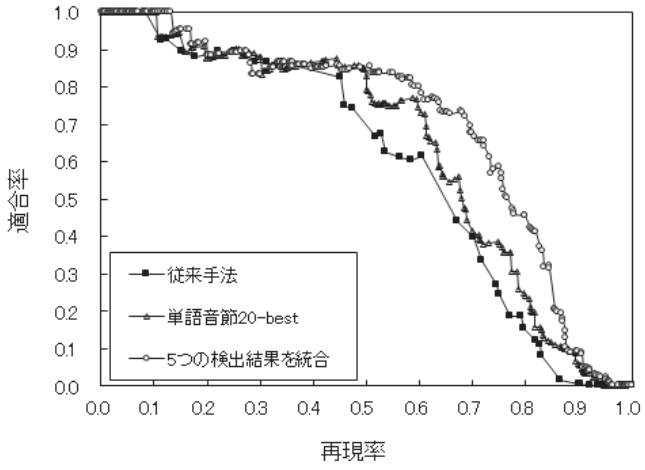


図 7 複数の検出結果を統合したときの再現率–適合率曲線
Fig. 7 Precision-Recall curve using fusion of several detection scores.

法は，2.5 節で述べたとおりである．最大 F 値と MAP の比較結果を**表 4** に示す．かっこ内の数字は，各統合に用いた検出結果単独での STD 性能と比較して何%性能が向上したかを表している．最上段の結果は，単語 1-best，単語 10-best，音節 1-best，音節 10-best の 4 通りの検出結果を統合しており，F 値では，この 4 通りの中で最も値が高かった単語 10-best 単独での F 値 65.4%よりも 5.1%改善されたことを示している．この表より，いずれの場合も性能が向上しており，最も良い最大 F 値を示したのは 5 つすべての検出結果を統合した結果であった．表 3 と表 4 の比較で示される統合の効果を検定してみると有意差は得られなかったが，F 値に関しては表 2 の従来方法と比較することで 5%の危険率で有意差があり，MAP に関しては 1%の危険率で有意差があった．

従来手法と，統合なしで最も高い F 値を示した単語音節 20-best の認識結果を用いて学習した V-P スコアでの結果と，5 つの検出結果を統合した結果の再現率–適合率曲線を**図 7** 示す．統合を行うことで性能が向上しており，検出結果の統合が STD 性能の改善に有効であることが分かる．特に適合率が低いときに再現率が向上している傾向にある．一般に，適合率が低いのは，多数の区間が検出されるよう検出のしきい値を下げた場合である．このとき，複数の検出を行っているとその中のどれかに正しい検出結果が含まれることが多くなり，再現率が向上することとなる．

3.5 処理時間

本手法では，1 時間の音声ドキュメントに対して 1 つの検索語を検索するのに約 20 秒の処理時間がかかっている．これは，音声ドキュメントを音素列で表現する場合と比較して約 13 倍の処理時間となっている．このため，大規模な音声ドキュメントに対する網羅的な検索語検出では，実用的な時間では処理できない．事前のインデクス作成などによる高速化が必要となる．あるいは，他の高速な方法で検出された候補区間の再評価では処理時間はあまり大きな問題にならないため，本手法が適用できると思われる．一方で，STD の用途としては，特定の音声ドキュメントに対して，検索語を検索したい場合や，内容検索のための事前のインデクス作成など，リアルタイムでの処理が求められない応用もあると考えられ，このような用途には実用的な手法になりうる．

4. おわりに

検索対象の音声ドキュメントの表現手法として音響情報をベクトル量子化して得られる VQ コード列を用い，あらかじめ学習しておく VQ コード列と音素の関連度（V-P スコア）に基づいて，テキスト入力された検索語との照合を行う手法を提案した．本手法では，話者ごとに VQ コードブックを作成し，各 VQ コードと音素の関連度を学習しておく必要がある．これを行うには，

