Combining State-level and DNN-based Acoustic Matches for Efficient Spoken Term Detection in NTCIR-12 SpokenQuery&Doc-2 Task

Shuji Oishi
Shizuoka University
3-5-1 Johoku, Hamamatsu-shi, Shizuoka
432-8561, Japan
oishi@spa.sys.eng.shizuoka.ac.jp

Tatsuya Matsuba
Shizuoka University
3-5-1 Johoku, Hamamatsu-shi, Shizuoka
432-8561, Japan
matsuba@spa.sys.eng.shizuoka.ac.jp

Mitsuaki Makino
Shizuoka University
3-5-1 Johoku, Hamamatsu-shi, Shizuoka
432-8561, Japan
makino@spa.sys.eng.shizuoka.ac.jp

Atsuhiko Kai
Shizuoka University
3-5-1 Johoku, Hamamatsu-shi, Shizuoka
432-8561, Japan
kai.atsuhiko@shizuoka.ac.jp

ABSTRACT

Recently, in spoken document retrieval task such as spoken term detection (STD), there has been increasing interest in using a spoken query. In 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 two spoken term detection methods which combine different types of side information for calculating integrated scores. Firstly, we propose combining feature-based acoustic match which is often employed in the STD systems for low resource languages, along with ASR-derived features. Secondly, we propose the method combining confidence measure of speech recognition with ASR. Both proposed methods consist of two-pass strategy. As the first pass, automatic transcripts for spoken document and spoken query are decomposed into corresponding acoustic model state sequences and used for spotting plausible speech segments. The experimental results showed that combination with feature-based acoustic match improves the STD performance compared to baseline system which uses the subword-level spotting alone.

Team Name

SHZU

Subtasks

SQ-STD

Keywords

spoken term detection, acoustic dissimilarity measure, posteriorgram, bottleneck feature, spoken query

1. INTRODUCTION

Spoken term detection (STD) is a task which locates a given search term in a large set of spoken documents. Typically, automatic speech recognition (ASR) is often employed as a frontend of STD system for its total performance in efficiency and accuracy. However, out-of-vocabulary (OOV) problem degrades recognition accuracy and affect the STD performance. Therefore, many approaches using a subword-unit based speech recognition system have been proposed[2, 3, 4, 5]. In our previous works[6, 7], DTW-based spotting and acoustic matches between feature representations derived from HMM's state-level alignments of automatic transcript has shown significant improvement. As another approach 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[8, 9, 10]. 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.

In this paper, we propose two STD methods which combine second-pass acoustic match scores derived from DNN-based feature extractor and side information derived from ASR output. Both approaches employ only one ASR system for providing a preliminary set of matching regions detected by a first-pass DTW-based spotting which is based on a state-level alignment between automatic transcripts of query and document. The first pass performs DTW-based spotting with ASR-derived acoustic dissimilarity which is syllable HMM state-level distance measure. The first proposed method employs DNN-based acoustic match as the second pass. The second pass performs frame-level acoustic match based on DNN-derived features such as bottle-neck feature and posteriorgram. The second proposed method integrates side information derived from ASR output for estimating final matching score. The second pass introduce a recognition confidence measure (RCM) in candidate regions by using lattices of spoken document. Finally, we obtain an integrated score with a logistic regression model which is trained with a development data.

The organization of this paper is as follows: our baseline system which is similar to the organizer's baseline of the NTCIR-12 SpokenQuery&Doc-2 task[1] is described in Section 2. Our proposed two-pass STD systems along with

the improved first-pass DTW-based spotter are described in Section 3. The experimental results in the NTCIR-12 SpokenQuery&Doc-2 task[1] are presented in Section 4. Finally, Section 5 gives a summary and conclusion of this work.

2. BASELINE SPOKEN TERM DETECTION SYSTEM

The baseline system 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[11], a similar system with the local distance measure based on phoneme-unit edit distance is used. In our baseline system, the local distance measure is defined by a syllable-unit acoustic dissimilarity as used in [12]. The distance between subwords x and y, $D_{sub}(x, y)$, is calculated by the DTW-based matching of two subword HMMs with the local distance defined by the distance between two HMM states. We define the distance between two states (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[6],[7]. Finally, a set of segments with a dissimilarity score less than a threshold is obtained as the retrieval result.

