MULTILAYER SUBWORD UNITS
FOR OPEN-VOCABULARY SPOKEN DOCUMENT RETRIEVAL

Shi-wook Lee*, Kazuyo Tanaka* **, Yoshiaki Itoh***

* National Institute of Advanced Industrial Science and Technology, JAPAN
** Institute of Library and Information Science, University of Tsukuba, JAPAN
*** Faculty of Software and Information Science, Iwate Prefectural University, JAPAN
s.lee@aist.go.jp, ktanaka@slis.tsukuba.ac.jp, y-itoh@iwate-pu.ac.jp

ABSTRACT
This paper describes the application of subword units in an effort of improving open-vocabulary spoken document retrieval performance in the case of highly corrupted recognition output. This paper presents the developed open-vocabulary spoken document retrieval system including the newly proposed subphonetic segment unit and combining multilayer subword units. Our experiments on Japanese spoken documents show that using the proposed subphonetic segment unit can improve retrieval performance, high precision and recall, and a combination of multilayer subword units is also effective.

1. INTRODUCTION
Since an enormous amount of human-generated information is spoken and much of them is stored in the form of audio signal, to find relevant information in spoken documents is a challenging task for modern multimedia information systems. There are two principal approaches to the task of Spoken Document Retrieval (SDR), the subword-based approach and the word-based approach. The word-based approach is to utilize a large vocabulary continuous speech recognition (LVCSR) system to convert the speech into text, to which well-established text retrieval methods can be applied.[1] An inevitable problem of this approach is the fact that the vocabulary size is limited. An alternative approach is to perform retrieval by subword-based transcriptions resulted from a subword recognizer. Subword-based SDR has the advantages that the recognizer is less expensive and open-vocabulary retrieval is possible, because the recognition component is not bound to any vocabulary. Our approach to SDR is based on a subword recognizer which initially transforms the spoken documents into subword sequences, where the retrieval process is conducted by calculating the distance between the parts of subword sequence extracted from subword recognizer. Since the system is based on directly matching subword sequence itself, the system can be free from the size of vocabulary and grammar and robust to recognition error.[2] Although word-based approaches have consistently outperformed subword approaches,[3] but it is helpful when the word-based speech recognition output is errorful and undesirable, for example Out-Of-Vocabulary (OOV) problems and multilingual tasks.[4] However, the main problem in SDR based on subword units is the fact that the recognition result is corrupted by a considerable amount of recognition errors. The contribution of this work is to provide a new subword unit and the application of multiple subword units to improve the performance of subword-based spoken document retrieval, therefore realizing open-vocabulary retrieval.

2. SUBWORD UNITS
The present authors have been developing a scheme for speech processing systems based on the universal phonetic code (UPC)[2]. All of the speech data are once encoded into UPC sequences, and then the speech processing systems, such as recognition, retrieval, and digestion, are constructed in the UPC domain. The international phonetic alphabet (IPA) or extended speech assessment methods phonetic alphabet (XSAMPA) is the candidate set for the UPC set. Here SAMPA is a machine-readable phonetic alphabet. The subphonetic segment (SPS) is derived from XSAMPA and is refined under the consideration of acoustic-articulatory effects. For example, the XSAMPA (i.e., IPA) contains partly extra-detailed categorization to be modeled in an engineering sense. Therefore, only primary IPA symbols are adopted, and minor phonetic variations are represented by statistical distributions in the acoustic domain. A simple example of an SPS converted from XSAMPA is given below. 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 dictionary.

- Speech: She had your dark ⋯
- XSAMPA-Phoneme: # S i h E del dZ @ r del d A kcl k ⋯
where \# denotes a pause or silence interval. A total of 429 SPSs are extracted from the 43 phonemes for Japanese (including 3 silence types).

