YLAB@RU at Spoken Term Detection Task in NTCIR-9

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

Toru Matsunaga
Ritsumeikan University
1-1-1 Noji-higashi,
Kusatsu-shi,
Shiga, 525-8577, Japan
cm014063@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

ABSTRACT

The information retrieval based on speech recognition is an important technique to easy access to large amount of multimedia contents including speech. The development of spoken term detection (STD) techniques, which detect a given word or phrase from spoken documents, is widely conducted. This paper proposes a new method of STD based on the vector quantization (VQ). 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 representation of VQ codes is an intermediate form between acoustic features such as MFCC parameters and sub-word symbols which are often used in conventional STD methods. The dependency of acoustic features on a speaker is avoided by the speaker-dependent VQ.

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

1. INTRODUCTION

Rapid increase of multimedia contents including spoken messages is raising the need of information retrieval for speech data to facilitate to access the data 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. SDR 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 SDR is how to detect OOV words[6].

Several methods have been proposed to avoid the OOV words problem. One of promising approaches is the use of speech recognition based on sub-word units, such as phonemes and syllables[3, 8]. Speech segments are detected by matching between two sub-word sequences which are obtained by input text query and speech recognition. In this approaches, the speech recognition converts speech into a sub-word sequence in a vocabulary-free manner, and can avoid the OOV word problem because a set of sub-word units cover all words. 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[2, 4, 5, 9]. The word spotting based on Hidden Markov Model (HMM) detects a speech segment similar to query text by calculating the likelihood for a sequence of acoustic parameters of the speech segment. The 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 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. In this paper, vector quantization (VQ) sequence is used as an alternative representation scheme of spoken documents for the STD. A new STD method detects speech segments of the term based on matching between the VQ sequence and input text query by defining the cooccurrence score between phonemes and VQ codes for each speaker.

2. METHOD

The flowchart of the proposed method is shown in Figure 1. The VQ process converts spoken documents in the database into sequences of VQ codes by a clustering technique. The spoken documents are also recognized by the

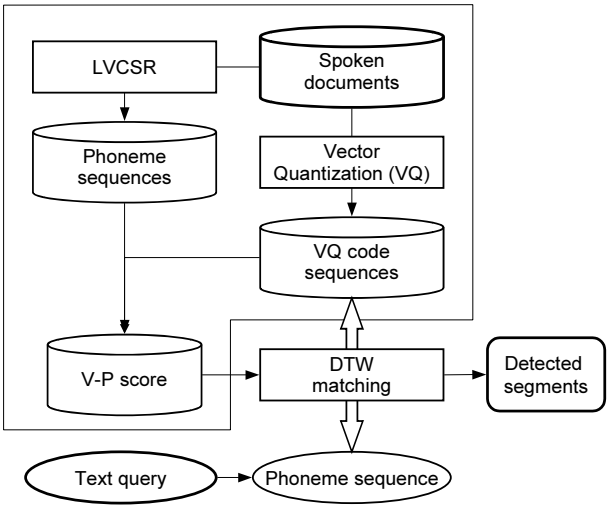

Figure 1: A flowchart of the proposed method.

large vocabulary continuous speech recognition (LVCSR). The recognized word sequences are automatically converted into sequences of phonemes. 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 msec frames and 10 msec intervals. MFCC parameter of 12 dimensions are obtained for each frame. The VQ process uses 24-dimensional parameters that are 12 MFCC parameters and the delta parameters as the feature vector of the frame.

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

$$s(v,p) = \log\left(\frac{C_v(p)}{N_v}\right) - \log\left(\frac{C_v(p_{best})}{N_v}\right) + 2.0, \qquad (1)$$

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 The 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 $p_j (1 \leq j \leq K)$ and K be the phonese sequnce of the input query and the number of the

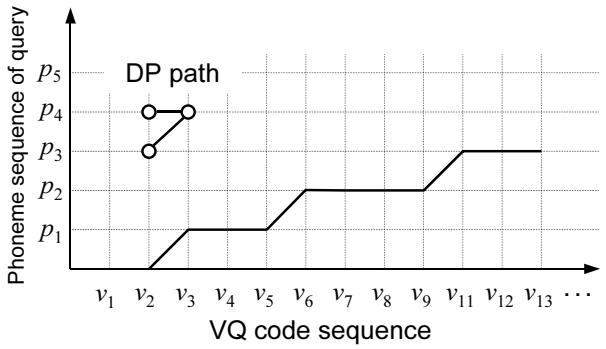

Figure 2: A sample of DTW matching.

