Combining Subword and State-level Dissimilarity Measures for Improved Spoken Term Detection in NTCIR-11 SpokenQuery&Doc Task

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@sys.eng.shizuoka.ac.jp

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. Recently, in such spoken document retrieval task, there has been increasing interest in using a spoken query not only for improving usability but also for low-resource languages which may have much errors by LVCSR systems. In this paper, we propose spoken term detection method using multiple scoring and dissimilarity measures for spoken query. Our proposed method is intended to convert the spoken query into a syllable sequence by LVCSR and do search that takes into account the acoustic dissimilarity on spoken documents' LVCSR transcripts. The experimental results showed that our proposed system improve the performance compared to baseline system.

Team Name

SHZU

Subtasks

SQ-STD

Keywords

spoken term detection, acoustic dissimilarity measure, distance between two distributions, spoken query

1. INTRODUCTION

Spoken term detection (STD) is a task which locates a given search term in a large set of spoken documents. A simple approach for STD is a textual search on Large Vocabulary Continuous Speech Recognizer (LVCSR) transcripts. However, the performance of STD is largely affected if the spoken documents include out-of-vocabulary (OOV) words or the LVCSR transcripts include recognition errors for in-vocabulary (IV) words. Therefore, many approaches using a subword-unit based speech recognition system have been proposed[1, 2, 3, 4]. The keyword spotting methods for subword sequences based on dynamic time warping(DTW)-based matching or n-gram indexing approaches have shown the robustness for recognition errors and OOV problems. Also, hybrid approaches with multiple speech recognition systems of word-based LVCSR and subword-unit based speech recognizer have shown the further performance improvement for both IV and OOV query terms[5, 6, 7].

In STD of text query, we have proposed an approach based on the two-pass spoken term detection method with state-level acoustic dissimilarity measures [8] and it is also used in combination with n-gram confidence-based scoring for improved STD accuracy[9, 10]. The method for text query based on new acoustic feature representation, which we call distribution-distance vector (DDV), has shown a significant improvement compared with simple DTW-based matching approaches[8, 9, 10].

In this paper, we propose a STD approach in which a spoken query is converted into a syllable sequence by automatic speech recognition and apply our STD method for text query by treating the syllable sequence as same as text query.

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 [6]. 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 state's output distributions. We define the distance between two Gaussian mixture models P and Q as

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

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 preprocessing stage, N-best recognition results for a spoken document archive are obtained by word-based and syllable-based speech recognition systems with N-gram language models of corresponding unit. Then, the word-based recognition results are converted into subword sequences.

At the stage of STD for spoken query input, the query is converted into a syllable sequence by word-based and syllable-based recognition, and the DTW-based word spotting with an asymmetric path constraint is performed. Then, the system checks if the query is composed only of words in LVCSR system's lexicon. If the query is judged as IV words, word-based recognition results (converted into syllable sequence) are used. Otherwise syllable-based recognition results are used. Finally, a set of segments with a dissimilarity measure less than a threshold is obtained as the retrieval result.

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 performance against recognition errors. One of the first pass methods simply performs the DTW-based query term spotting as described in Section 2. The second pass is a query term verifier which performs two kinds of detailed scoring (rescoring) for each candidate segment found in the first pass. The different approaches to scoring seqments at the first and second passes and their combinations are described in the following sections.

3.2 IV/OOV discrimination of spoken query term

In proposed STD system, when spoken query is inputted, the query is converted into a syllable sequence by word-based and syllable-based automatic speech recognizers. Some our results(runs) reported in Section 4 introduce an automatic discrimination of IV and OOV queries so that word-based and syllable-based transcripts are selectively used. We estimate if the query is composed only of words in LVCSR system's lexicon or not by using both of word-based and syllable-based recognition scores. If the query is composed only of vocabulary words (IV), word-based recognition results (converted into syllable sequence) are used. If the query is estimated to have OOV words, syllable-based recognition results are used. We used the log-likelihood ratio which is defined in Eq.(2) or Eq.(3) for IV/OOV discrimination. If the log-likelihood ratio is less than a threshold, then the query is discriminated as OOV.

$$LR_L = \log \frac{P_L(\boldsymbol{w})}{P_L(\hat{\boldsymbol{w}})} \tag{2}$$

where, $P_L(\boldsymbol{w})$ and $P_L(\hat{\boldsymbol{w}})$ are language likelihood scores of the best candidate $\boldsymbol{w}$ from word-based LVCSR system and the best candidate $\hat{\boldsymbol{w}}$ from syllable-based LVCSR system, respectively.