3. **SUBWORD BASED SPOKEN DOCUMENT RETRIEVAL**

This work focuses on the subword-based approach, where spoken documents are recognized as subword sequences and retrieval process is carried based on matching the Dynamic Programming scores of these subword sequences. Figure 1 shows the overall block diagram of the proposed SDR system. In the proposed system, the query for retrieval and the target database can be in either text or speech form, because the subword units have linguistic information.

![Block diagram of proposed SDR system based on subphonetic segments](image)

When subword sequences are recognized directly, with higher error rates than for words, selection of a good matching approach becomes much more important. The precisely proposed Shift-Continuous Dynamic Programming (Shift-CDP) is an algorithm that identifies similar parts between a reference pattern \( R_N \) and the input pattern sequence \( I_T \) synchronously. 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.

\[
R_N = \{R_0, \ldots, R_{\tau}, \ldots, R_{N-1}\} \quad (1)
\]

\[
I_T = \{I_0, \ldots, I_{\tau}, \ldots, I_{T-1}\} \quad (2)
\]

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 \([\cdot]\) 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.[5] The retrieved spoken documents are presented to the user in decreasing order of their DP score, given by follows:

\[
G(i, j) = \arg\min \left\{ \begin{array}{ll}
G(i-1, j-1) + D(s_i, s_j) \\
G(i-2, j-1) + D(s_i, s_j) \\
G(i-1, j-2) + 2 \cdot D(s_i, s_j)
\end{array} \right. \quad (3)
\]

\(G(i, j)\) denotes the cumulative distance up to reference subword \( s_j \) and input subword \( s_i \). \( D(\cdot) \) is local distance, which uses previously calculated distance matrix. The lower value of \( G(i, j) \) is ranked higher in retrieval result. Different subword unit can provide different types of information. Longer subword units can provide word or phrase information while shorter units can only represent word fragments. The trade-off is that the shorter units are more robust to errors and word variants than the longer units.[6] Here we assume that a highly reliable result is ranked highly both individual systems using different subword unit. In other words, the irrelevant document with lower score that is ranked high in retrieval outputs might be ranked low with high score in the other outputs. With this assumption, we newly propose the application of combination of DP score obtained from individual subword layer, phoneme\( (G_P) \) and SPS\( (G_S) \).

\[
G_{MULTI}(I, J) = (1 - \alpha) \cdot G_P(I, J) + \alpha \cdot G_S(I, J) \quad (4)
\]

4. **DISTANCE MEASURE**

In the Shift-CDP algorithm, the DP matching score is calculated using a pre-measured SPS distance matrix. Therefore, the system is directly influenced by the distance measure, and selecting a proper measure is important for the performance. Distance measures have been widely applied in a number of speech technologies. For speech coding, distance measures are used in the design scheme for vector quantization algorithms and as objective measures of speech quality. In speech and speaker recognition, the spectral difference between two speech patterns is measured to compare patterns and make similarity decisions. Motivated by these speech recognition techniques, some unit-selection algorithms for speech
synthesis and optimal-joining algorithms now use the distance measure between feature vectors. Here, the distance measures $D_{AB}$ between two multivariate Gaussian distributions, $N(\mu_A, \Sigma_A)$ and $N(\mu_B, \Sigma_B)$, are considered. The Bhattacharyya distance $D_{B_H A T}$, which is covered in many texts on statistical pattern recognition, is a separability measure between two Gaussian distributions:

$$D_{B_H A T} = \frac{1}{N} \sum_{n=1}^{N} \left( \frac{(\mu_{An} - \mu_{Bn})^2}{8} \right) \left( \frac{\Sigma_{An} + \Sigma_{Bn}}{2} \right)^{-1}$$

(5)

where $N$ is the number of HMM states and $N = 3$ states is used throughout this work. The first term of Eq. (5) provides the class separability from the difference between class means, while the second term gives the class separability from the difference between class covariance matrices. Here, considering the insufficiency of training data, the distance measure derived directly from the difference between class mean is adopted, that is, the first term of Eq. (5), as formulated below. This distance is very close to the weighted Mahalanobis distance.