(1) 検索対象の音声ドキュメントにおける話者情報が既知である，

(2) 同じ話者に対してまとまった量の音声ドキュメントが利用可能である，

(3) 大語彙連続音声認識が高い精度で可能である，

ことが前提となる．録音条件の良い講演音声や Web 上でアクセスできる音声ブログなどでは，これらの条件がある

程度満たされていることが多いと考えられる．提案手法では，音声認識の精度が高いほどSTDの精度も高くなる傾向があり，書き起こしテキストを使ってV-Pスコアを学習すればSTDの精度が向上することも確認している．

評価実験により，音素列間の照合を行う従来手法に比べて，提案手法の検出性能が高いことを示した．また，検出結果の統合を行い，さらに精度が改善されることを示した．検出結果の統合では，スコアの最大値をとることによって統合したが，統合したスコアの算出方法は今後の課題として残される [13]．本論文では，検索語がテキストで与えられた場合に，連続音声認識システムで未知語となる検索語に対する検出性能の改善に焦点を当てた．今後，処理時間の短縮や音声入力された検索語への対応などが課題としてあげられる．また，音素以外のサブワードの利用VQコードとの関連性の算出方法を検討し，さらなる性能向上も目指したい．

謝辞 本研究を行うにあたり工藤祐介氏に協力いただいた．また，「日本語話し言葉コーパス」および「NRCIR-9 SpokenDocタスクのCSJ Spoken Document Retrieval collection」を使用した．

参考文献

[1] Fiscus, J.G., Ajot, J., Garofolo, J.S. and Doddingtion, G.: Results of the 2006 Spoken Term Detection Evaluation, *Proc. 2007 Special Interest Group on Information Retrieval* (*SIGIR-07*) *Workshop in Searching Spontaneous Conversational Speech*, pp.51–57 (2007).

[2] 伊藤慶明，堀 貴明：音声認識の応用システム，日本音響学会誌，Vol.66, No.1, pp.36–40 (2010).

[3] 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, pp.1–8 (2010).

[4] 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.J86-D-II, No.10, pp.1369–1381 (2003).

[5] Yu, P. and Seide, F.: A Hybrid Word/Phoneme-Based Approach for Improved Vocabulary-Independent Search in Spontaneous Speech, *Proc. INTERSPEECH*, pp.293–296 (2004).

[6] 岩田耕平，伊藤慶明，小嶋和徳，石亀昌明，田中和世，李 時旭：語彙フリー音声文書検索手法における新しいサブワードモデルとサブワード音響距離の有効性の検証，情報処理学会論文誌，Vol.48, No.5, pp.1990–2000 (2007).

[7] 小野寺悠二，伊藤慶明，小嶋和徳，石亀昌明，田中和世，李 時旭：複数サブワード・言語モデルを用いた音声中の検索語検出の高精度化，第4回音声ドキュメント処理ワークショップ講演論文集 (2010).

[8] 中川聖一，岩見圭祐，藤井慶寿，山本一公：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，第4回音声ドキュメント処理ワークショップ講演論文集 (2010).

[9] Nishizaki, H., Furuya, Y., Natori, S. and Sekiguchi, Y.: Spoken Term Detection Using Multiple Speech Recognizers' Outputs at NTCIR-9 SpokenDoc STD subtask, *Proc. 9th NTCIR Workshop Meeting*, pp.236–241 (2011).

[10] Saito, H., Nakano, T., Narumi, S., Chiba, T., Kon'no, K. and Itoh, Y.: An STD system for OOV query terms using various subword units, *Proc. 9th NTCIR Workshop Meeting*, pp.281–286 (2011).

[11] Bulyko, I., Herrero, J., Mihelich, C. and Kimball, O.: Subword Speech Recognition for Detection of Unseen Words, *Proc. INTERSPEECH*, pp.2446–2449 (2012).

[12] Akbacak, M., Burget, L., Wang, W. and Hout, J.: Rich System Combination for Keyword Spotting in Noisy and Acoustically Heterogeneous Audio Streams, *Proc. ICASSP*, pp.8267–8271 (2013).

