YLAB@RU at Spoken Term Detection Task in NTCIR-10 SpokenDoc-2

Iori Sakamoto
Ritsumeikan University
1-1-1 Noji-higashi,
Kusatsu-shi,
Shiga, 525-8577, Japan
is019083@ed
.ritsumei.ac.jp

Kook Cho
Ritsumeikan University
1-1-1 Noji-higashi,
Kusatsu-shi,
Shiga, 525-8577, Japan
cho@slp.is
.ritsumei.ac.jp

Masanori Morise*
Ritsumeikan University
1-1-1 Noji-higashi,
Kusatsu-shi,
Shiga, 525-8577, Japan
morise@media
.ritsumei.ac.jp

Yoichi Yamashita
Ritsumeikan University
1-1-1 Noji-higashi,
Kusatsu-shi,
Shiga, 525-8577, Japan
yama@media
.ritsumei.ac.jp

ABSTRACT

The development of spoken term detection (STD) techniques, which detect a given word or phrase from spoken documents, is widely conducted in order to realize easy access to large amount of multimedia contents including speech This paper describes improvement of the STD method which is based on the vector quantization (VQ) and has been proposed in NTCIR-9 SpokenDoc. Spoken documents are represented as sequences of VQ codes, and they are matched with a text query to be detected based on the V-P score which measures the relationship between a VQ code and a phoneme. The matching score between VQ codes and phonemes is calculated after normalization for each phoneme in a query term to avoid biased scoring to particular phonemes. The score normalization improves the STD performance by 6% of F measure.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

spoken term detection, out-of-vocabulary, vector quantization, score normalization

Team Name

YLAB

Subtasks / Languages

Spoken Term Detection / Japanese

*Dr. Morise currently works for Yamanashi University.

External Resources Used

CSJ (Corpus of Spontaneous Japanese)

1. INTRODUCTION

Rapid increase of multimedia contents including spoken messages is raising the need of information retrieval for speech to facilitate to access the spoken documents that we want. Spoken term detection (STD) is a task of the information retrieval for speech data and finds words or phrases which match with a given query term[2]. STD can be accomplished by the combination of two techniques, automatic speech recognition (ASR) and text search. This simple approach is not sufficient because ASR can not recognize speech data completely without recognition errors and can not recognize out-of-vocabulary (OOV) words which are not contained in the ASR dictionary. One of important issues in STD is how to detect OOV words[8].

Several methods have been proposed to avoid the OOV word problem. One of promising approaches is the use of speech recognition based on sub-word units, such as phonemes and syllables[5, 11]. Speech segments are detected by matching between two sub-word sequences which are obtained from input text query and speech recognition. In this approach, the speech recognition converts spoken documents into sub-word sequences in a vocabulary-free manner, and can avoid the OOV word problem because a set of sub-word units cover all words or sentences. Theoretically, any word can be recognized correctly. However, the accuracy of the speech recognition based on sub-word units is lower than word-based speech recognition.

Another approach is a word spotting technique which has been widely studied in the 1980's[4, 6, 7, 13] in order to realize speech understanding based on keyword recognition rather than information retrieval. The word spotting based on Hidden Markov Model (HMM) detects speech segments similar to a query term by calculating the likelihood for sequences of acoustic parameters of the speech segment. The

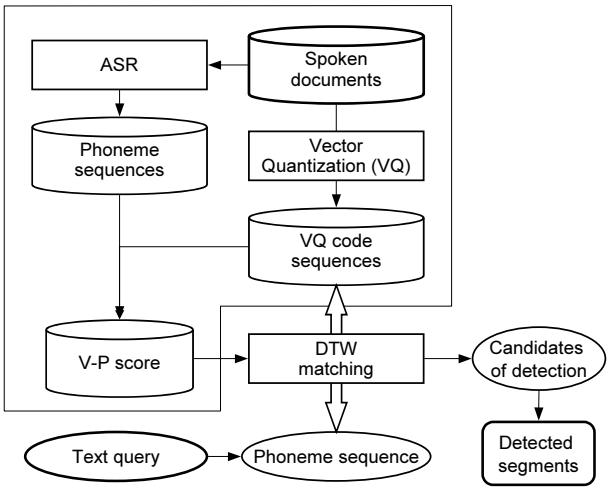

Figure 1: The overview of the STD method based on Vector Quantization.