phoneme in a query. Let $v_j (1 \leq i \leq L)$ 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) $S_{0,j} = 0.0 \quad (1 \leq j \leq K)$ (2)

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

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

4) $S_{i,j} = \max\left\{\begin{matrix} S_{i-1,j-1} \\ S_{i-1,j} \end{matrix}\right\} + s(v_i, p_j) \quad (1 \leq j \leq K)$ (3)

2.4 Term Detection

The continuous DTW matching calculates the score $S_{i,K}$ and determines the starting frame $start(i)$ of the matching segment frame by frame. The normalized score $\bar{S}(i)$ for the segment terminated by the i-th frame is given by

$$\begin{aligned} \bar{S}(i) &= \frac{1}{i - start(i) + 1} S_{i,K} \\ &= \frac{1}{i - start(i) + 1} \sum_{t=start(i)}^{i} s(v(t), p_i(t)), \end{aligned} \qquad (4)$$

where $v(t)$ is the VQ code of t-th frame and $p_i(t)$ is the phoneme that matches with t-th frame in the matching terminated by i-frame. If $\bar{S}(i)$ shows a local maximum and it is latger than a threshold, the segment from $start(i)$- to i-frames is a candidate of detection.

Preliminary experiments uncovers that the term detection only using a threshold for $\bar{S}$ 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 segments detected by a threshold logic for $\bar{S}$ are re-evaluated by three heuristic rules, C1, C2, and C3, described below to remove false detections.

2.4.1 *C1: The condition for the frame length of a phoneme*

A detected 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.

2.4.2 C2: The variance of matching scores

A detected segment is removed if the variance of the matching score between VQ codes and phonemes is larger than a threshold. The variance $V_S(i)$ for the segment terminated by i-frame is defined as

$$V_S(i) = \frac{1}{i - start(i) + 1} \sum_{t=start(i)}^{i} \left(s(v(t), p_i(t)) - \bar{S}(i) \right)^2 . \quad (5)$$

2.4.3 C3: The duration of a phoneme

A detected segment is removed if the duration of a phoneme in it is larger than a threshold. The duration is normalized by the total duration of the term. The duration difference $V_D(i)$ is defined as

$$V_D(i) = \frac{1}{K} \sum_{j=1}^{K} \left(\frac{D_l(p_j)}{L_l} - \frac{D_d(p_j)}{L_d} \right)^2 , \quad (6)$$

where $D_l(p)$ is the average frame length of a phoneme p which is obtained by speech data with correct phoneme labels, 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

$$L_l = \sum_{j=1}^{K} D_l(p_j). \quad (7)$$

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

$$L_d = \sum_{j=1}^{K} D_d(p_j). \quad (8)$$

2.5 Preconditions of the proposed method

The proposed method requires the speaker-dependent VQ codebook and the relationship between phonemes and VQ codes. This approach supposes that

1) The speaker of spoken documents to be retrieved is known.
2) Large amount of spoken documents are available for each speaker.
3) Automatic speech recognition gives high accuracy.

These preconditions are almost satisfied for lecture speech and spoken blog on the Web.

3. EVALUATION

3.1 Experiment Setup

The proposed method was evaluated on 177 spoken lectures in the CORE set of the Corpus of Spoken Japanese (CSJ)[7]. Each lecture in the CSJ is divided in segments, called Inter-Pausal Unit (IPU), by the pauses that no shorter than 200mssec. IPUs detected by proposed methods are judged whether the IPUs include a specified term or not, with the same measure as the formal run of the STD task. We selected 20 long words as query terms from the STD test collection which was proposed by the Spoken Document Processing Working Group[1]. The size of VQ codebook is 4096 for all speakers.

Table 1: STD results for different heuristic conditions.