3. PROPOSED SPOKEN TERM DETECTION METHODS

3.1 State-level DTW for detecting candidate regions (First pass)

Overview of our proposed STD systems are shown in Fig. 1 and 2. This section describes a state-level DTW (spotting) part which is commonly used as the first pass of those extended STD systems described in the next sections. The state-level DTW spotter directly accepts the HMM state sequences of ASR outputs for spoken query and documents as input, while our baseline system described in Section 2 employs state-level DTW only for calculating the local distance between two syllables, which are then used for syllable-level DTW spotter. In other words, the state-level DTW spotter uses acoustic dissimilarity Eq.(1) directly, unlike D_{sub} calculation for syllable-unit local distance in Baseline system.

Next sections (3.2 and 3.3) describe our proposed methods which combine different types of side information for calculating integrated scores for reducing the detection errors in the first pass.

3.2 Combination with DNN-based acoustic matches (Second pass)

3.2.1 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 HMM state sequences as described in Section 3.1. 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 features which are described in Section 3.2.2 and 3.2.3. Finally, we obtain an integrated score with a simple linear combination,

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

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

The second pass performs frame-level DTW by using frame-level local distances defined in the next sections. One of major issues of the second pass DTW is that the error of matched region caused by the first pass also affects the reliability of second-pass score and degrades the STD performance. We employ an endpoint-free DTW algorithm which allows extra matched regions ($+\beta$ frames in maximum) on both sides of hypothesized starting and ending points detected in the first pass.

3.2.2 Bottleneck feature

As illustrated in Fig 2, bottleneck feature (BNF) is extracted from a multi-layer perceptron (MLP), in which one of hidden layers has a small number of units, relative to the size of other hidden layer. Conventionally, DNN with a bottleneck layer is trained, so as to extract a small amount of features effective to identify output class. When consecutive speech frames are spliced as the input to the DNN, it is expected that the output from a bottleneck layer works as a non-linear compressor for the high-dimensional input feature and effectively represents the output class while removing the redundancy in the input features.

In our proposed system, similarly to the general DNN training method, a Deep Belief Network (DBN) is constructed to initialize all weights in the DNN with a bottleneck layer. The DBN is then fine-tuned as a classifier for context-dependent clustered triphone states with back-propagation method. DNN structure for BNF extraction is shown in Fig 2. The 42 dimensional BNF extracted from a bottleneck layer is used as the frame-level feature for the second-pass DTW in our STD system. The local distance between two BNF vectors is defined as the Euclidean distance.

3.2.3 Posteriorgram feature

In DTW-based STD approaches, posteriorgram feature is often used as an acoustic feature vectors for calculating local distance. Acoustic models based on GMM or Multilayer perceptron (MLP) are used to transform a speech feature into phonetic posteriors[13],[14].

In our proposed system, DNN to extract posteriorgram is trained in the same way as DNN in Sec 3.2.2, while the class for output nodes are replaced by monophone HMM states to provide a compact representation of 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:

