INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Combining state-level spotting and posterior-based acoustic match for improved query-by-example spoken term detection

Shuji Oishi[1], *Tatsuya Matsuba*[1], *Mitsuaki Makino*[1], *Atsuhiko Kai*[1]

[1]Graduate School of Integrated Science and Technology, Shizuoka University, Japan

oishi@spa.sys.eng.shizuoka.ac.jp, matsuba@spa.sys.eng.shizuoka.ac.jp,
makino@spa.sys.eng.shizuoka.ac.jp, kai@sys.eng.shizuoka.ac.jp

Abstract

In spoken term detection (STD) systems, automatic speech recognition (ASR) frontend is often employed for its reasonable accuracy and efficiency. However, out-of-vocabulary (OOV) problem at ASR stage has a great impact on the STD performance for spoken query. In this paper, we propose combining feature-based acoustic match which is often employed in the STD systems for low resource languages, along with the other ASR-derived features. First, automatic transcripts for spoken document and spoken query are decomposed into corresponding acoustic model state sequences and used for spotting plausible speech segments. Second, DTW-based acoustic match between the query and candidate segment is performed using the posterior features derived from a monophone-state DNN. Finally, an integrated score is obtained by a logistic regression model, which is trained with a large spoken document and automatically generated spoken queries as development data. The experimental results on NTCIR-12 SpokenQuery&Doc-2 task showed that the proposed method significantly outperforms the baseline systems which use the subword-level or state-level spotting alone. Also, our universal scoring model trained with a separate set of development data could achieve the best STD performance, and showed the effectiveness of additional ASR-derived features regarding the confidence measure and query length.

Index Terms: spoken term detection, spoken query, posteriorgram, acoustic similarity, score normalization

1. Introduction

Spoken term detection is a task which locates a given search term in a large set of spoken documents. Typically, automatic speech recognition (ASR) frontend is often employed for its STD performance in efficiency and accuracy. However, out-of-vocabulary (OOV) problem degrades recognition accuracy and affects the STD performance.

To deal with OOV problem, many approaches using a feature-based acoustic match have shown its effectiveness in low-resource STD tasks, as well as the robustness against the effects by difference of speaker and recording environments[1],[2],[3]. However, a feature-based approach is time-consuming and the approach alone couldn't outperform the STD performance of conventional ASR-based system for rich-resource language tasks. On the other hand, in the context of STD system for rich-resource language, related works using the approximate match between query and ASR-based automatic transcript with subword-level acoustic similarity have been proposed to deal with OOV problem. In [4], a syllable-level distance measure based on the Bhattacharyya distance derived from syllable-unit HMMs is used. In our previous works[5],[6], DTW-based spotting with a syllable-unit HMM state-level distance measure from word-level automatic transcript has shown significant improvement.

In this paper, we investigate the STD approaches which combine feature-based matching with the other ASR-derived features. An ASR acoustic model similarity based STD system is used as a baseline system. The proposed STD system is based on a two pass strategy. The first pass performs DTW-based spotting with ASR-derived acoustic dissimilarity which is syllable-unit HMM state-level distance measure. The second pass performs frame-level feature-based matching against candidate regions that is narrowed down by the first pass. We adopt posteriorgram feature derived from a DNN-based acoustic model, since previous studies have shown that it improves robustness for different speaker or recording environments[2][3].

We adopt two approaches for score-level system combination and for incorporating side information into scoring model. First, we integrate score with simple linear combination. Second, we integrate score with a logistic regression model which is trained with a large spoken document and automatically generated spoken queries. In contrast to conventional score fusion systems [7],[8] with multiple ASR systems, our regression model relies only on single ASR system and is trained along with the additional ASR-derived features regarding confidence measure and query length.

In this study, the experiments were conducted on the NTCIR-12 SpokenQuery&Doc-2 tasks[10] which target a spoken document collection: the Corpus of Spoken Document Processing Workshop (SDPWS). The experimental result shows that using an integrated score which is obtained by a logistic regression model from feature-based acoustic matching score and ASR-derived features attains the best STD performance. Also, the proposed approaches which combine feature-based acoustic matching and scoring with ASR-derived features are effective for OOV queries.

