Open-Vocabulary Spoken Document Retrieval based on new subword models and subword phonetic similarity

Kohei Iwata[†], *Yoshiaki Itoh*[†], *Kazunori Kojima*[†], *Masaaki Ishigame*[†]
Kazuyo Tanaka[‡], *and Shi-wook Lee*[*]

[†] Graduate School of Software and Information Sciences Studies, Iwate Prefectural University, Japan
[‡] Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan
[*] Information Technology Research Institute, AIST, Japan
g231d002@edu.soft.iwate-pu.ac.jp

Abstract

A new type of video retrieval system is proposed that identifies a target video section by searching for a word passage submitted as a quoted speech or text query. The proposed system has two unique characteristics. The first characteristic is that it is based on subword models such as phonemes, syllables, and morphemes so the system is able to deal with any type of query, including new words and personal names. The second characteristic is that the system relies on acoustic similarity between subword models. Furthermore, new subword models were constructed for the retrieval system to improve performance. The new models were based on two concepts: context-dependent models and more sophisticated in the time axis than phone models. Through experimentation, the effectiveness and scope of the proposed spoken document retrieval system were confirmed, and suitable subword models for the proposed method discussed.

Index Terms: Spoken Document Retrieval, open-vocabulary, phonetic similarity, subword model.

1. Introduction

Personal computers are equipped with multimedia applications – HDD and DVD recorders have become popular household appliances. Because a vast amount of data can be stored using these media, demand for retrieval and summarization of the data has been growing. Although program titles can be obtained through Internet, it is impossible to obtain what and where keywords are spoken in the program. In spoken document retrieval (SDR) systems, a target video segment is identified, based on a spoken or text query from a user. In order to retrieve a section of interest, query keywords submitted by numerous users are eventually recognized as proper nouns. However, special terms are not recognized by general speech recognizers because of their dictionary size.

In order to respond every query, we propose an open-vocabulary SDR system. The proposed system has two unique characteristics. First, the system uses subword models such as monophones and triphones. This approach to SDR is advantageous compared with that using speech recognition because it does not impose a vocabulary and any word – even if otherwise out-of-vocabulary (OOV) for a speech recognizer – could be a query word. In this paper, suitable subword models for the SDR system are discussed. The second characteristic of the proposed system is that it uses statistical phonetic similarities between subword models composed of Hidden Markov Models (HMMs). Most approaches using subword models use 'edit distance' [3], which determines the extent to which two subword models are the same. Retrieval performance in the edit distance approach depends on subword recognition errors. References [1][2] introduced comfusion matrix between subwords to improve the performance. The approach discussed here is an attempt to recover these errors more robustly by using the phonetic similarity between subword symbols that have been transformed, based on subword recognition. Phonetic similarity is obtained by determining the statistical distance between any two subword models. subword models are composed of particular HMM states. The statistical distance between each state of two subword HMMs is computed. All these distances are extracted and summed as the distance between two subword models. A distance matrix is composed of all the possible distances between any two subwords.

We propose two new subword models for the SDR system. In speech recognition, a context-dependent triphone model performs better than a monophone model. Thus, the proposed subword models are also context-dependent models. In a previous study [4], the sub-phonetic segment (SPS) model was better at retrieval than the triphone model. The SPS model is more sophisticated in the time-axis than the triphone model. Thus, two new models were devised that are more sophisticated than phone models in the time axis. One is created by dividing a triphone into two subword models and another by dividing a triphone into three subword models. They are called demi-phone and one-third (1/3) phone models.

First, this paper outlines the proposed approach to video data retrieval using speech or text query. Next, the proposed subword models and the method of their construction are described. Then, the method of computing subword similarities is explained. After that, the proposed method for retrieving the target word in the speech corpus is evaluated, and the results discussed. Finally, conclusions are presented.

2. A proposed spoken document retrieval system based on subword models

2.1. Outline of the proposed SDR system

In the proposed system, subword acoustic models, their language models, a subword distance matrix, and subword recognition results for spoken documents must be compiled. subword acoustic models composed of HMMs and language models are based on the speech corpus. All statistical phonetic distances between any

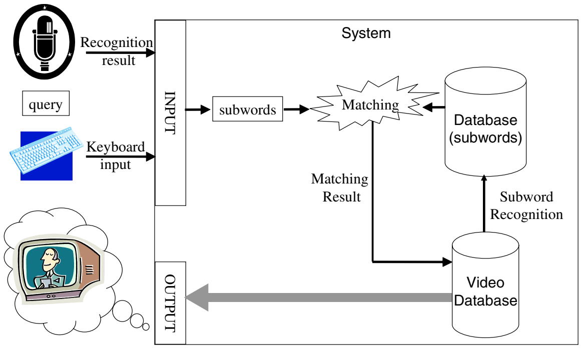

Figure 1: Outline of the proposed SDR system

two subword models are computed and stored in a distance matrix, representing subword similarity. All of the audio data in video data sets are transcribed beforehand into subword sequences by means of subword recognition.

