COMBINING MULTIPLE SUBWORD REPRESENTATIONS FOR OPEN-VOCABULARY SPOKEN DOCUMENT RETRIEVAL

Shi-wook Lee[1], *Kazuyo Tanaka*[2], *Yoshiaki Itoh*[3]

[1]National Institute of Advanced Industrial Science and Technology (AIST)
s.lee@aist.go.jp
[2]Institute of Library and Information Science, University of Tsukuba
ktanaka@ulis.ac.jp
[3]Faculty of Software and Information Science, Iwate Prefectural University
y-itoh@iwate-pu.ac.jp

ABSTRACT

This paper describes subword-based approaches for open-vocabulary spoken document retrieval. First, the feasibility of subword units in spoken document retrieval is investigated, and our previously proposed sub-phonetic segment units are compared to typical subword units, such as syllables, phonemes, and triphones. Next, we explore the linear combination of retrieval score from multiple subword representations to improve retrieval performance. Experimental evaluation of open-vocabulary spoken document retrieval tasks demonstrated that our proposed sub-phonetic segment units are more effective than typical subword units, and the linear combination of multiple subword representations resulted in a consistent improvement in the F-measure.

1. INTRODUCTION

Information retrieval has developed remarkably in recent years with the expansion of the World Wide Web and the improvement of mass-storage devices, enabling text databases to identify documents that are likely to be relevant to text queries. Since the accessible multimedia data containing spoken information has considerably increased, the demand for automatic retrieval methods has increased significantly. Such spoken documents, stored in the form of audio signals, could be recorded from various sources, such as news broadcasts on radio and television, voice/video e-mail, and multimedia material on the Web. Therefore, it has become necessary to be able to retrieve such spoken documents in response to a query.

The function of spoken document retrieval (SDR) is to retrieve the information content of multimedia data using a combination of automatic speech recognition and information retrieval techniques. There have been several different approaches for SDR. First, the keyword spotting technique is adopted for spoken documents to obtain a set of keyword transcriptions. Another approach is to use a large vocabulary continuous speech recognition (LVCSR) system to generate words, and then conduct conventional text retrieval. This word-based approach has been very popular and successful. In particular, research on the word-based approach has been promoted by the spoken document retrieval tracks of the Text REtrieval Conference (TREC) [1][2]. In this word-based approach, the main issue is to improve the performance of speech recognition so that the resulting text takes a lead role in the text retrieval process. Although this approach is successful and straightforward, it also has several difficulties. Contrary to the text retrieval, we must consider how to deal with the errors occurring from speech recognition in the text retrieval process in SDR. Another difficulty is vocabulary growth, since new words are introduced continuously from growing multimedia collections. These Out-Of-Vocabulary (OOV) words are not correctly recognized, and are therefore deleted or substituted. Furthermore, it leads to more significant problems when query words are OOV. With current LVCSR systems, there is a practical limitation of the vocabulary size. Although OOV words do not present a significant problem in the real task domain that is concluded from the TREC SDR task [1][2], further research is still needed to remove the effect of OOV and realize the ultimate results in SDR. Several researchers have reported on strategies to minimize the effect of the OOV words. The first approach, query expansion, is a technique where query terms are automatically expanded or modified with additional query information from the collection. However, the use of subword representations for spoken document retrieval is helpful to solve the OOV word problem. One way is to use phoneme sequences generated by a phoneme recognizer.[3] Here, they introduced Probabilistic String Matching (PSM), where query words are spotted in document phoneme sequences that are corrupted by recognition error. In [4], phone n-grams are used as the indexing terms. This previous research indicated that a subword-based approach could be an optimal solution to the OOV word problem. Another merit of using subword units in the recognition is that the size of vocabulary needed to cover the language can be adjusted. The subword-based SDR is attractive even though it is outperformed by word-based systems based on subword units.[1][2] However, the main problem with subword units is the high rate of subword

recognition errors, compared to word-based recognition. There is also a trade-off between the size of the subword unit and its recognition accuracy. Therefore, the choice of subword units needed to perform effective SDR and its feasibility study are essential. Our approach here is based on subword sequences generated by a post-processing subword recognizer and their sequential matching by the following Shift-Continuous Dynamic Processing (Shift-CDP). Since the system is based on matching subword sequences directly, the system is not constrained in terms of vocabulary or grammar, and is robust with respect to recognition error. The paper first discusses the feasibility of several subword units in an SDR system. The use of a proposed subword unit, Sub-Phonetic Segment [5], is explored and compared through experimental evaluation. In the experiments, both text and speech are accepted as query input in Japanese SDR tasks. Finally, the linear combination of information from different subword units is further adopted to improve retrieval performance.