2. Baseline spoken term detection system

We compare with two baseline systems. The first baseline system (Baseline1) adopts a DTW-based spotting method which performs matching between subword sequences of query term and spoken documents and outputs matched segments. In NTCIR-9 SpokenDoc STD baseline system[9], a similar system with the local distance measure based on phoneme-unit edit distance is used. In our Baseline1 system, the local distance measure is defined by a syllable-unit acoustic dissimilarity as used in [4]. The distance between subwords x and y, $D_{sub}(x, y)$, is calculated by the DTW-based matching of two

http://dx.doi.org/10.21437/Interspeech.2016-1259

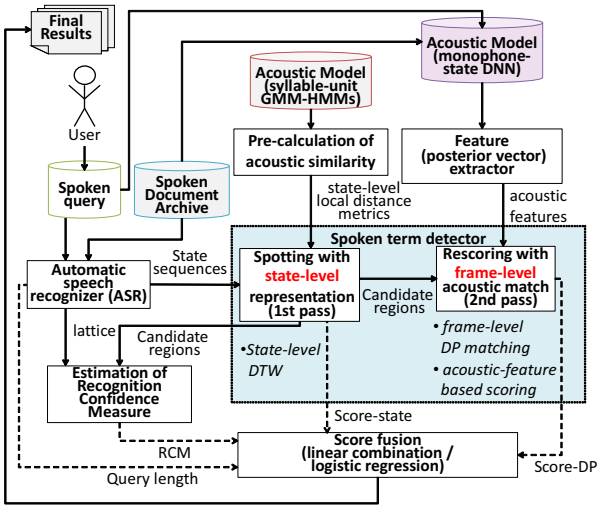

Figure 1: Overview of proposed STD system

subword HMMs with the local distance defined by the distance between two state's output distributions. We define the distance between two Gaussian mixture models P and Q as

$$D_{BD}(P,Q) = \min_{u,v} BD(P^{\{u\}}, Q^{\{v\}}) \quad (1)$$

where, $BD(P^{\{u\}}, Q^{\{v\}})$ denotes the Bhattacharyya distance between the u-th Gaussian component of P and the v-th Gaussian component of Q.

At the first stage of preprocessing, 1-best recognition results for a spoken document archive are obtained by an ASR system with a word N-gram language model. Then, the word-based recognition results are converted into syllable sequences by using pronunciation dictionary.

At the stage of STD for spoken query input, the query is first transcribed by ASR system and then decomposed into a syllable sequence. Next, a DTW-based word spotting is performed by using an acoustic dissimilarity as local distance measure[5],[6]. Finally, a set of segments with a dissimilarity score less than a threshold is obtained as the retrieval result.

Baseline1 adopts the spotting method based on syllable-unit local distance measure $D_{sub}(x,y)$. On the other hand, the second baseline system (Baseline2) adopts a DTW-based spotting method for the syllable-unit HMM state sequences, that is shorter than syllable unit, to elaborate spotting unit. The DTW-based spotting method for the syllable-unit HMM state sequences uses acoustic dissimilarity eq.(1) directly in the first pass, unlike subword-level local distance D_{sub} calculation in Baseline1 system or the use in the second pass only[6].

3. Proposed spoken term detection method

3.1. Proposed system overview

Overview of our proposed STD system is shown in Fig. 1. The system adopts two-pass strategy for both efficient processing and improved STD against recognition errors. The first pass performs DTW-based spotting method for the syllable-unit HMM state sequences as described in Section 2. The second pass performs frame-level acoustic matching against candidate regions those are narrowed down by the first pass. As shown in Fig. 1, we adopt DNN-based phonetic posterior feature which is described in Section 3.2. Finally, an integrated score is obtained by a logistic regression model as shown in Fig. 1 together with the additional ASR-derived features regarding the recognition confidence measure and query length. We have also compared an alternate method which is score fusion with a simple linear combination,

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

where the first pass score is $Score_{state}$ and the second pass score is $Score_{DP}$, respectively.

3.2. Posterior-based acoustic match

In DTW-based STD approaches, posteriorgram feature is often used as an acoustic feature vectors for calculating local distance. Acoustic model based on GMM or Multilayer perceptron (MLP) is used to transform the MFCC feature into phonetic posteriors[11],[12].

We use deep neural network (DNN) to model the distribution of acoustic feature in monophone HMM states and its output serves as a posterior feature vector. The local distance between two posterior vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as the negative log of the inner product :

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

The DNN is structured in 7 layers (an input layer, 5 hidden layers, and output layer). The number of units of output layer is 145. The DNN is trained on Restricted Boltzmann Machine (RBM) as pre-training and then discriminatively trained for monophone states by cross-entropy criterion.

The second pass performs frame-level DTW by using posterior vector-based local distance. One of major issues of the second pass DTW is the error of matched region caused by the first pass DTW detection error. We employ an end-point free DTW algorithm which allows extra matched regions ($+\beta$ frames in maximum) on both sides of hypothesized starting and ending points.