word spotting takes much time to calculate the likelihood while the accuracy is high. This characteristics of the word spotting is not appropriate to the on-line STD task for large database of spoken documents.

Representation scheme of spoken documents is a crucial issue for STD with high accuracy. The sub word is one of symbolic representations of spoken documents. Some acoustic properties are possibly missed in the process of conversion from an acoustic parameter sequence into a sub-word symbol. On the other hand, the acoustic parameters have full acoustic properties of speech, but they consume much time for the STD task. The authors have proposed a method of STD based on vector quantization (VQ) which uses the sequence of VO codes as an alternative representation scheme of spoken documents for the STD[12, 10]. A query term is detected by matching between the VQ sequence and input text query by defining the cooccurrence score between phonemes and VQ codes for each speaker. This paper describes a normalization method of matching score which evaluates equally phonemes in the query to reject candidates with unnatural structure of phoneme durations.

2. STD METHOD BASED ON VECTOR QUANTIZATION

The overview of the STD method based on VQ is shown in Figure 1. The VQ process converts spoken documents in the database into sequences of VQ codes by a clustering technique for each speaker. The automatic speech recognition (ASR) also converts the spoken documents into sequences of phonemes with time alignment information. The V-P score, which is defined as a cooccurrence score of a phoneme for a VQ code, is trained for each speaker-dependent VQ codebook. The continuous DTW matching technique compares the VQ code sequences of spoken documents with the phoneme sequence of input text query, and detects segments using some threshold logics.

2.1 Vector Quantization of Acoustic Parameters

Spoken documents are analyzed with 20 [ms] frames and 10 [ms] intervals. MFCC parameter of 12 dimensions are obtained for each frame. The feature vector of a frame consists of 60 parameters including 12 MFCCs of two preceding and two following frames as well as the current frame. The VQ process converts a 60-dimensional parameter vector into a VQ code frame by frame.

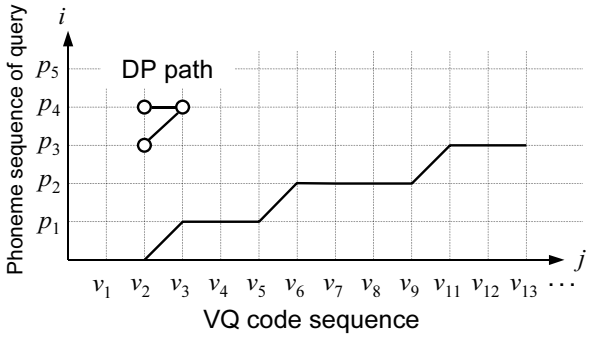

Figure 2: A sample of DTW matching.

2.2 V-P score: cooccurrence score of a phoneme for a VQ code

The V-P score, which is the cooccurrence score of a phoneme for a VQ code, is beforehand trained to compare VQ code sequences of spoken documents with the input query. The V-P score $s(v,p)$ of a phoneme p for a VQ code v is defined based on the occurrence count of the phoneme for the VQ code, as

$$s(v,p) = \log\left(\frac{C_v(p)}{N_v}\right) - \log\left(\frac{C_v(p_{best})}{N_v}\right), \quad (1)$$

where $C_v(p)$ is the number of frames which are labeled with the phoneme p and is quantized into the VQ code v, N_v is the total number of frames of v, and p_{best} is the phoneme which appears most in v.