$$LR_{AL} = \log \frac{P_A(X_q|\boldsymbol{w})P_L(\boldsymbol{w})}{P_A(X_q|\hat{\boldsymbol{w}})P_L(\hat{\boldsymbol{w}})} \tag{3}$$

where, $P_A(X_q|\boldsymbol{w})$ and $P_A(X_q|\hat{\boldsymbol{w}})$ are acoustic likelihood scores of the best candidate $\boldsymbol{w}$ from word-based LVCSR system and the best candidate $\hat{\boldsymbol{w}}$ from syllable-based LVCSR system, respectively.

3.3 N-gram confidence-based scoring

For finding the occurrence of certain subword sequence from the lattice, n-gram confidence-based relevance scoring has been effectively used to deal with the recognition error problem[12]. We adopts the n-gram confidence-based scoring method as an additional filtering process which precedes or follows the two-pass spotting and rescoring passes mentioned in the previous section. The relevance score is compared with a threshold parameter to filter out unlikely speech segments before the two-pass match is performed.

Let the $\hat{Q} = \{w_1, \cdots, w_M\}$ be the estimated subword sequence of a spoken query term and $\{w_i, \cdots, w_{i+n-1}\}$ $(i = 1, \cdots, M-n+1)$ denote partial n-grams of the query term. We define the relevance score R_{n-gram} of speech segment $\boldsymbol{X_s}$ and query term $\hat{Q}$ for each order of n as

$$R_{n-gram} = \sum_{i=1}^{M-n+1} \sum_{\hat{\boldsymbol{W}} \in \boldsymbol{W}(\boldsymbol{X_s})} CM(\hat{\boldsymbol{W}})C(\hat{\boldsymbol{W}}, \{w_i, \cdots, w_{i+n-1}\}) \tag{4}$$

where $C(\hat{\boldsymbol{W}}, \{w_i, \cdots, w_{i+n-1}\})$ is the occurrence count of n-gram $\{w_i, \cdots, w_{i+n-1}\}$ in sentence hypothesis $\hat{\boldsymbol{W}}$ which is included in subword lattice $\boldsymbol{W(X_s)}$, and $CM(\hat{\boldsymbol{W}})$ denotes the confidence score of sentence $\hat{\boldsymbol{W}}$ as the posteriori probability in lattice $\boldsymbol{W(X_s)}$. The final relevance score is obtained by

$$Score_{CM}(\boldsymbol{X_s}, \hat{Q}) = \sum_{n=1}^{N} a_n R_{n-gram} \tag{5}$$

where a_n is a weight parameter. In practice, Eq. (4) is equivalently calculated by efficient forward-backward algorithm from a subword lattice[13].

3.4 Rescoring with state-level representation (2nd pass)

As described in Section 2, the first-pass query term spotting performs DTW-based matching by using the subword-level local distance metric $D_{sub}(x, y)$. The output is a set of aligned subword sequences which have the dissimilarity score below a threshold. The second pass first expands the aligned subword sequences into the corresponding HMM state sequences and calculates dissimilarity score based on a state-level local distance metric.

A simple approach to calculate dissimilarity score between HMM state sequences is the DTW matching based on the local distance measure defined in (1). The dissimilarity scores obtained for each candidate segments are compared with a threshold. We refer to this dissimilarity score as $Score_{BD}$.

Our previous study introduced new acoustic dissimilarity score based on a distance-vector representation which is defined for each HMM state. Like a structural feature representation proposed in [14] and a self similarity matrix in [15], we can consider a feature representation for each HMM state based on the distances between a target state and all states in a set of subword-unit HMMs. It is expected that such structural feature can estimate more robust acoustic dissimilarity measure for comparing the subword sequences including recognition errors.

Let the $\boldsymbol{P} = \{P_s\}(s = 1, 2, \cdots, S)$ be a set of all distributions in subword-unit HMMs. We define a distance vector for the HMM state s as

$$\boldsymbol{\phi}(s) = (D_{BD}(P_s, P_1), D_{BD}(P_s, P_2), \cdots, D_{BD}(P_s, P_S))^{\mathrm{T}} \tag{6}$$

We refer to this vector representation as distribution-distance vector (DDV).