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

As illustrated in Fig 3, the number of units in output layer

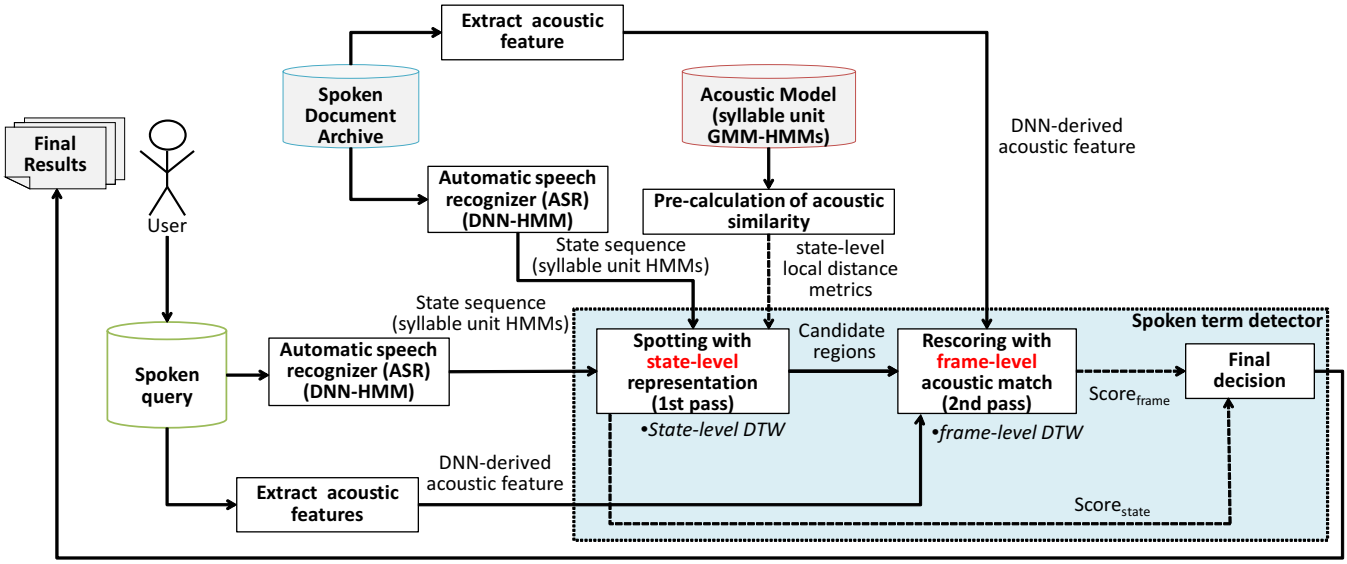

Figure 1: Overview of proposed STD system with the second-pass DNN-based acoustic matches

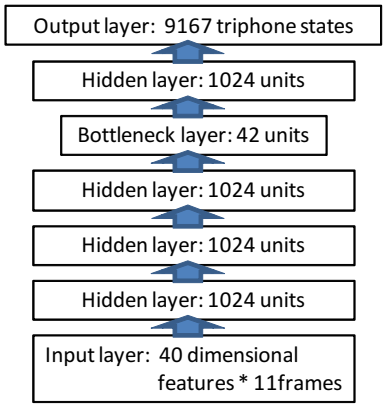

Figure 2: DNN structure to extract bottleneck feature

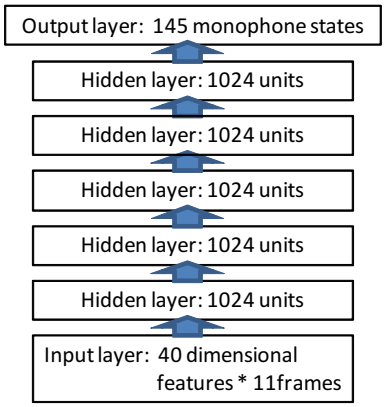

Figure 3: DNN structure to extract posteriorgram feature

(i.e., dimension of feature vector) is 145.

3.3 Combination with ASR confidence measure (Second pass)

Since our STD systems depend on ASR system as a frontend of spoken term detection, the speech recognition accuracy greatly affects the search performance. In our previous work[7], introducing a confidence measure that estimates the plausibility that a query term exists in a short speech segment has shown to improve the STD performance. However, previous study has used the confidence measure only for filtering the candidates of matched segments, rather than directly introducing to the final detection score. We can expect that the improvement of search performance is obtained by integrating side information by ASR output such as confidence measure into the final scoring.

3.3.1 *System overview*

Overview of our proposed combination with confidence measure STD system is shown in Fig. 4. The flow of the system is the following.

1. Perform the DTW-based spotting for the HMM state sequences and obtain a set of candidate region and the dissimilarity score $Score_{state}$.

2. Estimate confidence measure (RCM) in candidate region by using recognition lattice of spoken document.

3. Combine $Score_{state}$ and RCM by using logistic regression model and the combined score is compared with a threshold for a final decision.

3.3.2 *Estimation of confidence measure from the lattice*

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 combined with $Score_{state}$ obtained by the first-pass DTW-based spotting. A fusion score is obtained by logistic regression model for integrating side information from ASR output as well as for score normalization.

