# Spoken Term Detection Using Distance-Vector based Dissimilarity Measures and Its Evaluation on the NTCIR-10 SpokenDoc-2 Task

Naoki Yamamoto Shizuoka University 3-5-1 Johoku,Hamamatsu-shi,Shizuoka 432-8561,Japan yamamoto\_nao@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. One of the most popular approaches performs indexing based on the sub-word sequence which is converted from the recognition hypotheses from LVCSR decoder for considering recognition errors and OOV problems. In this paper, we propose acoustic dissimilarity measures for improved STD performance. The proposed measures are based on a feature sequence of distance-vector representation, which consists of all the distances between two possible combinations of distributions in a set of subword unit HMMs and represents a structural feature. The experimental results showed that our two-pass STD system with new acoustic dissimilarity measure improve the performance compared to the STD system with a conventional acoustic measure.

#### Team Name

SHZU

## Subtasks

Spoken Term Detection

# **Keywords**

spoken term detection, distance between two distributions, distance measure between two structures, acoustic similarity

# 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 invocabulary (IV) words. Therefore, many approaches using a subword-unit based speech recognition system have been proposed[2, 4, 5, 9]. 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[10, 11, 12].

In this paper, we introduce a keyword verifier which utilizes new acoustic dissimilarity measures based on different types of local distance metrics derived from a common set of subword-unit acoustic models for improved STD. In general, the STD approaches based on subword sequences assumes a predefined local distance measure between subword units and some cost parameters. However, the performance is degraded if the automatic transcripts have many recognition errors including insertions and deletions as in the recordings of spontaneous speech. To address the lack of acoustic information in subword sequences which are derived from LVCSR or subword-unit based speech recognition results, we extend the local distance measure to account for state-level acoustic dissimilarity based on the subword-unit HMMs which are commonly used for speech recognition systems. We also introduce a keyword verifier which aims at the detailed matching between query term and subword sequences based on the proposed state-level acoustic dissimilarity measures. It should be noted that our approach is different from the hierarchical approach which uses frame-level acoustic match[9] which consumes time and is solely based on the subwordbased (N-best) transcripts. Thus, it's easy to extend our method by hybrid speech recognition approaches and fast indexing with table lookup methods.

Related works using the acoustic similarity for STD task are roughly divided into two types: STD systems for text query input (e.g. [3]) and those for spoken query input or unsupervised spoken keyword spotting (e.g. [6, 7, 8]). Typically, the former systems use certain information about confusability between subwords. In [11], a syllable-level distance measure based on the Bhattacharyya distance derived from syllable-unit HMMs is used. Though our proposed acoustic measures is also based on subword-unit HMMs, the state-level local distance instead of subword-level one is used for evaluating the match between query and subword sequences. Also, new feature vector representation for each state in subword-unit HMMs is constructed based on the distances of all possible pairs of distributions in a set of subword-unit HMMs. This feature representation is related to the idea of using an invariant structural feature for removing acoustic variations caused by non-linguistic factors[13, 14] and it is expected that the proposed feature is effective for erroneous transcripts. Recently, similar idea of using structural feature for acoustic dissimilarity estimation is effectively applied to the systems of latter type. In [7], a speech segment is represented as the posteriorgram sequence of GMM or HMM states, and evaluate the similarity between query term and speech segments by using a self similarity matrix. The result showed the robustness to the various language conditions that are different from the training data.

# 2. PROPOSED SPOKEN TERM DETECTION METHOD

## 2.1 Proposed system overview

Overview of our proposed STD system is shown in Figure 1. The system adopts two-pass strategy for both efficient processing and improved STD performance against recognition errors. The first pass performs the DTW-based keyword spotting as described in Section 2.2. The second pass is a keyword verifier which performs two kinds of detailed scoring (rescoring) for each candidate segment found in the first pass. The detailed procedure for STD is as follows.

- 1. Perform the 1st-pass keyword (query term) spotting based on the DTW-based matching with an asymmetric path constraint shown in Figure 2, and obtain a set of candidate segments.
- 2. Perform the DTW-based matching for the HMM state sequences between query and candidate segments with the state-level local distance measure defined in Section 2.2.2 and a symmetric path constraint shown in Figure 3. This step yields a dissimilarity score  $Score_{BD}$  for each candidate segment.
- 3. Calculate the acoustic dissimilarity score *Score*<sub>DDV</sub> using a distance-vector representation as feature (described in Section 2.3.1 and 2.3.2).
- Combined score is calculated for each candidate segment and the score is compared with a threshold for a final decision.