3.3. Integrated score by logistic regression model

There are some difficulties in deciding an optimized threshold, because $Score_{DP}$ and $Score_{state}$, which influence final decision, are easy to change by ASR condition such as mismatch between linguistic/acoustic model and target document. Therefore, we adopt an approach which learns scoring model which discriminates positive and negative detections from a lot of samples and expect that the approach gives a normalized score.

We use a large spoken document and automatically generated spoken queries as development set in advance to train the logistic regression model. The model is supervised learned by using matching scores derived from true and false candidates, which are detected by the first-pass spotting process, together with some ASR-derived features as described below. Spoken queries are extracted from a large spoken documents which is employed for learning acoustic model of ASR. We automatically extract spoken queries by using results of alignment which is employed for learning acoustic model. Spoken queries are all noun phrases and those which only appear once in one lecture are excluded.

We estimate the probability of correct detection by using the trained logistic regression model. As a cue that affects the variability in matching scores of correct detection, we focus on recognition confidence measure (RCM) and query length. We introduce them to the logistic regression model as additional ASR-derived features to estimate an integrated score by logistic regression.

Table 1: Example of binary features for query length
(If the number of morae in a query term is L_k and less, the k-th feature represents 1, otherwise 0)

Query term	Length (#morae)	Binary feature L_k			
		4	6	8	10
Ni ho n ji n (Japanese)	5	0	1	1	1
Shi zu o ka da i ga ku (Shizuoka university)	8	0	0	1	1
A ri ga to u go za i ma su (Thank you)	10	0	0	0	1

We estimate the recognition confidence measure (RCM) by calculating the average of posterior probability of the highest likelihood recognition in candidate region detected by the first pass. The RCM score is assumed to combine with $Score_{state}$ obtained by DTW-based spotting.

Given a syllable sequence of candidate region $B = \{B_1, \cdots, B_Y\}$, the confidence of candidate region is defined as,

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

where, $X = \{X_1, \cdots, X_Y\}$ is speech segment corresponding to the B, and $P(B_k|X_k)$ denotes the posterior probability of syllable B_k on 1-best recognition result in lattice. For simplicity, we approximate the calculation of RCM by

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

where, $T = \{i, \cdots, j\}$ is the speech frames corresponding to the B, and $P(s|t)$ denotes the posterior probability of phone s at frame t in lattice.

Another feature we introduced is query length in terms of the number of morae included in the corresponding automatic transcript. Mora is a linguistic unit in Japanese language and often used as a convenient way to describe the length. We introduce binary features for representing query length. As illustrated in Table 1, if the number of query's morae is a specific number of morae (L_k) such as 4, 6, 8, 10 morae and less, then the feature represents 1, otherwise, the feature represents 0.

4. Experiments

4.1. Experimental setup

In the evaluation experiment of STD by spoken queries (SQ-STD task evaluation set), we have verified the robustness of the proposal method by using a target document collection used in the NTCIR-12 SpokenQuery&Doc-2 tasks: the lecture of Spoken Document Processing Workshop (SDPWS, 107 lectures, about 29 hours). As with NTCIR-12 SpokenQuery&Doc-2 STD evaluation, the Inter-Pausal Units (IPU) are used as the basic units to be searched and the retrieval result of the IPU is regarded as correct if it includes the query term. In NTCIR-12 SpokenQuery&Doc-2 task, some queries are composed of two or more kinds of terms. To accommodate such queries, we split the query into terms by using the automatic transcript of spoken query and performed a search for each term. Finally, we have obtained search result of the query composed of some terms by using each term's search result. Regarding how to split the query, we have split them by the pauses which are no shorter than 200 msec. For the evaluation of SQ-STD task, we use 162 query terms which are spoken by 10 speakers including 59 OOV query terms which were used for the formal-run in NTCIR-12 SpokenQuery&Doc-2 SQ-STD task. These reference automatic transcriptions of the evaluation set were provided from NTCIR-12 organizer. They provided reference automatic transcriptions by using DNN-HMM acoustic model which was trained with the Corpus of Spontaneous Japanese (CSJ, 950 lectures)[13]. They used KALDI toolkit[14] to train the acoustic model.