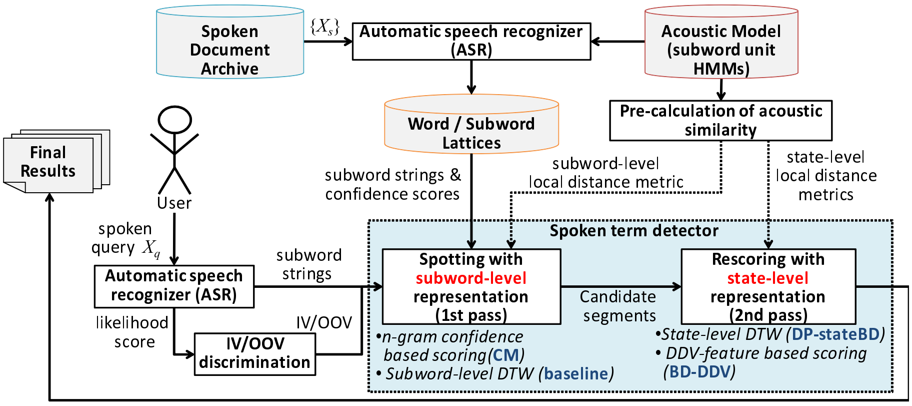
Spoken Document Archive
$\{X_s\}$
Automatic speech recognizer (ASR)
Acoustic Model (subword unit HMMs)
Word / Subword Lattices
Pre-calculation of acoustic similarity
Final Results
User
subword strings & confidence scores
subword-level local distance metric
state-level local distance metrics
spoken query X_q
Spoken term detector
subword strings
Automatic speech recognizer (ASR)
Spotting with subword-level representation (1st pass)
Candidate segments
Rescoring with state-level representation (2nd pass)
likelihood score
IV/OOV
IV/OOV discrimination
•n-gram confidence based scoring(CM)
• Subword-level DTW (baseline)
•State-level DTW (DP-stateBD)
• DDV-feature based scoring (BD-DDV)

Figure 1: Overview of proposed STD system

To simplify the calculation of dissimilarity score using the DDV representation, we can utilize the alignment between two state sequences obtained as a result of calculating $Score_{BD}$. Let the $F = c_1, c_2, \cdots, c_k, \cdots, c_K$ be the state-level alignment and the $c_k = (a_i, b_j)$ represents the correspondence between the i-th state in HMM state sequence $A = a_1, a_2, \cdots, a_I$ and the j-th state in HMM state sequence $B = b_1, b_2, \cdots, b_J$. We investigate the following definition based on the DDV representation.

$$Score_{DDV\_L2max} = \frac{\max_{1\le k\le K}\left\{\sum_{s=1}^{S}|\psi_s(c_k)|^2\right\}^{1/2}}{K\cdot S} \qquad (7)$$

where $\psi_s(c_k)$ is the s-th element of the vector $\phi(a_i) - \phi(b_j)$. $Score_{DDV\_L2Max}$ uses the maximum value of all L2 norms in DDV feature vector sequences and thus it emphasizes the most dissimilar part in a subword sequence.

Finally, the above mentioned dissimilarity scores based on the state-level representations can be combined as

$$Score_{fusion} = \alpha \cdot Score_{BD} + (1-\alpha)\cdot\tau\cdot Score_{DDV} \qquad (8)$$

where $\alpha (0 \le \alpha \le 1)$ is a weight coefficient and τ is a constant for adjusting the score range. To reduce the computational cost, the local distance values between states can be prepared beforehand by using a set of subword-unit HMM parameters.

4. EVALUATION

4.1 Experimental setup

We compared two baseline methods described in Section 2 (baseline1,baseline2) and seven two-pass methods described in Section 3 (SHZU1-7). SHZU1,2 and 3 have been submitted to the NTCIR-11 SpokenQuery&Doc SQ-STD task formal run and their labels correspond to the run IDs SHZU-SPK1,2 and SHZU-TXT3 presented in the overview paper[16]. Additional four runs (SHZU4-7) are performed after the formal run submission. In all conditions, we used reference automatic transcription recognized by matched models (REF-WORD-MATCH, REF-SYLLABLE-MATCH) for target documents. The differences of these conditions are the recognition method of the query and IV/OOV discrimination method of the query. Table 1 shows the differences of each methods.