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

where  $\alpha(0 \le \alpha \le 1)$  is a weight coefficient and  $\tau$  is a constant for adjusting the score range. Figure 4 shows the concept of calculating the combined score.

To reduce the computational cost, the local distance values required in Step 1-3 are prepared beforehand by using a set of subword-unit HMM parameters.

## 2.2 Keyword Spotting System (1st Pass)

#### 2.2.1 Keyword Spotting

Our 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 the baseline systems for both of NTCIR-9 SpokenDoc and NTCIR-10 SpokenDoc2 STD subtasks [3, 1], a similar

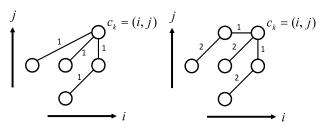


Figure 2: Asymmetric Figure 3: Symmetric path constraint path constraint

method with the local distance measure based on phonemeunit edit distance is used. In our system, the local distance measure is defined by a syllable-unit acoustic dissimilarity as described in Section 2.2.2, and a look-up table is precalculated from an acoustic model.

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 query input, the query term is converted into a syllable sequence, and the DTW-based word spotting with an asymmetric path constraint as shown in Figure 2 is performed. If the term consists of in-vocabulary (IV) words, word-based recognition results (converted into syllable sequence) are used. If the term consists of out-ofvocabulary (OOV) words, syllable-based recognition results are used. Finally, a set of segments with a spotting score (dissimilarity) less than a threshold is obtained as the candidate segments for the second pass.

# 2.2.2 Acoustic dissimilarity based on subword-unit HMM

In [11], the local distance measure is based on the Bhattacharyya distance between two distributions and derived from the acoustic model parameters of syllable-unit HMMs. The Bhattacharyya distance between two distributions Pand Q is expressed as follows when they are multivariate Gaussian distributions.

$$BD(P,Q) = \frac{1}{8}(\mu_P - \mu_Q) \left( \frac{\sum_P + \sum_Q}{2} \right)^{-1} (\mu_P - \mu_Q)^t + \frac{1}{2} \log \left( \frac{|(\sum_P + \sum_Q)/2|}{|\sum_P |^{1/2}|\sum_Q |^{1/2}} \right)$$

where  $\mu$ . is the mean vector and  $\sum_{i}$  is the covariance matrix of each distribution, respectively.

Since each subword-unit HMM has multiple states and state-level distribution is modeled as Gaussian mixture model (GMM) in general, the definition of distance between two HMMs is not straightforward. Therefore, first we define the between-state distance between two GMMs P and Q as

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

where the superscript notations u and v denote a single Gaussian component of each GMM.

Then, we calculate the subword-level distance  $D_{sub}(x, y)$  by the DTW-based matching between two HMM state sequences which correspond to two subwords x and y, respectively, with the local distance defined in equation (1) and a symmetric DTW path constraint shown in Figure 3. The

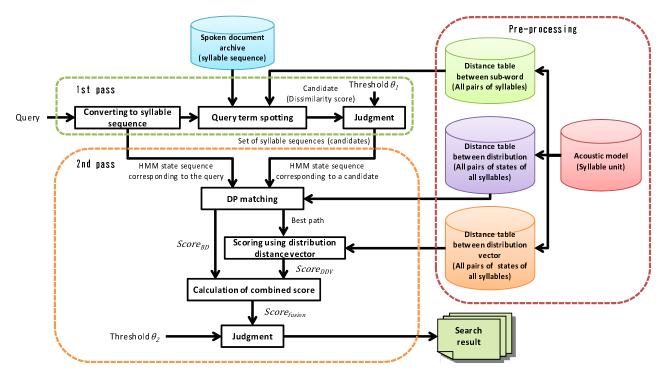


Figure 1: Overview of proposed STD system

distance  $D_{sub}(x, y)$  is used as the local distance of the DTWbased matching at the first pass (Step 1).