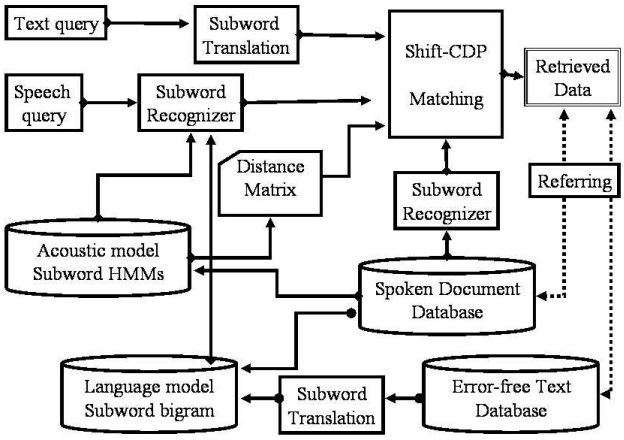

Fig. 1. *Block diagram of proposed SDR system based on subword units*

2. SYSTEM OVERVIEW

The development of SDR is very similar to text retrieval, except for a number of difficulties in actual application such as the accurate detection of word boundaries, recognition errors, and acoustic mismatching. Unlike text retrieval, spoken document retrieval must deal with these transcription errors. For this reason, most of the research on SDR using LVCSR has recently focused on new techniques such as relevance feedback and query expansion. However, a spoken document database is assumed to include a small number of OOV words, such as names, places, and special technical terms. Such OOV words will be susceptible to poor retrieval performance due to errors in speech recognition. Furthermore, when a query is proven to be an OOV word, the retrieval fails. The SDR system proposed here seeks to automatically detect the words (or multiword phrases) embedded in a large corpus of spoken documents. First, the retrieval system represents spoken documents as a linear sequence of subwords. The Shift-CDP then performs a subword match between query terms and documents. Since a similarity between document and query is based on only subword sequences and used to score and rank the documents in retrieval, the system needs a fixed number of subwords as the vocabulary so that open-vocabulary retrieval tasks can be performed. The system also works effectively when the quality or environmental conditions of the input and stored speech data differ considerably, because the subword does not use acoustic models in the matching process. Fig. 1 presents the overall block diagram of the proposed SDR system.

3. SUBWORD UNITS

Here, we explore the feasibility of typical subword units for SDR. In current speech recognition, the typical subword units are phonemes, syllables, or triphones (context-dependent phonemes). Triphones are effective subword units for LVCSR due to the representation of co-articulation effects. The difficulty with triphones is the large number of parameters to be estimated against the limitation of training data. From this reason, we have proposed sub-phonetic segments (SPSs) as new subword units for SDR.[5] The SPSs are derived from phonemes, and refined under the consideration of acoustic co-articulatory effects. The advantage of training SPS models is that pronunciation variation is trained directly into the acoustic model, and does not need to be modeled separately in the vocabulary. A graphical description of typical subword units and SPSs, re-estimated from a phoneme sequence consisting of stationary and non-stationary segments, is given in Fig. 2.

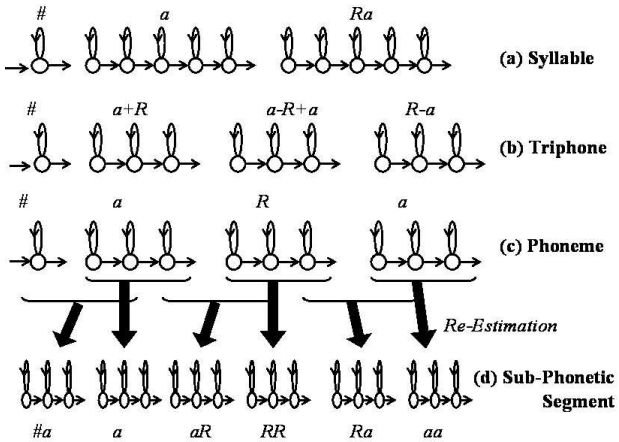

Fig. 2. *Graphical description of subword units and Sub-Phonetic Segmentation*

Japanese Newspaper Articles Sentences (JNAS)[6] are used for training the acoustic models of each subword unit. Also, to improve baseline subword recognition accuracy, subword bigram language models were estimated from the corpus used in training acoustic models. Theoretically, 1610

SPSs can be extracted from the 43 Japanese phonemes. However, some concatenations of phonemes do not exist in real language. The remarkably fewer Japanese SPSs is due to the fact that most Japanese syllables consist of 1 consonant and 1 vowel (C+V). Therefore, concatenations of consonants are very rare in Japanese. Table 1 summarizes the amount of training material and the number of each Japanese subword unit adopted in this work.