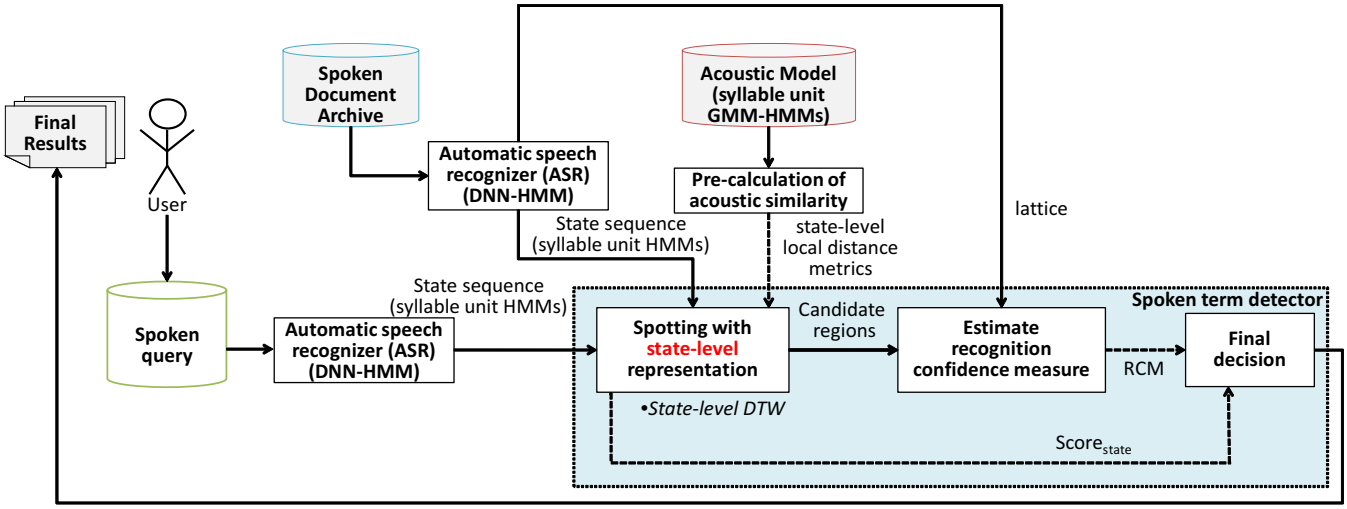

Figure 4: Overview of proposed STD system with the second-pass ASR confidence scoring integration

Given a syllable sequence of a query term $B = \{B_1, \cdots, B_Y\}$, the confidence of a candidate speech segment X is estimated as,

$$RCM(B) = P(B|X) \tag{4}$$

where $P(B|X)$ denotes the posterior probability of the syllable sequence B. In general, this measure can be estimated from the lattice output from ASR system. For simplicity, we approximate the calculation of RCM by

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

where $T = \{i, \cdots, j\}$ is the speech frames of candidate segment detected by the first-pass of our STD system, and $P(s|t)$ denotes the posterior probability of phone s at frame t in lattice.

4. EVALUATION

4.1 Experimental setup

We compared the baseline method described in Section 2 and proposed methods described in Section 3. In all experimental conditions, we used reference automatic transcription recognized by Kaldi toolkit[18] with matched models (K-REF-WOR D-MATCH) for target documents and spoken query transcription provided from NTCIR-12 organizer.

In NTCIR-12 SpokenQuery&Doc-2 task[1], 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 separately. Finally, we obtained a search result by combining all the results by considering AND condition. The spoken query including ASR output of pause (silence) which is no shorter than 200 msec are considered as containing multiple query terms and those output are segmented by the pause.

When we submitted a formalrun result (run ID: SHZU3), DNN to extract BNF is trained with PDNN toolkit[16] and amount of training data was reduced to 234 lecture speeches. After the formalrun, additional experiments were performed by using DNNs for both BNF and posteriorgram extraction which are trained on 910 lecture speeches of CSJ corpus[17] with 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 × 11frames as input features to DNN.

