Using Acoustic Dissimilarity Measures Based on State-level Distance Vector Representation for Improved Spoken Term Detection

Naoki Yamamoto and Atsuhiko Kai
Graduate School of Engineering, Shizuoka University, Japan
E-mail: yamamoto_nao@spa.sys.eng.shizuoka.ac.jp, kai@sys.eng.shizuoka.ac.jp

***Abstract*—This paper proposes a simple approach to subword-based spoken term detection (STD) which uses improved acoustic dissimilarity measures based on a distance-vector representation at the state-level. Our approach assumes that both the query term and spoken documents are represented by subword units and then converted to the sequence of HMM states. A set of all distributions in subword-based HMMs is used for generating distance-vector representation of each state of all subword units. The element of a distance-vector corresponds to the distance between distributions of two different states, and thus a vector represents a structural feature at the state-level. The experimental result showed that the proposed method significantly outperforms the baseline method, which employs a conventional acoustic dissimilarity measure based on subword unit, with very little increase in the required search time.**

I. Introduction

Spoken term detection (STD) is a task which locates a given search term in a large set of spoken documents. To deal with out-of-vocabulary (OOV) problems and recognition errors, many approaches using a subword-unit based speech recognition system have been proposed [1]–[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]–[7].

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.

Related works using the acoustic similarity for STD task are roughly divided into two types: STD systems for text query input (e.g. [8]) and those for spoken query input or unsupervised spoken keyword spotting (e.g. [9]–[11]). Typically, the former systems use certain information about confusability between subwords. In [6], 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, the 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 [12], [13] 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 [10], 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.

In this study, the experiments were conducted on a NTCIR-9 SpokenDoc STD subtask [8] which targets a document collection of the Corpus of Spontaneous Japanese(CSJ). The experimental results show that the proposed method significantly outperforms the baseline methods, which employ either a edit distance or conventional acoustic dissimilarity measure based on subword unit, with very little increase in the required search time. It should be noted that our approach is different from the hierarchical approach which uses frame-level acoustic match [4] which consumes time and is solely based on the subword-based (N-best) transcripts. Thus, it's easy to extend our method by hybrid speech recognition approaches and fast indexing with table lookup methods.

II. Baseline spoken term detection system

A. Baseline system overview

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 [8], a similar system with the local distance measure based on phoneme-unit edit distance is used. In our system, the local distance measure is defined by a syllable-unit acoustic dissimilarity as described in Section II-B, 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 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-Of-Vocabulary (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 retrieval result.

B. Acoustic dissimilarity based on subword-unit HMM

In [6], the local distance measure is based on the Bhattacharyya distance between two distributions and derived from the acoustic model parameters of syllable-unit HMMs. We define the between-state distance between two GMMs P and Q as

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

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.

Then, we calculate the between-subword distance $D_{sub}(x,y)$ by the DTW-based matching of two subword HMMs with the local distance defined in (1) and a symmetric DTW path constraint.

III. Proposed spoken term detection method

A. 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. The first pass performs the DTW-based keyword spotting as described in Section II. 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 spotting and obtain a set of candidate segments (same as the baseline system described in Section II).

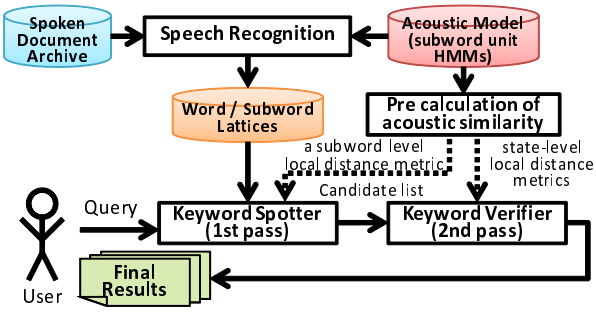

Fig. 1. Overview of proposed STD system

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 (1) and obtain the dissimilarity score $Score_{BD}$ for each candidate segment.
3) Calculate the acoustic dissimilarity score $Score_{DDV}$ using a distance-vector representation as feature (described in Section III-B and III-C).
4) 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 $\alpha (0 \leq \alpha \leq 1)$ is a weight coefficient and τ is a constant for adjusting the score range.

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.