Table 1. *Training material used for acoustic and language models, and the number of each subword unit*

No. of sentence	28152
Length	52.07 hours
No. of syllables	225
No. of phonemes	43
No. of triphones	7241
No. of SPSs	411

4. SHIFT-CONTINUOUS DYNAMIC PROGRAMMING

When subword sequences are recognized directly (with higher error rates than for words), selection of a good matching approach becomes much more important. The previously proposed Shift-CDP is an algorithm that identifies similar parts between a reference pattern R_N and the input pattern sequence I_T.[7] The pre-fixed part of the reference pattern, called the unit reference pattern (URP), is shifted from the start point of the reference pattern to the end by a certain number of frames. The matching results for each URP in the reference pattern are then compared and integrated.

$$R_N = \{R_0, \cdots, R_\tau, \cdots, R_{\tau+r}, \cdots, R_{N-1}\} \quad (1)$$

$$I_T = \{I_0, \cdots, I_t, \cdots, I_{t+i}, \cdots, I_{T-1}\} \quad (2)$$

The first URP is taken from R_0 in the reference pattern R_N. The next URP is then composed of the same number of N_{URP} frames from the $(N_{shift}+1)_{th}$ frame. In the same way, the k_{th} URP is composed of N_{URP} frames from the $k \times (N_{shift}+1)_{th}$ frame. Thus, the number of URPs becomes $[N/N_{shift}]+1$, where [] indicates any integer that does not exceed the enclosed value. Shift-CDP is then performed for all URPs in the reference R_N. It is not necessary to normalize each cumulative distance at the end frame of a URP because all URPs are of the same length. Actually, Shift-CDP is a very simple and flat algorithm that performs CDP for each URP, and integrates the results.[7] The retrieved spoken documents are presented to the user in decreasing order of their DP score, given as follows.

$$G(i,r) = argmin \begin{cases} G(i-1,r-1) + D(s_i, s_r) \\ G(i-2,r-1) + D(s_i, s_r) \\ G(i-1,r-2) + 2 \cdot D(s_i, s_r) \end{cases} \quad (3)$$

$G(i,r)$ denotes the cumulative distance up to reference subword s_r and input subword s_i. $D(\cdot)$ is local distance, which uses a previously calculated distance matrix. Here, the distance measure D_{AB} between two multivariate Gaussian distributions, $N(\mu_A, \Sigma_A)$ and $N(\mu_B, \Sigma_B)$, is given as follows:

$$D_{AB} = \frac{1}{N} \sum_{n=1}^{N} (\mu_{An} - \mu_{Bn})^2 \left(\frac{\Sigma_{An} + \Sigma_{Bn}}{2} \right)^{-1} \quad (4)$$

where N is the number of HMM states.

4.1. Linear combination of multiple subword units

Different subword units can convey different types of information. Longer subword units can capture word or phrase information while shorter units can only model word fragments. The trade-off is that the shorter units are more robust to errors and word variants than the longer units. The lower value of $G(I,R)$ is ranked higher in retrieval results. Here, we assume that a highly reliable result is ranked high in both systems using a different single subword unit. In other words, the irrelevant document with a lower score that is ranked high in retrieval results might be ranked lower in the other outputs. With this assumption, we utilize the linear combination of the Shift-CDP score obtained from the different subword units, $G^n(I,R)$, where n denotes an individual subword unit.

$$G_c(I,R) = \sum_n w_n G^n(I,R) \quad (5)$$

$$\sum_n w_n = 1 \quad (6)$$

5. EXPERIMENTAL EVALUATIONS

As illustrated in Fig. 1, the proposed system can perform retrieval in response to both speech and text queries. All experiments are conducted using speech queries. A set of 10 key-phrase queries, uttered twice by 5 male speakers (100 $input\ queries$), are prepared to perform SDR evaluation experiments. Each spoken query has 9 relevant documents in a 2000 target database (3.29 $hours$). The underlying recognition system for decoding subwords is a single pass beam search decoder, which is based on the $Julius$ system.[8][9] Table 2 summarizes the recognition performance in each subword unit.

Table 2. *Baseline recognition performance for individual subword units; the value within parentheses indicates the recalculated result after triphone is converted into phoneme*

	Correct(%)	Accuracy(%)
SPS	72.35	65.56
Triphone	51.64 (77.98)	45.16 (71.01)
Monophone	73.26	70.70
Syllable	63.56	60.96

Recall and precision rates, which are commonly used in information retrieval, are used as evaluation measures. Also, the F-measure that takes into account both recall and precision is adopted.