[13] Mamou, J. et al.: System Combination and Score Normalization for Spoken Term Detection, *Proc. ICASSP*, pp.8272–8276 (2013).

[14] 神田直之，糸山克寿，奥乃 博：音声中の任意検索語検出のための未知語区間推定に基づく選択的インデックス統合法，情報処理学会論文誌，Vol.55, No.3, pp.1201–1211 (2014).

[15] Norouzian, Z. and Richard, R.: An Approach for Efficient Open Vocabulary Spoken Term Detection, *Speech Communication*, Vol.57, pp.50–62 (2014).

[16] 松永 徹，趙 國，山下洋一：音響情報の話者依存ベクトル量子化を用いた音声検索語検出，第5回音声ドキュメント処理ワークショップ講演論文集，No.SDPWS2011-07 (2011).

[17] Yamashita, Y., Matsunaga, T. and Cho, K.: YLAB@RU at Spoken Term Detection Task in NTCIR9-SpokenDoc, *Proc. 9th NTCIR Workshop Meeting*, pp.287–290 (2011).

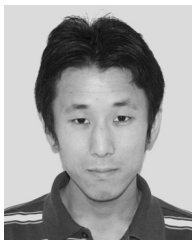
[18] 中川聖一：パターン情報処理，丸善 (1999).

[19] Linde, T., Buzo, A. and Gray, R.M.: An algorithm for vector quantizer design, *IEEE Trans. Commun.*, Vol.COM-28, No.1, pp.84–95 (1980).

[20] 福田 隆，新田恒雄：音声認識のための特徴パラメータ正準化法の検討（認識・理解・対話），情報処理学会研究報告，Vol.2004-SLP-51, No.5, pp.19–24 (2004).

[21] 西崎博光，胡 新輝，南條浩輝，伊藤慶明，秋葉友良，河原達也，中川聖一，松井知子，山下洋一，相川清明：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, pp.1–8 (2011).

[22] 西崎博光，秋葉友良，相川清明，伊藤慶明，河原達也，胡 新輝，中川聖一，南條浩輝，山下洋一：NTCIR-10 SpokenDoc-2 Spoken Term Detectionタスクの結果と知見，日本音響学会2013年秋季研究発表論文集，No.3-8-6, pp.107–110 (2013).

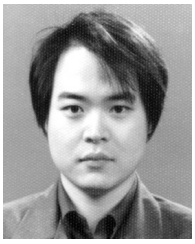


坂本 伊織

平成24年立命館大学情報理工学部メディア情報学科卒業．平成26年同大学大学院博士前期課程修了．現在，村田機械株式会社勤務．在学中は音声ドキュメント検索に関する研究に従事．

松永 徹

平成 22 年立命館大学情報理工学部メディア情報学科卒業．平成 24 年同大学大学院博士前期課程修了．現在，日本電気株式会社勤務．在学中は音声ドキュメント検索に関する研究に従事．



趙 國

平成 10 年立命館大学理工学部電気電子工学科卒業．平成 12 年同大学大学院博士前期課程修了．平成 15 年同大学院博士後期課程単位取得退学．平成 17 年同大学情報理工学部ポスドク研究員，平成 18 年同助手，平成 24 年同客員研究員．平成 26 年東亜大学助教授，現在に至る．博士（工学）．音声・音響情報処理に関する研究に従事．ISCA 会員．

山下 洋一 （正会員）

昭和 57 年大阪大学工学部電子工学科卒業．昭和 59 年同大学大学院前期課程修了．同年同大学産業科学研究所文部技官，平成 5 年同助手，平成 6 年同講師，平成 9 年立命館大学理工学部助教授，平成 13 年同教授，平成 16 年同大学情報理工学部教授，現在に至る．博士（工学）．音声情報処理に関する研究に従事．電子情報通信学会，日本音響学会，人工知能学会，ISCA，IEEE 各会員．