B. Distance vector representation

The distance $D_{BD}(P,Q)$ in (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 [12] and a self similarity matrix in [10], 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 P_s be a distribution corresponding to a state in a subword-unit HMM, and 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

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

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

C. 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 (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 $F = c_1, c_2, \cdots, c_k, \cdots, c_K$ be the state-level alignment obtained in Step 2 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 $B = b_1, b_2, \cdots, b_J$. 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.

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

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

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

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.

IV. Experiments

A. Experimental setup

Our target document collection is CORE lectures (177 lectures, about 44 hours) of the Corpus of Spontaneous Japanese (CSJ). As with NTCIR-9 SpokenDoc STD evaluation [8], the Inter-Pausal Units (IPU) are used as the basic unit to be searched and a retrieval result of an IPU is regarded as correct if it includes the query term. The term set is composed of 50 queries (IV:19, OOV:31) which were used for the formal-run (CSJ-CORE set) in the NTCIR-9 SpokenDoc STD subtask.

We used both of word-based and syllable-based reference automatic transcriptions distributed at NTCIR-9 SpokenDoc evaluation. These reference automatic transcriptions include N-best results (N=10) using a triphone acoustic model and word/syllable n-gram language models.

As for the acoustic model which used at the calculation of the acoustic dissimilarity, an independent set of syllable (mora)-unit HMMs (133 units in total) is used. The models are trained in the same way as the NTCIR-9 reference models are trained. Although a common set of subword HMMs can be used both for performing speech recognition and estimating acoustic dissimilarity, we adopt the syllable-unit model to make the size of a distance table more compact and to make the 2nd-pass search more efficient. Also, we prepared the local distance tables, which are used in our baseline and proposed systems as described in Section II-B and III. Thus, the experiments were performed under open conditions for the spoken documents used in the evaluation. Table I shows the speech recognition performance of two acoustic models: the reference (triphone) acoustic model (RECOG-AM) for providing automatic transcriptions and the syllable-unit acoustic model for providing the distance tables of acoustic dissimilarity (DIST-AM).

TABLE I
SPEECH RECOGNITION PERFORMANCE[%].
"Syl.Corr." and "Syl.Acc." denotes the syllable-based correct rate and accuracy, respectively.

	Word-based LM		Syllable-based LM	
AM	Syl.Corr.	Syl.Acc.	Syl.Corr.	Syl.Acc.
triphone (RECOG)	86.5	83.0	81.8	77.4
syllable (DIST)	82.5	78.2	75.1	72.1

As measures of search performance, we use Recall, Precision, F-measure(max), and MAP. F-measure(max) is the maximum value of F-measure when the threshold is adjusted.

B. Effect of the proposed method and DDV-based scores

Fig. 2 shows Recall-Precision curves of baseline method and proposed methods with three types of DDV-based score definitions. All proposed methods are based on the STD using the combined score described in Section III-A. The parameters of the 1st-pass threshold and a weight coefficient for the combined score are adjusted for each set of IV and OOV queries.

These results show that our proposed method outperforms the baseline system which uses the 1st pass only. This seems that new acoustic dissimilarity measure based on the DDV representation could reject many unreliable candidates including recognition errors than using a conventional measure. As for the proposed methods, the parameter of the 1st-pass threshold was adjusted to attain the best F-measure value for the final output in the second pass. In case of the two-pass method with a $Score_{DDV\_L1Max}$, the precision performance as well as the other evaluation measures were significantly improved, while the recall and precision were about 81% and 3%, respectively, at the 1st-pass output. It seems that the dissimilarity is well represented by focusing on those parts which are most far in a candidate segment when used with the DDV-based feature representation.

So far, typical speeding-up strategies such as indexing have not been applied in our STD system. However, our two-pass approach can take advantage of the progressive search with an efficient scoring with distance look-up tables. In case of parallel processing with two CPU cores (Xeon X5560 2.80GHz), it took about 0.7 seconds per query for the STD of CORE lectures (177 lectures, about 44 hours). The breakdown of this processing time was 0.68 seconds for the 1st pass and 0.02 seconds for the 2nd pass.