## 2.3 Keyword Verifying System (2nd Pass)

#### 2.3.1 Distance vector representation

The distance  $D_{BD}(P,Q)$  in equation (1) only depends on the parameters of two distributions which correspond to a pair of aligned states in DTW-based matching of HMM state sequences. Like a structural feature representation proposed in [13] and a self similarity matrix in [7], 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  $\mathbf{P} = \{P_s\}(s = 1, 2, \dots, S)$  be a set of all distributions in subword-unit HMMs. We define a distance vector for the HMM state s as

$$\phi(s) = (D_{BD}(P_s, P_1), D_{BD}(P_s, P_2), \cdots, D_{BD}(P_s, P_S))^{\mathrm{T}}$$
(2)

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

#### 2.3.2 *Keyword verifier based on distance vector sequences*

We can replace the local distance measure used by the DTW-based matching in Step 2 with a new dissimilarity measure based on the DDV representation in equation (2). To simplify the calculation of dissimilarity score using the DDV representation, we utilize the alignment between two state sequences obtained by the DTW process in Step 2.

Let the  $F = c_1, c_2, \dots, c_k, \dots, c_K$  be the state-level alignment obtained in Step 2 and the  $c_k = (a_i, b_j)$  represents the correspondence between *i*-th state in HMM state sequence  $A = a_1, a_2, \dots, a_I$  and the *j*-th state in HMM state sequence  $B = b_1, b_2, \dots, b_J$ . In our proposed system, two state sequences correspond to a query and candidate segment respectively, which are identical to the input for the DTW-based matching in Step 2. We investigate the following three types of definitions as the dissimilarity score for a candidate segment.

$$Score_{DDV\_L1} = \frac{\sum_{k=1}^{K} \sum_{s=1}^{S} |\psi_s(c_k)|}{K \cdot S}$$
(3)

$$Score_{DDV\_L2} = \frac{1}{K} \sum_{k=1}^{K} \left\{ \frac{1}{S} \sum_{s=1}^{S} |\psi_s(c_k)|^2 \right\}^{1/2}$$
(4)

$$Score_{DDV\_L1Max} = \frac{\max_{1 \le k \le K} \sum_{s=1}^{S} |\psi_s(c_k)|}{K \cdot S}$$
(5)

where  $\psi_s(c_k)$  is the s-th element of the vector  $\phi(a_i) - \phi(b_j)$ . We use these definitions as a dissimilarity score because these scores take a value closer to zero as two state sequences A and B become acoustically similar.  $Score_{DDV\_L1}$  represents a normalized score of accumulated L1 norms between two DDV sequences, while  $Score_{DDV\_L2}$  represents a normalized score of accumulated L2 (Euclidean) norms (although not strictly L2 norm since a normalization term 1/S is included). On the other hand,  $Score_{DDV\_L1Max}$  uses the maximum value of all L1 norms in a DDV sequence and thus it emphasizes the most dissimilar part in a subword sequence.

Figure 4 shows the concept of the detailed scoring process at the second pass (Step 2-4 described in Section 2.1).

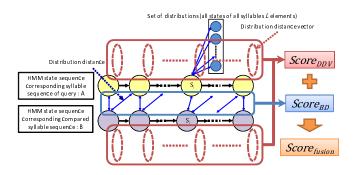


Figure 4: Concept of the detailed scoring process at the second pass

# 3. EVALUATION

### **3.1** Experimental setup

We prepared a set of subword-unit HMMs which are used in calculating the acoustic dissimilarities between subwords and states. We used a training set which is identical to the condition for training acoustic models used in NTCIR-10 SpokenDoc2 baseline system. Table 1 shows the specifications of the acoustic model used for calculating the distance between the distributions. Each HMM has five states and three output distributions for a part of mora categories (/a/, /i/, /u/, /e/, /o/, /N/, /q/, /sp/, /silB/, and /silE/), seven states and five output distributions for the other mora categories.

Two kind of acoustic models are used for NTCIR tasks:

- SHZU-1 Syllable-unit HMMs that were trained using the CSJ corpus [16], while initial HMMs were trained using two commonly-used read speech databases: ASJ-PB(phonetically balanced sentences of continuous speech uttered by 30 males and 34 females) and JNAS(Japanese Newspaper Article Sentences, 15911 sentences by male speakers and 15860 sentences by female speakers).
- **SHZU-2** Syllable-unit HMMs that were trained by the flat start method using only the CSJ corpus.