In order to adjust the parameters such as weight α (Eq.(2)), we used dryrun query set of the NTCIR-12 SpokenQuery&Doc-2 SQ-STD task as a development set. For training the logistic regression model described in Section 3.3, dryrun query set as well as the first-pass DTW scores and RCM scores from positive and negative examples of retrieval result were used as a development set data.

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 Evaluation results

Table 1 and 2 show the results for dryrun (development set) and formalrun (evaluation set), respectively.

SHZU1,SHZU2 and SHZU3 have been submitted to the NTCIR-12 SpokenQuery&Doc-2 SQ-STD task formal run. Additional four systems (baseline, state_spot+BNF, state_spot +post, state_ spot+RCM) are evaluated after the formal run submission. SHZU1 and SHZU2 perform DTW-based spotting method for the HMM state sequences described in Section 3.1. SHZU3 and additional two systems(State_spot+BNF and State_spot+post) represent the system based on the combined score by using linear combination described in Section 3.2. As mentioned in the previous section, the DNN for BNF extraction for SHZU3 system was trained with reduced amount of training data for missing a deadline for formalrun submission, while the State_spot+BNF used the DNN trained with much more data compared to SHZU3. In addition, SHZU3 system didn't employ an endpoint-free DTW algorithm as described in Section 3.2, while additional two systems (State_spot+BNF and State_spot+post)

employed the endpoint-free DTW. The parameter β which determines the maximum frame length of extra matched regions in the second-pass acoustic match was empirically set to 30. State_spot+RCM is the system with confidence measure described in Section 3.3. Only state_spot+RCM system used the combined score by logistic regression model described in Section 3.3.

It should be noted that only for SHZU2 and SHZU3 the parameters of the first-pass threshold are determined so that the maximum F-measure is obtained for dryrun query set (therefore, only those two systems are denoted as dryrun_opt). On the other hand, the first-pass threshold in all other systems are simply decided to limit the number of candidates per query to 1000. The result of "ALL" query type consist of the results of both IV (in vocabulary) and OOV (out of vocabulary) queries.

As for the effect of the combination with DNN-based acoustic matches, the proposed methods except SHZU3 are better than baseline. The reason that SHZU3 doesn't improve the performance can be explained by the fact that the DNN for BNF extraction was trained with a reduced amount of data and the second-pass DTW doesn't allow extra matched regions (i.e., insufficient endpoint-free region), while all addtional systems have been improved at these points. The results show that the runs with state_spot+post exhibit the best STD performance for ALL or OOV in MAP. State_spot+post attained the best STD performance for OOV in both dryrun and formalrun.

As for the effect of the combination with ASR confidence measure, the proposed method hardly improves the baseline performance. This may be due to the fact that the development data for learning logistic regression model is too small and the lack of ASR-related features which affect the score, since the factors that affect the STD performance come in a variety of types and they may not be sufficiently modeled only by RCM. In fact, our recent work obtains some promising results by incorporating other ASR-related features in addition to RCM.

5. CONCLUSIONS

In this paper, we introduced two spoken term detection methods which only depend on single ASR system and additional second-pass rescoring system with DNN-based acoustic matches or ASR confidence measures. We proposed combining feature-based acoustic matches with the ASR-derived features and using integrated score is obtained by a logistic regression model with the recognition confidence measure (RCM). The experimental results showed that combining a feature-level matching of posteriorgram and BNF 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. On the other hand, combination with RCM doesn't improve baseline methods.

Our proposed systems have used the first-pass DTW with a state-level local distance derived from syllable-unit GMM-HMM with MFCC feature, while the automatic transcripts as input for the first-pass were given by a DNN-HMM-based decoder. Since our experiment showed much lower accuracy for GMM-HMM-based ASR than DNN-HMM-based ASR, we should replace the local distance with more accurate model-based one. Also, 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 [7].

6. REFERENCES

[1] 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).

[2] Y. Itoh, H. Nishiaki, X. Hu, H. Nanjo, T. Akiba, T. Kawahara, S. Nakagawa, T. Matsui, Y. Yamashita and K. Aikawa : "Constructing Japanese Test Collections for Spoken Term Detection," In Proc. of Interspeech, pp.677-680 (2010).