$$D_{AB} = \frac{1}{N} \sum_{n=1}^{N} \left( \frac{(\mu_{An} - \mu_{Bn})^2}{8} \right) \left( \frac{\Sigma_{An} + \Sigma_{Bn}}{2} \right)^{-1}$$

(6)

5. EXPERIMENTAL EVALUATION

Japanese Newspaper Articles Sentences[7] are used for training 43 phonemes and 429 SPSs acoustic models. Also, in order to increase baseline subword recognition accuracy, phoneme and SPS bigram language models were estimated from the same corpus used in training acoustic models. A set of 10 keyphrase queries uttered 2 times by 5 male speakers('100 input queries') are prepared to perform SDR evaluation experiments. Each query has 9 relevant documents in 2000 target database('3.29 hours'). The underlying recognition system for decoding subword is a single pass beam search decoder, which is based on the Jidix system[8] system. Table 1 summaries the recognition performance in each subword units.

<table>
<thead>
<tr>
<th>Type</th>
<th>Correct(%)</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phoneme</td>
<td>71.01</td>
<td>53.26</td>
</tr>
<tr>
<td>SPS</td>
<td>60.69</td>
<td>51.52</td>
</tr>
</tbody>
</table>

As an evaluation measure, recall and precision rate are used, that is commonly used in information retrieval.

Also the F-measure that takes into account both recall and precision is adopted.

$$Recall = \frac{Num. of relevant doc. retrieved}{Total num. of relevant doc.}$$

(7)

$$Precision = \frac{Num. of relevant doc. retrieved}{Total num. of doc. retrieved}$$

(8)

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

(9)

5.1. Subword-based retrieval

Figure 2 shows the baseline SDR performance, recall-precision curves according to subword units. In figure 2, the curve closer to the upper right-hand corner of the graph indicates the better SDR performance. The SPS-based SDR is remarkably outperformed phoneme-based results. These results are referable to the number of subwords, regarded as the quantity of information.

Figure 2: Baseline subword-based SDR performance, Recall-Precision rate curves according to the mixture number and subword type, phoneme(PHO) and SPS; the value within parentheses indicates maximum F-measure

In order to confirm the effectiveness of the proposed SPS unit, the SDR for error-free text queries is also conducted and compared. From figure 3, the performance for speech queries with SPS unit is slightly worse than that for error-free text queries.

5.2. Multilayer subword-based retrieval

As shown in figure 4, the retrieval performance using multilayer subword units is better than the baseline performance using phoneme and SPS individually. The recall rate, precision rate and F-measure using multilayer subword units($G_{MULTI}$) are graphed in figure 5 by changing the weight $\alpha$ in Eq.(4). Compared to the performance of SPS-based(SPS) and phoneme-based(PHO)
systems, it is confirmed that the proposed method combining multilayer subword units improved retrieval performance.

![Figure 3](image3.png)

**Figure 3:** Comparison of SDR performance to different input queries, speech and error-free text queries; Mixture=16

![Figure 4](image4.png)

**Figure 4:** Comparison of Recall-Precision rate curves between the proposed method using multilayer subword unit and the baseline using phoneme(PHO) and SPS

6. CONCLUSIONS

In this paper we have presented new methods for open-vocabulary spoken document retrieval. Experiments on the Japanese retrieval show that the retrieval performance can be improved remarkably using newly proposed SPS unit. Combining multilayer subword units, the performance is slightly better than just using individual subword unit. From the experimental evaluation we confirmed that using the proposed SPS unit and the combination of multilayer subword units is effective for open-vocabulary spoken document retrieval.

![Figure 5](image5.png)

**Figure 5:** Comparison of maximum F-measure of multilayer subword based SDR(proposed) by changing the weight value $\alpha$; Also the Recall and Precision rate of proposed method are plotted.

7. REFERENCES