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$$

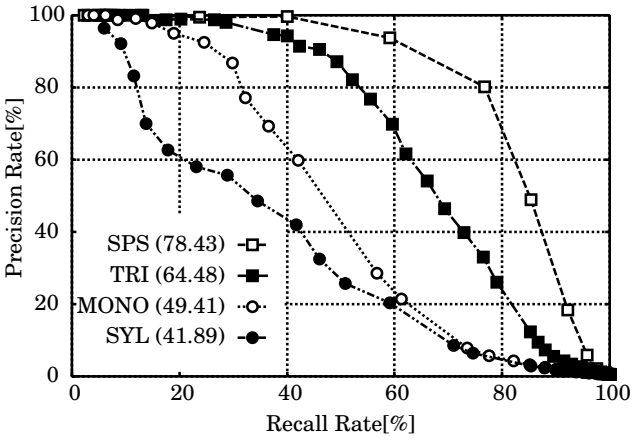

Fig. 3. *Comparison of the Recall-Precision curve for each subword unit; the value within parentheses indicates maximum F-measure*

Fig. 3 presents baseline SDR performance, recall-precision curves according to individual subword units, and their combination. From the results, the SPS-based SDR is remarkably superior to other systems. These results refer to the number of subwords, regarded as the quantity of information. The maximum F-measure, using a linear combination of subword units with the best-fitted weight, is plotted in Fig. 4. Comparing the performance of individual systems, the linear combination of subword units improves the retrieval performance, up to 89.07 for the F-measure. The linear combination of subword units should be helpful if the different subword units behave differently from each other. When the different subword units make different errors, combining the results provides an opportunity to improve the results, since some units are performing well. In addition, linear combination methods avoid the need to commit a priori to a single subword unit representation.

6. CONCLUSIONS

In this paper, we have described the development of an open-vocabulary spoken document retrieval system based on subword units. First, the proposed sub-phonetic segment units are presented and compared to typical subword units by experimental evaluation. Next, the linear combination of retrieval scores from multiple subword representations is explored to improve retrieval performance. The experiments confirmed that our proposed sub-phonetic segment units are more effective than typical subword units, and that the linear

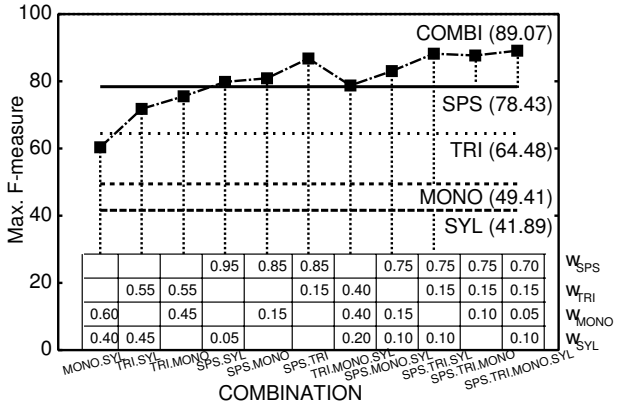

Fig. 4. *Comparison of maximum F-measure of linear combinations, by the best-fitted weighting value indicated in the inner table.*

combination of multiple subword representations can significantly improve the retrieval performance, up to 89.07 for the F-measure.

7. ACKNOWLEDGEMENT

This research was supported in part by Grant-in-Aid For Scientific Research(B)(1) Project No. 15300026.

8. REFERENCES

[1] E. Voorhees, et al.,"Overview of the Seventh Text REtrieval Conference", *Proc. of TREC-7*, pp.1-24,1998

[2] J.S. Garofolo, et al., "The TREC SDR Track: A Success Storyh In Eighth Text Retrieval Conference", *Proc. of TREC-8*, pp. 107-129, 2000

[3] M. Wechsler, "Spoken Document Retrieval Based on Phoneme Recognition", *Ph.D. thesis*, Swiss Federal Institute of Technology (ETH), Zurich, 1998

[4] K. Ng., "Subword-based approaches for Spoken Document Retrieval", *Ph.D. thesis*, Massachusetts Institute of Technology, Cambridge,MA, 2000

[5] K. Tanaka, et al., "Speech data retrieval system constructed on a universal phonetic code domain", *Proc. of ASRU2001*, pp. 1-4, 2001

[6] K.Itoh et al.,"JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *J. Acoust. soc. Japan (E)*, Vol.20, No.3, pp.199-206, March, 1999.

[7] Y. Itoh, et al., "Automatic Labeling and Digesting for Lecture Speech Utilizing Repeated Speech by Shift CDP", *Proc. of EUROSPEECH2001*, pp. 1805-1808, 2001

[8] T. Kawahara, et al.,"Sharable software repository for Japanese large vocabulary continuous speech recognition",*Proc. of ICSLP1998*, pp.3257-3260, 1998

[9] http://julius.sourceforge.jp/en/julius.html