Table 1: Conditions of LVCSR system used for transcribing spoken query and documents

	Language model (LM)	LM. unit/ IV/OOV descri.
baseline1	nounLM	word
baseline2	nounLM	syllable
baseline3	manual	word
SHZU1	nounLM	word & syllable (Eq.(2))
SHZU2	refLM	word & syllable (Eq.(2))
SHZU3	manual	word
SHZU4	nounLM	word & syllable (Eq.(3))
SHZU5	refLM	word & syllable (Eq.(3))
SHZU6	nounLM	word
SHZU7	nounLM	syllable

In Table 1, Language model (LM) column shows the language model to be used in LVCSR system for transcribing the query. In Language model (LM) column, "nounLM" denotes the word-unit and syllable-unit n-gram language models trained by noun phrases in CSJ corpus, while "refLM" denotes the matched word-unit and syllable-unit language models provided by task organizers, and "manual" shows that we used manual (correct) transcription. LM unit column shows the recognition unit used in LVCSR system for transcribing target document and query. LM unit column also shows the definition used in IV/OOV discrimination of the query. In LM unit column, "word" shows that we treated all query as IV word and "syllable" shows that we treated all query as OOV word.

In order to adjust the parameters such as thresholds, we used dry run query set of the NTCIR-11 SpokenQuery&Doc SQ-STD task as a development set. All parameters are adjusted without distinction between IV and OOV queries. Adjusting the weight α in Eq.(8), $\alpha = 1$ exhibited best performance for the dry run query set. Therefore, we used the $Score_{BD}$ only as $Score_{fusion}$.

Table 2: IV/OOV discrimination accuracy[%]

	% correct
SHZU1	88.18
SHZU2	61.08
SHZU4	84.73
SHZU5	56.65

Table 3: STD performance of spoken query task (SQ-STD) [%]

Query set		Recall	Precision	F-measure
IV	baseline1	46.71	6.56	11.51
	baseline2	34.76	5.06	8.83
	SHZU1	50.07	40.86	45.00
	SHZU2	39.95	34.37	36.95
	SHZU4	49.68	44.54	46.97
	SHZU5	42.42	40.05	41.20
	SHZU6	51.60	40.23	45.21
	SHZU7	30.41	29.24	29.81
OOV	baseline1	8.81	10.20	9.46
	baseline2	6.10	5.81	5.95
	SHZU1	11.19	20.00	14.35
	SHZU2	10.85	20.51	14.19
	SHZU4	11.19	20.00	14.35
	SHZU5	10.51	11.19	10.84
	SHZU6	11.19	20.00	14.35
	SHZU7	13.22	4.59	6.81
All	baseline1	46.18	6.56	11.50
	baseline2	34.35	5.06	8.82
	SHZU1	49.52	40.65	44.65
	SHZU2	39.52	34.31	36.73
	SHZU4	49.13	44.37	46.63
	SHZU5	41.96	39.71	40.80
	SHZU6	49.95	40.70	44.86
	SHZU7	30.11	28.56	29.32

Table 4: STD performance of text query task (SQ-STD) [%]

Query set		Recall	Precision	F-measure
IV	baseline3	64.66	40.58	49.86
	SHZU3	71.72	65.57	68.51
OOV	baseline3	24.75	36.50	29.50
	SHZU3	24.75	58.40	34.76
All	baseline3	64.03	40.59	49.69
	SHZU3	71.04	65.53	68.18

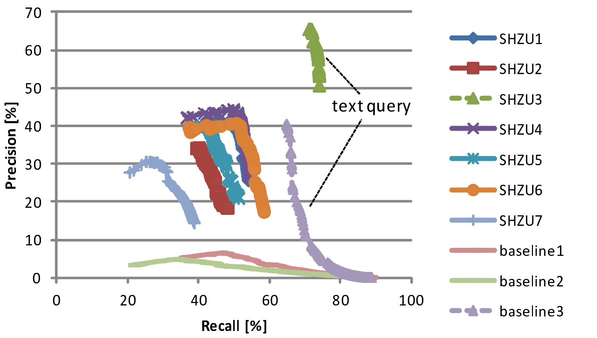

Figure 2: Recall-Precision curves(IV query terms)

4.2 NTCIR-11 SQ-STD task results

Table 2 shows IV/OOV discrimination accuracy for each method. Since the IV rate of query set for NTCIR-11 SpokenQuery&Doc SQ-STD task formal run is 97.54%, and the query terms are dominated by IV cases, the introduction of IV/OOV discrimination does not produce significant result.