2.3 Continuous DTW matching

The continuous DTW matching compares the VQ code sequences with the phoneme sequence of the input query. Figure 2 shows a sample of DTW matching and the DP path. The V-P score is used as a local score between a VQ code and a phoneme. Let $p_j (1 \le j \le K)$ and K be the phoneme sequence of the input query and the number of the phoneme in a query. Let $v_j (1 \le i \le L)$ and L be a VQ code sequence and the number of the VQ code in a spoken documents. The maximum accumulated score $S_{i,K}$ at the frame i for the input query is calculated as follows.

1) $S_{0,j} = 0.0 \quad (1 \le j \le K)$ (2)

2) repeat 3),4),5) for $i = 1, 2, \cdots, L$

3) $S_{i,0} = 0.0$

4) repeat 5) for $j = 1, 2, \cdots, K$

5) $$S_{i,j} = \begin{cases} S_{i-1,j-1} + s(v_i, p_j) & (\bar{S}_{i-1,j-1} > \bar{S}_{i-1,j}) \\ S_{i-1,j} + s(v_i, p_j) & (\bar{S}_{i-1,j-1} \le \bar{S}_{i-1,j}) \end{cases} \quad (3)$$

The $\bar{S}_{i,j}$ is a duration-normalized accumulated intermediate score and is defined as

$$\bar{S}_{i,j} = \frac{1}{i - start(i,j) + 1} S_{i,j}, \quad (4)$$

where $start(i, j)$ indicates the starting frame which is determined by backward tracking from the matching between the i-th phoneme and the j-th VQ code. The continuous DTW matching calculates the duration-normalized accumulated score $\bar{S}(i)(= S_{i,K})$. If $\bar{S}(i)$ shows a local maximum and it is larger than a threshold, the segment from $start(i, K)$- to i-frames is a candidate of detection.

2.4 Re-evaluation of Candidates Considering Duration Structure

Preliminary experiments uncovers that the term detection only using a threshold of $\bar{S}(i)$ generates many false detections, which include

(1) Too small number of frames match with a phoneme in the term.

(2) A few phonemes in the term occupies most of the detected segment.

Speech segment candidates detected by a threshold logic for matching score $\bar{S}(i)$ are re-evaluated in terms of duration of the phonemes.

First of all, a candidate segment is removed if the frame length of a phoneme in it is lower than a threshold. The threshold is set for each phoneme based on the average frame length of the phoneme.

Secondly, the distortion of phoneme duration in a candidate segment is calculated and candidates are re-evaluated by the unified score which incorporates both matching score and naturalness of phoneme duration structure.

2.4.1 *Distortion of Phoneme Duration*

The distortion of phoneme duration, $D(i)$, is defined as

$$D(i) = \frac{1}{K}\sum_{j=1}^{K}\left(\frac{d_l(p_j)}{L_l} - \frac{d_d(p_j)}{L_d}\right)^2, \tag{5}$$

where $d_l(p)$ is the average frame length of a phoneme p which is obtained by speech recognition results of the same speaker, and $d_d(p)$ is the frame length of a phoneme p in the detected segment. The estimated total duration L_l of the term is given by

$$L_l = \sum_{j=1}^{K} d_l(p_j). \tag{6}$$

The actual total duration L_d of the detected segment is given by

$$L_d = \sum_{j=1}^{K} d_d(p_j). \tag{7}$$

2.4.2 *Unified Score*

The unified score evaluates candidates for detection in terms of both matching and naturalness of duration structure. The matching score $\bar{S}(i)$ is already normalized by the total duration of the candidate segment as shown in the equation (4) and is is again normalized based on the statistical distribution of matching scores for all candidate segments. The statistically normalized matching score $P_{\bar{S}}(i)$ is calculated as follows.

$$P_{\bar{S}}(i) = \frac{\bar{S}(i) - \mu_{\bar{S}}}{\sigma_{\bar{S}}}, \tag{8}$$

$$\mu_{\bar{S}} = \frac{1}{N}\sum_{i=1}^{N}\bar{S}(i), \tag{9}$$

$$\sigma_{\bar{S}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{S}(i) - \mu_{\bar{S}})^2}, \tag{10}$$

where N is the number of candidate segments. For the distortion of phoneme duration $D(i)$, the statistically normalized duration distortion $P_D(i)$ is calculated in the same manner. The unified score which evaluates candidates is defined as

$$P(i) = P_{\bar{S}}(i) - P_D(i). \tag{11}$$

The final decision of detection is executed by threshold logic for the unified score $P(i)$.