We used both of word-based and syllable-based reference automatic transcriptions ("REF-WORD-MATCHED" and "REF-SYLLABLE-MATCHED") distributed by organizers. The 10-best hypotheses are used for the first pass described in Section 2.2.1.

Table 2 shows the speech recognition performance for CSJ CORE lectures using three acoustic models: the reference (triphone) acoustic model (RCG-AM) used by NTCIR-10 organizers for providing automatic transcriptions and the syllable-unit acoustic models for providing the distance tables of acoustic dissimilarity (SHZU1-AM and SHZU2-AM) in our system. Note that SHZU1-AM and SHZU2-AM are only used for calculating acoustic dissimilarity and not used for preparing automatic transcriptions.

## **3.2** Evaluation results

#### 3.2.1 Comparison of dissimilarity measures

Table 3 and Figure 5 show the performance of baseline and our systems for NTCIR9 SpokenDoc STD subtask. The

Table 1: Specifications of the HMM used in calculating the distance between the distribution

Category/Unit	133 syllables(morae)
# of states	7 or 5
# of output states	5 or 3
Output distribution	32 mixture, normal
	(diagonal covariance matrix)
Feature parameter	38 dimensions $(MFCC + \Delta MFCC)$
	$+\Delta\Delta MFCC + \Delta Power + \Delta\Delta Power)$

Table 2: Speech recognition performance for CSJ CORE lectures[%]. "Syl.Corr." and "Syl.Acc." denotes the syllable-based correct rate and accuracy, respectively. In case of word-based language model (LM), all words were converted to syllable sequences.

	Word-ba	sed LM	Syllable-based LM	
AM	Syl.Corr.	Syl.Acc.	Syl.Corr.	Syl.Acc.
RCG-AM (triphone)	86.5	83.0	81.8	77.4
SHZU1-AM (syllable)	82.6	78.3	75.2	72.3
SHZU2-AM (syllable)	82.5	78.2	75.1	72.1

NTCIR baseline and our baseline system (1st pass only) are compared with the proposed methods which use three types of DDV-based score definitions described in Section 2.3.2 at the second pass. Note that our baseline system is similar to the organizer's baseline system in that they are based on the DTW-based matching of subword sequences. Major differences are as follows: the organizer's baseline result is based on the transcriptions of REF-SYLLABLE [3] and uses phoneme-based edit distance, while our baseline (1st pass) system is based on the hybrid use of the **REF-SYLLABLE** and **REF-WORD** transcriptions and uses syllable-based acoustic dissimilarity. These results show that the two-pass method with a  $Score_{DDV\_L1Max}$  outperforms the others. So the proposed system with  $Score_{DDV_{-L1Max}}$ was used for the NTCIR-10 evaluations described in the next subsection.

#### 3.2.2 NTCIR-10 STD task results

The evaluation results for CSJ (large-size) task are shown in Table 4 and Figure 6. The decision point for calculating

Table 3: Spoken term detection performance of NTCIR-9 SpokenDoc STD subtask[%].

	Recall	Precision	F-measure	MAP	
NTCIR-9 baseline <sup>*</sup>	NA	NA	52.7	59.5	
Our baseline (1st pass only)	50.6	80.1	62.0	63.2	
$Score_{DDV-L1}$	61.2	70.9	65.7	62.4	
$Score_{DDV_{-L2}}$	58.1	77.3	66.3	62.7	
Score <sub>DDV_L1Max</sub>	58.1	85.2	69.1	63.5	

\* SpokenDoc STD subtask (formal-run of CORE set)[3]

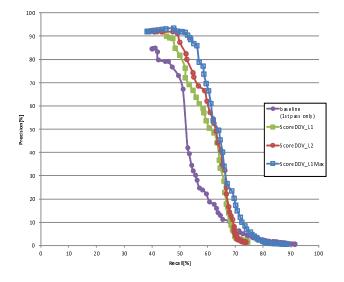


Figure 5: Recall-Precision curves for the CORE formal-run query set in NTCIR-9 SpokenDoc STD subtask

"spec. F" was decided by the result of the CORE formal-run query set in the NTCIR9 SpokenDoc STD subtask. The parameters (1st pass threshold, weight coefficient and 2nd pass threshold) were adjusted for each set of IV and OOV queries to attain the best F-measure value for the final output in the 2nd pass.