Table 3 shows the results (recall, precision, and F-measure(max)) for spoken query STD methods (baseline1,2 and SHZU1-7 excluding SHZU3). Table 4 shows the results for text query STD methods (baseline3 and SHZU3). F-measure(max) is the maximum value of F-measure when the threshold is adjusted. Fig. 2,3,4 shows the recall-precision curves of each method. The result shows that the two-pass methods (SHZU1-7) outperforms the baseline methods which use only the first pass.

The result shows that the runs with nounLM (SHZU1,4) exhibit better performance in compared with the runs with refLM (SHZU2,5). As for the effect of IV/OOV discrimination, the runs with Eq.(3) (SHZU4,5) achieved better performance in compared with the runs with Eq.(2) (SHZU1,2) despite the loss of IV/OOV discrimination accuracy. It should be noted that the run (SHZU6) which didn't apply IV/OOV discrimination (which correspond to IV/OOV discrimination accuracy of 97.54%) doesn't improve the performance in compared with SHZU4. The difference of performance between SHZU3 and the other two-pass methods is quite large. This seems to be due to the poor accuracy of LVCSR system for spoken queries.

As described above, we adjusted parameter without distinction between IV and OOV queries. Therefore, by adjusting the parameters separately for each of IV and OOV query types, it would be possible to improve the performance. We also adjusted the weight α to 1 and didn't use $Score_{DDV}$ in Eq.(8). However, in STD by text query of NTCIR-10 SpokenDoc-2 moderate-size task [17] in which a subset of spoken documents for NTCIR-11 is used, it is shown that it is effective to use the $Score_{DDV}$. For your information, we show the effect of the score weight parameter α for NTCIR-10 SpokenDoc-2 moderate-size task in Fig. 5 [10, 18]. Although detailed analysis is required in terms of different consequence between prior NTCIR-10 and current NTCIR-11 task, it seems that the result is consistent with the fact that our DDV-based approach exhibited more significant effect for OOV query terms as shown in Fig. 5.

5. CONCLUSIONS

In this paper, we introduced spoken term detection method using multiple scoring and dissimilarity measures for spoken query. Experimental result shows that two-pass spoken term detection method in which the state-level acoustic dissimilarity and using the language model trained by noun phrases for recognition of various types of spoken query is effective in improving STD performance.

Since our method is a simple extension of the conventional DTW-based method, it is straightforward to combine with indexing techniques (e.g. [6]) for speeding up our STD system. Also, an automatic estimation of optimal parameters, such as a score threshold and weight, or score normalization

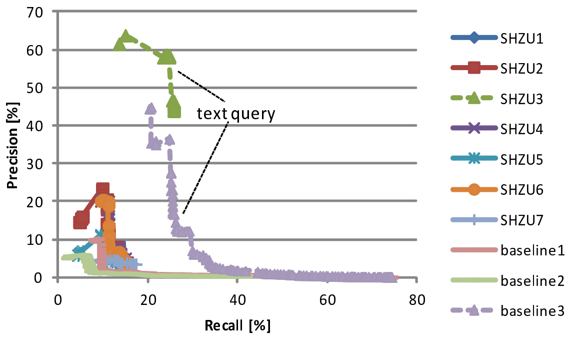

Figure 3: Recall-Precision curves(OOV query terms)

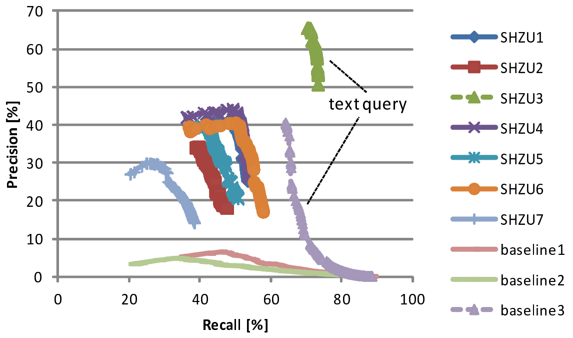

Figure 4: Recall-Precision curves(all)

methods[19] are necessary to achieve further improvement and robustness for spoken documents in the real world.