For training the logistic regression model described in Section 3.3, a part of the CSJ corpus was used as a development data. The development data was divided into two parts: a target spoken document set for generating examples of matched scores by running STD and the other document set for extracting examples of spoken queries. The Core set of the CSJ (CSJ-CORE, 177 lectures, about 44 hours) was selected as the target spoken document set. For generating positive and negative examples, we use 620 spoken query terms including 163 OOV query terms, which are selected by a commonly used tf-idf criterion from a manual text transcription and excluding terms which are either too long(more than 13 morae) or too short(less than 3 morae). The spoken query terms are automatically extracted as described in Section 3.3 by using a subset of CSJ corpus (910 lectures), which is also used for learning both of acoustic model for ASR and DNN-based feature extractor described in Section 3.2. The word accuracies for based automatic transcriptions were 81.3% for SDPWS and 74.0% for CSJ-CORE, respectively.

We have trained monophone-state DNN which are used for posterior feature extraction by using the KALDI toolkit. The experiment uses 40 dimensional features that applied Linear Discriminant Analysis(LDA) to 39 dimensional MFCC features (MFCC+power+ΔMFCC+Δpower+ΔΔ MFCC+ ΔΔ power) with speaker level Cepstral Mean and Variance Normalization(CMVN), and we use 40 dimensional $\times$ 11frames as input features to DNN. The posterior feature vector has 145 dimensional features which correspond to the number of monophone states.

As a measure of search performance, we use F-measure(max) and MAP. F-measure(max) is the maximum value of F-measure when the threshold is adjusted. MAP is the Mean of Average Precision of all queries per query.

4.2. SQ-STD task result

Table 2 compares the STD performance of two baseline systems (first-pass only) and the proposed systems which consist of two-pass detection and verification steps with different combining methods for SDPWS target document collection. Baseline1 and Baseline2 are described in Section 2. State_spot+post represents the system based on the combined score by using linear combination described in Section 3. State_spot+RCM and State_spot+post+RCM+Mora represent the systems that combine the state_spot score with only RCM or 3 features (post, RCM and Mora) described in Section 3.3 by using logistic regression, respectively. A system denoted as "+Mora" adds a set of binary features on query length for score fusion as described in Section 3.3.

All queries consist of IV(in-vocabulary) queries and OOV queries. The parameters of the first-pass threshold are decided to separate the upper 1000 candidates per query. The weight coefficient α for the linear combination is determined by optimizing the F-max performance for development set. The pa-

Table 2: Formal-run(evaluation set) STD performance

querytype	system	F(max)	MAP
ALL	Baseline1(syll_spot)	38.32	64.42
	Baseline2(state_spot)	45.94	66.03
	state_spot+post(LC, α=0.6)	42.75	72.27
	state_spot+RCM(LR)	45.88	65.38
	state_spot+post+RCM+mora (LR)	**51.64**	**72.73**
IV	Baseline1(syll_spot)	45.34	73.14
	Baseline2(state_spot)	**59.46**	74.52
	state_spot+post(LC, α=0.6)	51.19	80.06
	state_spot+RCM(LR)	59.38	73.93
	state_spot+post+RCM+mora (LR)	58.83	**80.02**
OOV	Baseline1(syll_spot)	20.51	49.20
	Baseline2(state_spot)	19.71	51.20
	state_spot+post(LC, α=0.6)	22.25	58.66
	state_spot+RCM(LR)	19.55	50.46
	state_spot+post+RCM+mora (LR)	**31.79**	**60.01**

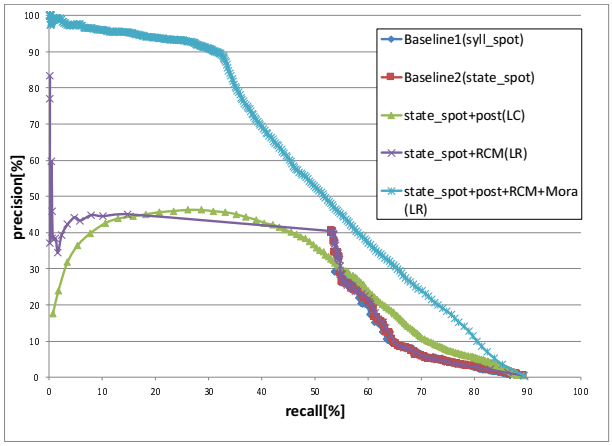

Figure 2: Recall-Precision curves of different STD systems (ALL queries)

rameter β which determines the maximum frame length of extra matched regions in the second-pass acoustic match was empirically set to 30.

The result shows that the proposed system outperforms the baseline systems which use only the first pass. In the proposed methods, especially, the second pass approaches which combine feature-based acoustic match improve the performance in MAP. On the other hand, the STD performance of state_spot+post and state_spot+RCM are lower than that of Baseline2(state_spot) in F-measure(F-MAX). However, state_spot+post+RCM+Mora which uses an integrated score that combined all knowledge information by using logistic regression model attained the best STD performance while improving the STD performance for ALL and OOV queries.