The evaluation results for SDPWS (moderate-size) task are shown in Table 5 and Figure 7. The decision point for calculating "spec. F" was decided by the result of the NT-CIR10 SpokenDoc2 SDPWS dry-run query set.

The curves of "baseline1-3" show the results provided by organizers [1]. Baseline systems perform the DTW-based word spotting with phoneme-based edit distance. The "baseline1" system calculates over the syllable-based transcriptions, "baseline2" system calculates over the word-based transcriptions, and "baseline3" system calculates over the word-based and syllable-based transcriptions.

Table 5 shows that our two-pass systems (SHZU-1 and SHZU-2) significantly improve the STD performance compared with one-pass only systems (SHZU-1(1pass) and SHZU-2(1pass)) which are similar to the organizer's baseline3 system. The SHZU-1 system attains a slightly better performance in terms of F-measure and MAP than the SHZU-2 system in Table 5, while the SHZU-1 system is slightly worse than the SHZU-2 system in Table 4. One reason for only a slight difference between the SHZU-1 and SHZU-2 STD performances is explained by insignificant difference in the speech recognition performance between two acoustic models used in these systems as shown in Table 2.

The results show that the performance of "baseline2" and "baseline3" are better than our proposed methods, especially for SDPWS task. One of the reasons for this is thought to be the wrong use of the transcriptions provided by the NTCIR organizers because the difference between the organizer's baseline3 system and our systems (1st. pass only) are very similar but their results differ significantly. The main difference between the baseline3 and our system (1st. pass only) are only the definition of local distance for the DTW

Tab	le 4:	$\mathbf{STD}$	$\mathbf{results}$	$ {\bf for} $	$\mathbf{CSJ}$	(large-size)	$\operatorname{task}$
-----	-------	----------------	--------------------	---------------	----------------	--------------	-----------------------

System	max F.[%]	$\operatorname{spec.F}[\%]$	MAP
baseline1	42.32	40.71	0.500
baseline2	52.52	48.22	0.507
baseline3	54.25	50.46	0.532
SHZU-1	49.44	47.56	0.423
SHZU-2	51.14	44.20	0.510

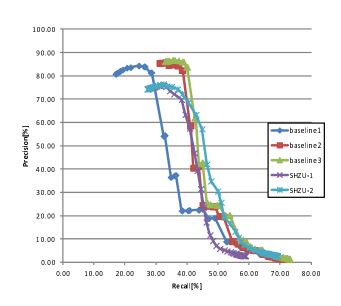


Figure 6: Recall-Precision curves for CSJ (largesize) task

matching and the unit of subword, that is the phoneme v.s. the syllable. Also, comparison between the NTCIR10 runs of organizer's baseline and our system showed that our proposed method often incorrectly judged the IV query as the OOV query, while the word-based recognition results are used for IV queries and syllable-based recognition results are used for OOV queries in our system. Therefore, we conducted additional experiments using the REF-WORD-MATCHED transcription only, which is similar to the organizer's baseline2 condition. The bottom lines in Table 5 show the additional results obtained by our systems based on the REF-WORD-MATCHED transcriptions instead of the hybrid use of the REF-SYLLABLE-MATCHED and REF-WORD-MATCHED transcriptions (the upper four SHZU systems in the middle of the table). The comparison between two SHZU-1(1st. pass) systems in this table reveals that only the change of transcriptions (not using REF-SYLLABLE-MATCHED) greatly improve the STD performance. Accordingly, our two-pass system attains a performance comparable with the baseline2 system, while the performance of the 1st. pass is still worse, and the performance approached to those of the baseline3 system. These result seem promising since the speech recognition performances of used acoustic models (SHZU1-AM and SHZU2-AM) are worse than the RCG-AM used for preparing the transcriptions by organizer's, but our two-pass systems still improved the performance.