3. PROPOSED METHOD: INTRA-PHONEME NORMALIZATION

The STD method which is described in Section 2 uses the vector quantization and calculates the matching score between VQ codes and phonemes. In usual STD methods, the HMM-based decoder converts speech into a sequence of phonemes or words considering time-variable properties of speech. The VQ process independently converts a feature vector of acoustic parameters into a VQ code frame by frame. A sequence of VQ codes for a spoken document has weak constraints on time structure although the feature vector consists of a segment of 5 frames. Our proposed STD method more likely generates false detections with unnatural time structure. In order to reduce such false detections, we introduced the additional metric measuring unnaturalness of duration structure for candidates, which is mentioned in 2.4.1. However, our method still detects some false segments with unnatural duration structure.

In this paper, we propose a new method of calculating matching score for more reduction of false detection. The new matching score is defined as

$$\bar{N}(i) = \frac{1}{K}\sum_{j=1}^{K}\left(\frac{1}{N}\sum_{m=m_j}^{m_j+n_j-1} s(v_m, p_j)\right), \tag{12}$$

where n_j is the number of frames matching with j-th phoneme of the query term, and m_j is the starting frame of the segment of j-th phoneme. This equation executes an intra-phoneme normalization by averaging scores within a phoneme to avoid that some particular phonemes yield too more effects on the total score. The new STD method uses this score for calculating the unified score, mentioned in 2.4.2, by replacing $\bar{S}(i)$ by $\bar{N}(i)$ and statistically normalizing $\bar{N}(i)$ in the same manner as 2.4.2 to get $P_{\bar{N}}(i)$. The new final score is defined as

$$P(i) = P_{\bar{N}}(i) - P_D(i). \tag{13}$$

4. EVALUATION

4.1 Experiment Setup

The proposed method was evaluated on 177 spoken lectures in the CORE set of the Corpus of Spoken Japanese (CSJ)[9]. Each lecture in the CSJ is divided in segments, called Inter-Pausal Unit (IPU), by the pauses that no shorter

Table 1: Performance of STD methods.

	F-measure[%]	MAP[%]
baseline	60.9	50.0
old version	59.3	61.1
proposed	65.9	67.5

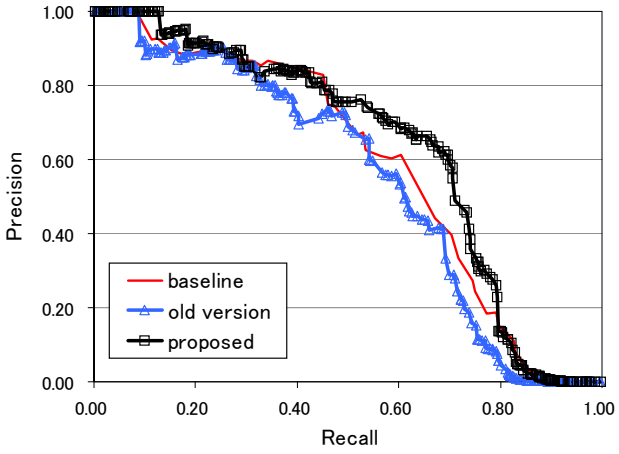

Figure 3: Recall and precision of STD methods.

than 200 [ms]. IPUs detected by proposed methods are judged whether the IPUs include a specified query term or not, with the same measure as the formal run of the STD task[2]. The query terms are 50 terms in the STD test collection which were proposed by the Spoken Document Processing Working Group[1]. The size of VQ codebook is 1024 for all speakers.