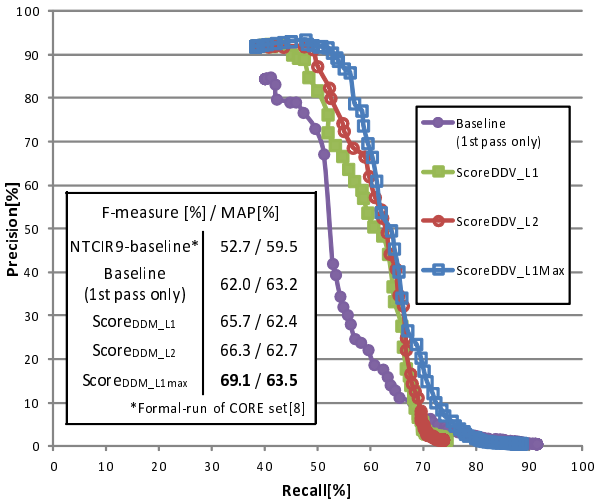

Fig. 2. Recall-Precision curves of different STD systems for NTCIR-9 SpokenDoc STD subtask

C. Effect of the weight of combined score

In the previous section, all the STD performance measures were presented with the adjusted parameters which maximize the performance in terms of F-measure value. The parameters include the 1st-pass threshold θ_1, the weight coefficient α, and the 2nd-pass threshold θ_2. Therefore, we analyze the performance when changing the score binding weight α which significantly affect the scoring with distribution distance vector. Fig. 3 shows the F-measure values of the STD system with $Score_{DDV\_L1Max}$ changing the weight parameter α between 0 and 1. As described before, word-based transcriptions are used for the IV queries and syllable-based transcriptions are used for the OOV queries. If α equals to one, only $Score_{BD}$ is used as the score at the second pass. On other hand, if α equals to zero, only $Score_{DDV}$ is used.

The result shows that the influence of the weight parameter is small for IV queries. But for OOV queries, the best performance is achieved by combining the two types of scores. Also, the result shows that the performance is superior to the baseline even at the point of $\alpha = 1$ or $\alpha = 0$. It suggests that the state-level matching across the subword unit is effective rather than using a subword-unit local distance for the STD task, even when the acoustic dissimilarity between distributions are identically estimated.

V. Conclusions

In this paper, we introduced new acoustic dissimilarity measure for subword-unit HMM and proposed a two-pass spoken term detection method in which the state-level acoustic dissimilarity is effectively incorporated into the scoring process. The new acoustic dissimilarity measure is based on a distance vector representation and the elements consist of the distance between a distribution and other all distributions which correspond to a set of states in subword-unit HMMs. The experimental results with NTCIR-9 SpokenDoc STD subtask showed that our proposed method was significantly improved compared with baseline methods which use only subword-level local acoustic dissimilarity measure.

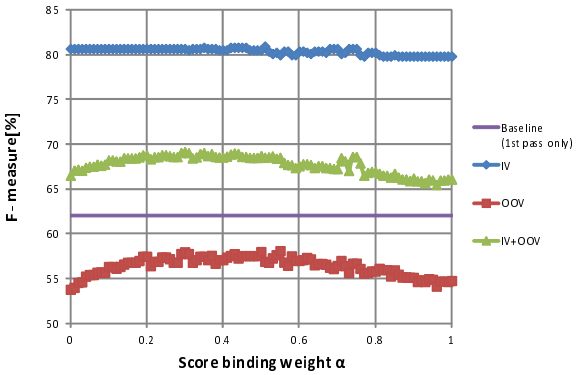

Fig. 3. Effect of the score weight parameter α (STD system with $Score_{DDV\_L1Max}$). The curve "IV" and "OOV" show the breakdown of STD performance for in-vocabulary queries and out-of-vocabulary queries, respectively.

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 methods [15] are necessary to achieve the further improvement and the robustness for the spoken documents in the real world.

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).

[8] 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).

[9] Y. Zhang and J. R. Glass: "Unsupervised Spoken Keyword Spotting via Segmental DTW on Gaussian Posteriorgrams," Proc. of ASRU, pp.398-403 (2009).

[10] 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).

[11] H. Lee, et al.: "Open-Vocabulary Retrieval of Spoken Content with Shorter/Longer Queries Considering Word/Subword-based Acoustic Feature Similarity," Proc. of Interspeech (2012).

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

[13] T. Murakami et al.: "Japanese vowel recognition based on structural representation of speech," Proc. of EUROSPEECH, pp.1261-1264 (2005)

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

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