[3] K. Iwami, Y. Fujii, K. Yamamoto and S. Nakagawa. : "Out-of-vocabulary term detection by n-gram array with distance from continuous syllable recognition results," Proc. of Spoken Language Technology Workshop, pp.212-217 (2010).

[4] N. Ariwardhani, M. Kimura, Y. Iribe, K. Katsurada, T. Nitta : "Phoneme Recognition Based on AF-HMMs with an Optimal Parameter Set," Proc. of NCSP, pp.170-173 (2012).

[5] N. Kanda, H. Sagawa, T. Sumiyoshi, Y. Obuchi : "Open-vocabulary keyword detection from super-large scale speech database," Proc. of MMSP, pp.939-944 (2008).

[6] 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).

[7] 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)

[8] 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).

[9] 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).

[10] 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).

[11] T. Akiba, H. Nishizaki, K. Aikawa 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).

[12] 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).

[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

Table 1: Dryrun(development set) STD performance [%]

querytype	system	run ID	F-measure(max)	MAP
ALL	baseline(syll_spot)	additional	66.37	70.46
	state_spot	SHZU1	66.67	72.57
	state_spot(dryrun_opt)	SHZU2	66.67	56.41
	state_spot(dryrun_opt)+BNF	SHZU3	65.08	71.23
	state_spot+BNF	additional	67.29	76.94
	state_spot+post	additional	68.65	79.89
	state_spot+RCM	additional	65.82	73.36
IV	baseline(syll_spot)	additional	85.29	86.51
	state_spot	SHZU1	85.63	87.43
	state_spot(dryrun_opt)	SHZU2	85.67	76.64
	state_spot(dryrun_opt)+BNF	SHZU3	83.90	86.47
	state_spot+BNF	additional	85.37	90.80
	state_spot+post	additional	83.71	90.69
	state_spot+RCM	additional	84.74	87.70
OOV	baseline(syll_spot)	additional	39.33	51.74
	state_spot	SHZU1	38.21	55.23
	state_spot(dryrun_opt)	SHZU2	38.14	32.82
	state_spot(dryrun_opt)+BNF	SHZU3	39.74	53.46
	state_spot+BNF	additional	41.23	60.77
	state_spot+post	additional	46.19	67.29
	state_spot+RCM	additional	40.36	56.62

Table 2: Formalrun(evaluation set) STD performance [%]

querytype	system	run ID	F-measure(max)	MAP
ALL	baseline(syll_spot)	additional	38.32	64.42
	state_spot	SHZU1	45.94	66.03
	state_spot(dryrun_opt)	SHZU2	9.02	47.34
	state_spot(dryrun_opt)+BNF	SHZU3	7.75	43.02
	state_spot+BNF	additional	48.59	69.05
	state_spot+post	additional	42.75	72.27
	state_spot+RCM	additional	45.10	66.08
IV	baseline(syll_spot)	additional	45.34	73.14
	state_spot	SHZU1	59.46	74.52
	state_spot(dryrun_opt)	SHZU2	12.39	57.23
	state_spot(dryrun_opt)+BNF	SHZU3	10.12	53.00
	state_spot+BNF	additional	57.35	77.48
	state_spot+post	additional	51.19	80.06
	state_spot+RCM	additional	57.74	74.28
OOV	baseline(syll_spot)	additional	20.51	49.20
	state_spot	SHZU1	19.71	51.20
	state_spot(dryrun_opt)	SHZU2	4.18	30.06
	state_spot(dryrun_opt)+BNF	SHZU3	3.84	25.61
	state_spot+BNF	additional	25.26	54.33
	state_spot+post	additional	22.25	58.66
	state_spot+RCM	additional	24.76	51.78

unsupervised speech pattern discovery," Proc. ICASSP, pp. 4366-4369, (2010).

[15] F. Wessel, R. Schluter, K. Macherey and H. Ney : "Confidence measures for large vocabulary continuous speech recognition," IEEE Trans. on Speech and Audio Processing, Vol.9, No.3, pp.288 - 298 (2001).

[16] Y. Miao:"Kaldi+PDNN: Building DNN-based ASR Systems with Kaldi and PDNN," arXiv preprint arXiv:1401.6984 (2014).

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

[18] 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," IEEE ASRU (2011).