ask			
System	max F.[%]	$\operatorname{spec.F}[\%]$	MAP
baseline1	25.08	24.70	0.317
baseline2	37.58	37.46	0.358
baseline3	39.36	39.16	0.393
$SHZU-1(1st pass)^+$	25.24	20.85	0.335
$SHZU-2(1st pass)^+$	24.20	21.63	0.334
SHZU-1	28.62	27.75	0.337
SHZU-2	27.40	23.55	0.319
$SHZU-1(1st pass)^{+\#}$	33.71	-	0.382
$SHZU-2(1st pass)^{+\#}$	32.53	-	0.386
SHZU-1 <sup>+#</sup>	37.85	-	0.359
$SHZU-2^{+\#}$	38.18	-	0.400

Table 5: STD results for SDPWS (moderate-size) task

\* The upper four systems (SHZU-1 and SHZU-2) are based on the hybrid use of the REF-SYLLABLE-MATCHED and REF-WORD-MATCHED transcriptions, while the bottom four systems (marked by a superscript <sup>#</sup>) are based on the REF-WORD-MATCHED transcription only.

<sup>+</sup> These results have not been submitted to the NTCIR-10 formal run and included for reference.

## 4. CONCLUSIONS

We participated in NTCIR10 SpokenDoc-2 STD task. In this paper, we proposed a method for evaluating acoustic dissimilarity between two sub-word sequences based on a sequence of distance-vector representation, which consists of all the distances between two possible combinations of distributions in a set of sub-word unit HMMs and represents a structural feature.

Since our method is a simple extension of the conventional DTW-based method, it is straightforward to replace the 1st. pass with more improved method or to combine with indexing techniques (e.g. [11]) for speeding up our STD system. Also, an automatic estimation of optimal parameters, such as a score threshold and weight, or score normalization[15] are necessary to achieve the further improvement and the robustness for the spoken documents in the real world.

#### 5. **REFERENCES**

- Tomoyosi Akiba, Hiromitsu Nishizaki, Kiyoaki Aikawa, Xinhui Hu, Yoshiaki Itoh, Tatsuya Kawahara, Seiichi Nakagawa, Hiroaki Nanjo, Yoichi Yamashita : "Overview of the NTCIR-10 SpokenDoc-2 Task," Proc. of the 10th NTCIR Workshop Meeting, (2013).
- [2] Y. Itoh, et al.: "Constructing Japanese Test Collections for Spoken Term Detection," Proc. of Interspeech, pp.677-680 (2010).
- [3] 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).
- [4] K. Iwami, et al.: "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).
- [5] N. Ariwardhani, et al.: "Phoneme Recognition Based on AF-HMMs with an Optimal Parameter Set," Proc. of NCSP, pp.170-173 (2012).
- [6] Y. Zhang and J. R. Glass: "Unsupervised Spoken

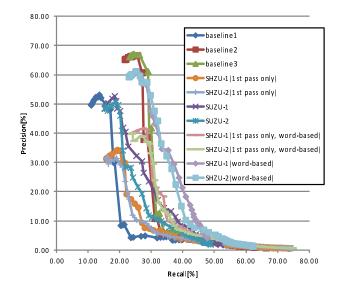


Figure 7: Recall-Precision curves for SDPWS (moderate-size) task

Keyword Spotting via Segmental DTW on Gaussian Posteriorgrams," Proc. of ASRU, pp.398-403 (2009).

- [7] 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).
- [8] Lee. H, et al.: "Open-Vocabulary Retrieval of Spoken Content with Shorter/Longer Queries Considering Word/Subword-based Acoustic Feature Similarity," Proc. of Interspeech (2012).
- [9] N. Kanda, et al.: "Open-vocabulary keyword detection from super-large scale speech database," Proc. of MMSP, pp.939-944 (2008).
- [10] 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).
- [11] 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).
- [12] 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).
- [13] N. Minematsu et al.: "Structural representation of the pronunciation and its use for CALL," Proc. of Spoken Language Technology Workshop, pp.126–129 (2006).
- [14] T. Murakami et al.: "Japanese vowel recognition based on structural representation of speech," Proc. of EUROSPEECH, pp.1261-1264 (2005)
- [15] B. Zhang, et al.: "White Listing and Score Normalization for Keyword Spotting of Noisy Speech," Proc. of Interspeech (2012).
- [16] K. Maekawa, et al.: "Spontaneous speech corpus of Japanese," Proc. of LREC, pp.947-952 (2000).