Our STD method based on VQ needs phoneme label information of spoken documents to define the V-P score. We used the 1-best results of continuous phoneme recognition which were provided by the task organizer of NTCIR-9 SpokenDoc[3]. The evaluation measures are F-measure and mean average precision(MAP).

4.2 Evaluation Results

Table 1 and Figure 3 compare evaluation results for several STD methods. The baseline STD is based on matching between two phoneme sequences, the phoneme sequence of input query and the phoneme sequences recognized by ASR for spoken documents, using the edit distance as local score of DTW. The old version and proposed STD are the methods described in Section 2 and 3.

The performance of the old version is comparable to the baseline method. The proposed method improves the performance by about 6% to other two methods.

4.3 Formal Run Result

For the formal run of SDPWS[2], we trained the V-P score using the 1-best results of REF-SYLLABLE-UNMATCHED, which are produced by ASR with the syllable-based trigram language model trained with unmatched language resources.

Figure 4 and Table 2 show the STD performance for SDPWS data in the formal run[2]. We compared the proposed method with the baseline performance BL-3 which was provided by the task organizer and detected query terms based on the DP-based detection for two kinds of phoneme sequences generated by the continuous syllable recognition and the continuous word recognition with matched language models[2].

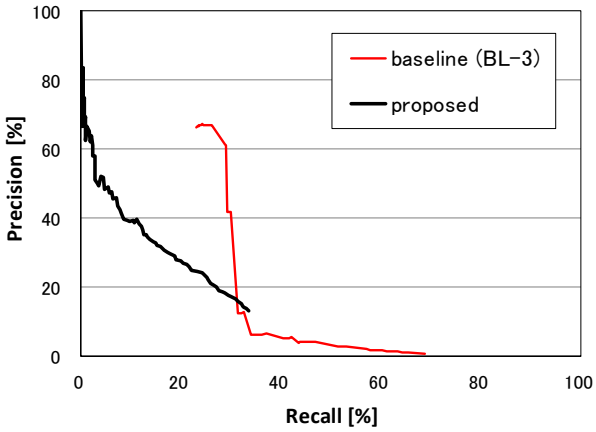

Figure 4: Recall and precision for SDPWS data in the formal run.

Table 2: STD Performance for SDPWS data in the formal run.

	F-measure[%]		MAP[%]
	(max)	(spec.)	
baseline(BL-3)	39.36	39.16	39.3
proposed	24.10	24.04	22.1

The performance is degraded in comparison with the results mentioned in 4.2 not only for the proposed method but also for the baseline. We guess that the accuracy of speech recognition is lower for SDPWS data rather than for CSJ data. Our STD method expects that the accuracy of speech recognition is so high that the V-P score can be trained with high accuracy. The low performance of speech recognition degrade our method based on VQ more than the baseline method of phoneme matching.

5. INEXISTENT SPOKEN TERM DETECTION

We conducted the inexistent spoken term detection (iSTD) that is a newly introduced in NTCIR-10 SpekenDoc-2 and is a task of judging whether a query term is existent or inexistent[2]. The collection of spoken documents is the same as data in 4.1 and consists of 177 spoken lectures in the CORE set of CSJ. the measure for judging the existence of the term is the same as the unified score described in Section 3. Figure 5 shows our result of iSTD. The maximum F-measure is 73.3%.