$S(i)$	C1	C2	C3	recal	prec.	F-meas.
○	-	-	-	0.409	0.005	0.013
○	○	-	-	0.623	0.007	0.015
○	-	○	-	0.345	0.662	0.454
○	-	-	○	0.221	0.804	0.346
○	○	○	-	0.559	0.332	0.417
○	○	-	○	0.317	0.760	0.447
○	-	○	○	0.480	0.721	0.577
○	○	○	○	0.509	0.773	0.613

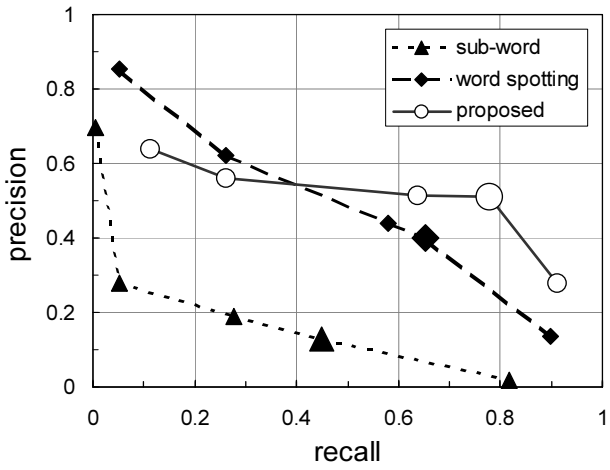

Figure 3: Recall and precision of STD methods.

The proposed method needs phoneme label information of spoken documents to define the V-P score. The phoneme labels should be automatically generated by the ASR system. In this evaluation, in order to investigate the potential performance of the proposed method, we compare two kind of label information, the correct labels by hand and the force alignment results by ASR. The evaluation measures are precision, recall, and F-measure.

3.2 Evaluation Results

Table 1 lists evaluation results for combination of heuristics. The term detection using only the average score $\bar{S}(i)$ generated many false detections and results in vary low precision rate. The introduction of three heuristic rules drastically improves the performance. The C2 condition is most effective.

Figure 3 and Table 2 compare the performance of the proposed method for conventional STD methods. The 'sub-word' STD is based on matching between two phoneme sequences, the phoneme sequence of input query and the phoneme sequences recognized by ASR for spoken docu-

Table 2: Comparison of STD methods.

	F-measure	time[min]
sub-word	0.225	0.50
word spotting	0.501	240.20
proposed	0.613	1.66

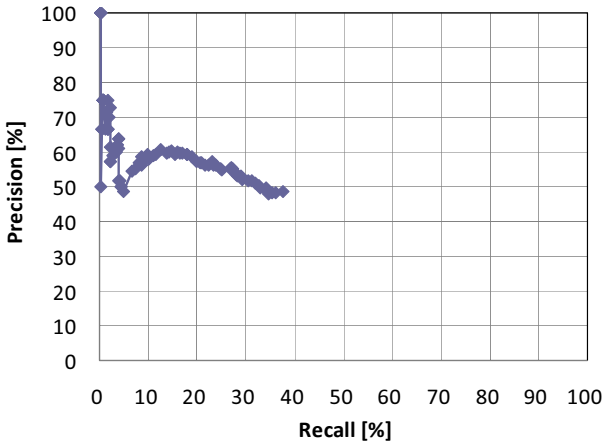

Figure 4: Recall and precision in the formal run.

Table 3: Performance of the proposed method.

F-measure(max)	F-measure(spec.)	MAP
0.425	0.425	0.344

ments. ASR was carried out with a monophone acoustic model and a language model of the syllable trigram. The word spotting is based on the likelihood calculated by phoneme HMMs of the monophone. In Figure 3, large marks indicate the maximum points of F-measures. The proposed method consumes more process time for the sub-word method and drastically reduces process time for the word spotting. The F-measure of the proposed method is comparable or improved to the word-spotting and much better than the sub-word method.

3.3 Formal Run Result

Figure 4 and Table 3 show the performance of the proposed method in the formal run. The performance is degraded in comparison with the results mentioned in 3.2 because the phoneme sequences, which are used to define the V-P score and are not used for matching with an input query, are obtained by ASR in the formal run evaluation. In Figure 4, some results does not appeared for low precision and high recall rate because we submitted the data of IPUs which were detected by the definition of our system and did not include results for relaxed thresholds.

4. CONCLUSIONS

This paper describes a new STD method based on the VQ representation of spoken documents. We are going to try to improve the performance by considering the definition of the VP score and introducing VQ for speech segments. The proposed method is conceptually similar to the word spotting based on VQ coding of spoken documents and the discrete HMM. Future plans include a comparison to the word spotting based on the discrete HMM.

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

[3] 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.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] 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.