Also, STD performance is often measured in terms of receiver operating characteristic (ROC) curve. ROC curve is drawn by changing the detection threshold. ROC curves in Fig.3, 4 and 5 respectively show that the State_spot+post+RCM+Mora method attains the best STD performance. The proposed method significantly improved the precision in low-recall region.

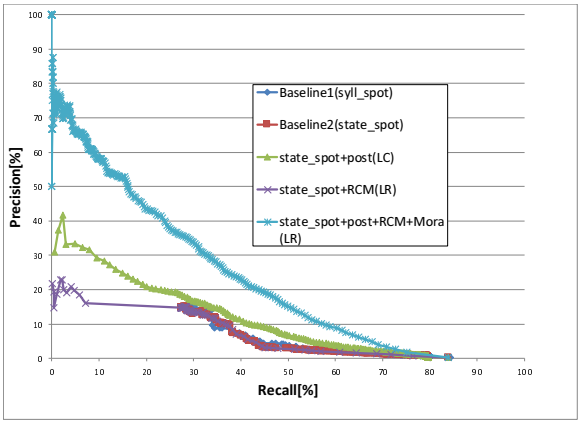

Figure 3: Recall-Precision curves of different STD systems (OOV queries)

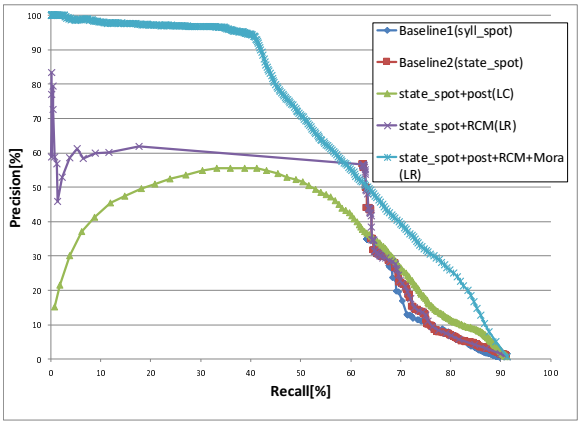

Figure 4: Recall-Precision curves of different STD systems (IV queries)

5. Conclusions

In this paper, we proposed combining feature-based acoustic matches with the other ASR-derived features to solve OOV problem and introduced a logistic regression modeling for incorporating simple ASR-derived features from single ASR system. The experimental results showed that combining a feature-level matching of posterior feature vectors with ASR frontend-based spotting improves the STD performance compared with baseline methods which use only spotting with subword-level or state-level local acoustic dissimilarity measure. In addition, using an integrated score which is obtained by a logistic regression model derived from feature-based acoustic matching score and ASR-derived features significantly improves the STD performance compared with baseline methods for all query without a significant decline in the STD performance of IV queries.

The recognition confidence measure (RCM) used as additional ASR-derived feature is based only on the ASR output of target document, and doesn't cope with the similarity between query and document. Therefore, we expect further improvement of the STD performance by using recognition confidence measure considering similarity with the query [6].

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 25330128.

7. References

[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, and Language Processing, Vol.23, (2015).

[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] M. Akbacak, L. Burget, W. Wang, and, J. van Hout : "Rich system combination for keyword spotting in noisy and acoustic heterogeneous audio streams," Proc. of ICASSP, (2013)

[8] J. van Hout, L. Ferrer, D. Vergyri, N. Scheffer, Y. Lei, V. Mitra, and S. Wegmann : "Calibration and multiple system fusion for spoken term detection using linear logistic regression," Proc. of ICASSP, (2014)

[9] T. Akiba, H. Nishiaki, K. Aikawa, T. Kawahara, and T. Matsui: "Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop," Proc. of NTCIR-9 Workshop Meeting, pp.223-235, (2011).

[10] National Institute of Informatics, http://research.nii.ac.jp/ntcir/ntcir-12/

[11] T. J. Hazen, W. Shen, and C. White: "Query-by-example spoken term detection using phonetic posteriorgram templates," Proc. ASRU, pp. 421-426, (2009).

[12] Y. Zhang and J. Glass : "Towards multi-speaker unsupervised speech pattern discovery," Proc. ICASSP, pp. 4366-4369, (2010).

[13] National Institute for Japanese Language : "Corpus of spontaneous Japanese: CSJ," http://www.ninjal.ac.jp/english/products/csj/, (2004).

[14] 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).