6. CONCLUSIONS

This paper describes a method of normalizing a matching score for VQ-based STD which we had proposed. The intra-phoneme normalization which averages scores for each phoneme in a query term improves the performance of STD by about 6% of F-measure and MAP. However, the proposed method shows the low performance for SDPWS data in the formal run. We conducted the iSTD for CSJ data and ob-

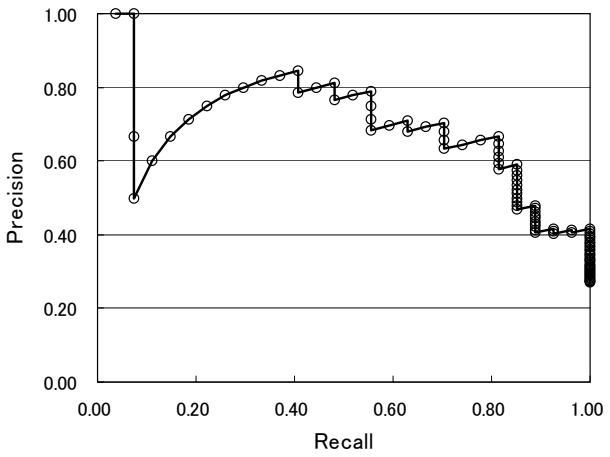

Figure 5: Recall and precision of iSTD.

tained the maximum F-measure of 73.3%. The automatic definition of the threshold is a future work for iSTD.

7. REFERENCES

[1] T. Akiba, K. Aikawa, Y. Itou, T. Kawahara, H. Nanjo, H. Nishizaki, N. Yasuda, Y. Yamashita, and K.Iou. Developing an sdr test collection from japanese lecture audio data. In Proceedings of the 1st Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC2009), page 6, 2009.

[2] T. Akiba, H. Nishizaki, K. Aikawa, X. Hu, Y. Itoh, T. Kawahara, S. Nakagawa, H. Nanjo, and Y. Yamashita. Overview of the ntcir-10 spokendoc2 task. In Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies, 2013.

[3] T. Akiba, H. Nishizaki, K. Aikawa, T. Kawahara, and T. Matsui. Overview of the ir for spoken documents task in ntcir-9 workshop. In Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2011.

[4] E. M. Hofsteller and R. C. Rose. Techiniques for task independent word spotting in continuous speech messages. In Proceedings of ICASSP 1992.

[5] Y. Itoh, T. Otake, K. Iwata, K. Kojima, M. Ishigame, K. Tanaka, and S. wook Lee. Two-stage vocabulary-free spoken document retrieval —subword identification and re-recognition of the identified sections—. In Proceedings of INTERSPEECH 2006, pages 1161–1164, 2006.

[6] G. J. F. Jones, J. T. Foote, K. Sparck, and S. J. Young. Video mail retrieval: The effect of word spotting accuracy on precision. In Proceedings of ICASSP 1995, pages 309–312, 1995.

[7] K. M. Knill and S. J. Young. Fast implementation methods for viterbi-based word-spotting. In Proceedings of ICASSP 1996, pages 522–525, 1996.

[8] B. Logana and J. M. V. Thong. Confusion-based query expansion for oov words in spoken document retrieval. In Proceedings of ICSLP 2002, pages 1997–2000, 2002.

[9] K. Maekawa, H. Koiso, S. Furui, and H. Isahara. Spontaneous speech corpus of japanese. In Proceedings of LREC, pages 947–952, 2000.

[10] T. Matsunaga, K. Cho, and Y. Yamashita. Spoken term detection using the segment quantization of acoustic features of spoken documents. In Proceedings of the 6th Spoken Document Processing Workshop, SDPWS2012-7, pages 1–7.

[11] T. Mertens, R. Wallace, and D. Schneider. Cross-site combination and evaluation of subword spoken term detection systems. In Proceedings of Content-Based Multimedia Indexing (CBMI), pages 61–66, 2011.

[12] Y. Yamashita, T. Matsunaga, and K. Cho. Ylab@ru at spoken term detection task in ntcir9-spokendoc. In Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, pages 287–290, 2011.

[13] Y. Yamashita and R. Mizoguchi. Keyword spotting using f0 contour matching. In Proceedings of 5th Conference on Speech Communication and Technology (Eurospeech '97), volume 1, pages 271–274, 1997.