6. REFERENCES

[1] Y. Itoh, et al.: “Constructing Japanese Test Collections for Spoken Term Detection,” Proc. of INTERSPEECH, pp.677-680 (2010).

[2] K. Iwami, et al.: “Out-of-vocabulary term detection by n-gram array with distance fromcontinuous syllable recognition results,” Proc. of Spoken Language Technology Workshop, pp.212-217 (2010).

[3] N. Ariwardhani, et al.: “Phoneme Recognition Based on AF-HMMs with an Optimal Parameter Set,” Proc. of NCSP, pp.170-173 (2012).

[4] N. Kanda, et al.: “Open-vocabulary keyword detection from super-large scale speech database,” Proc. of MMSP, pp.939-944 (2008).

[5] K.Iwami, et al.: “Efficient out-of-vocabulary term detection by N-gram array in deices with distance from a syllable lattices,” Proc. of ICASSP, pp.5664-5667 (2011).

[6] S.Nakagawa. et al.: “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).

[7] H. Nishizaki, et al. : “Spoken Term Detection Using Multiple Speech Recognizers’ Outputs at NTCIR-9 SpokenDoc STD subtask,” Proc. of NTCIR-9 Workshop Meeting, pp.236-241 (2011).

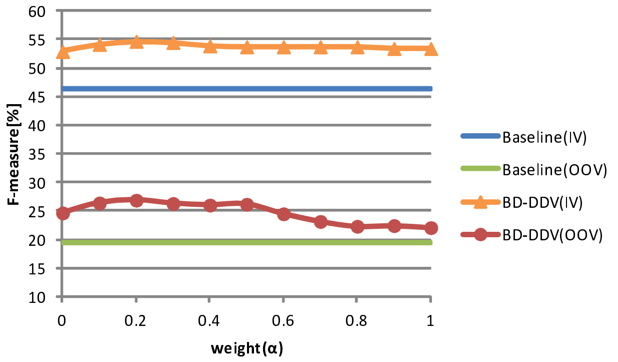

Figure 5: Effect of the score weight parameter α (STD system with $Score_{DDV\_L2Max}$). The curve “IV” and “OOV” show the breakdown of STD performance for in-vocabulary queries and out-of-vocabulary queries, respectively. (NTCIR-10 SpokenDoc-2 moderate-size task)

[8] N. Yamamoto and A. Kai : “Spoken Term Detection Using Distance-Vector based Dissimilarity Measures and Its Evaluation on the NTCIR-10 SpokenDoc-2 Task,” Proc. of the 10th NTCIR Workshop Meeting, pp.648-653, (2013).

[9] 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 (2013).

[10] 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).

[11] T. Akiba, et al.: “Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop,” Proc. of NTCIR-9 Workshop Meeting, pp.223-235 (2011).

[12] H. Lee, P. Chou and L. Lee : “Open-vocabulary retrieval of spoken content with shorter/longer queries considering word/subword-based acoustic feature similarity,” Proc. of INTERSPEECH (2012).

[13] 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).

[14] N. Minematsu et al.: “Structural representation of the pronunciation and its use for CALL,” Proc. of Spoken Language Technology Workshop, pp.126–129 (2006).

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

[16] Tomoyosi Akiba, Hiromitsu Nishizaki, Hiroaki Nanjo and Gareth J. F. Jones : “Overview of the NTCIR-11 SpokenQuery&Doc task,” In Proceedings of the NTCIR-11 Conference, Tokyo, Japan, (2014).

[17] Tomoyosi Akiba, Hiromitsu Nishizaki, Kiyoaki Aikawa, Xinhui Hu, Yoshiaki Itoh, Tatsuya Kawahara, Seiichi Nakagawa, Hiroaki Nanjo and Yoichi Yamashita : “Overview of the NTCIR-10 SpokenDoc-2 Task,” Proc. of the 10th NTCIR Workshop Meeting, (2013).

[18] M. Makino, N. Yamamoto and A. Kai : "Using Acoustic Dissimilarity Measures Based on Distance Vector Representation for Improved Spoken Term Detection by Text and Spoken Queries", Proc. of the eighth annual Spoken Document Processing Workshop(SDPWS) (2014). (in Japanese)

[19] B. Zhang, et al.: "White Listing and Score Normalization for Keyword Spotting of Noisy Speech," Proc. of INTERSPEECH (2012).