Figure 1 outlines the proposed SDR system based on subword models. To retrieve a target section, the system allows both text and speech queries. When a text query is made, the text is automatically converted to a subword sequence according to conversion rules. For speech queries, the system performs subword recognition and transcribes the speech into a subword sequence in the same manner as it would audio data in a video data set. Here, subword language models are based on subword bigrams and trigrams from JNAS corpus training data [5]. The system then retrieves the target video segment using Continuous Dynamic Programming (CDP) algorithms to compare a queried subword sequence with all the subword sequences in the video data sets. The local distance in the CDP algorithm is checked against the distance matrix. This system identifies those multiple video segments with a high degree of correlation to the query word.

2.2. The proposed new subword models

2.2.1. New subword models

Recently, speech recognition technology has been developed, which relies on acoustic triphone models that consider both the preceding and following phones. Two new subword models are proposed here, based on context-dependent models such as triphone models. In a previous study [4], the XSAMPA-based SPS model [6] is better at retrieval than the triphone model. Each SPS model expresses phonetic feature segments that are more sophisticated in the time axis than the triphone model. Thus, the two proposed subword models are more sophisticated in the time axis than phone models. The first model is labeled 'demi-phone'. Each triphone model is divided into two demi-phone models – a model of the initial and final parts. For example, the triphone model 'a-k+i' is divided into 'a1k' and 'k2i', which indicate the initial and final parts respectively, of the triphone 'a-k+i'. The second is called the 1/3 phone model. A triphone model is divided into three 1/3 phone models. Figure 2 represents each subword expression and concept of the subword boundary of the phone sequence 'aki'.

2.2.2. Preliminary evaluation of subword models

We constructed each subword acoustic model and subword language model by using the JNAS training database and the Hidden Markov Model Toolkit (HTK) as a training tool. The JNAS contains approximately 150 sentences spoken by each of 306 speakers (153 males, 153 females). The conditions for feature extraction are listed in Table 1. The frame interval is usually 10 ms in most speech recognition systems. Because the proposed subword models are sophisticated in the time axis, a more detailed feature sequence was essential. In the preliminary experiment, performance at 5 ms frame intervals was better than at 10 ms frame intervals when using those subword models, and performance at 10 ms frame intervals was better when using the monophone and triphone models.

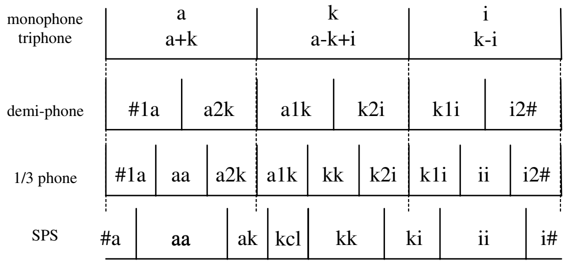

Figure 2: Each subword expression of the word "aki"

Table 1: The conditions for feature extraction

sampling	16 KHz, 16 bits
feature vector	12-dimentional MFCC_E_D_N_Z
window	hamming window
window length	25 ms
frame interval	10 / 5 ms

Table 2 compares each subword model. In Japanese, there are 43 monophone models. Logically, there are 43^3 (79,507) triphones in Japanese. Approximately 8,000 triphones appear in the JNAS training data. Likewise, there are approximately 1,300, 1,400, and 400 demi-phone, 1/3 phone, and SPS models, respectively. In the table, the sophistication level represents the fineness of each time-axis subword model when the phone model is set at 1.0. As a triphone model is divided into two demi-phones, the sophistication level of the demi-phone model becomes 2.0. Perplexity corresponds to the subword perplexity of each subword language model. In the table, context-based models show low perplexity. Furthermore, perplexity is related to sophistication level. The recognition rates were obtained from preliminarily evaluation experiments for each subword model, using the ETL-DB [7]. These results indicate subword recognition rates have deep relationship with the number of subword and perplexity.

2.3. Phonetic similarity between subword models

Phonetic similarity between subword models was introduced into the matching process. The introduction of subword similarity enables flexible and robust spoken document retrieval, compared with approaches that rely on subword frequency or edit distance (whether or not two subwords are the same). The method for determining phonetic similarity is described in this section.

Each subword is composed of HMMs with N states. The phonetic similarity is found by defining the statistical distance between any two subwords models. First, the statistical distance between the Gaussian distributions composing each state is computed. We use the Bhattacharya distance Eq.(1), one of the most representa-

Table 2: Comparison of each subword model

	num. of subword	sophistication level	perplexity	recognition rate
monophone	43	1.0	7.91	73.52
triphone	7,956	1.0	4.73	55.89
demi-phone	1,333	2.0	2.97	65.08
1/3 phone	1,374	3.0	2.02	70.30
SPS	423	2.2	2.65	77.84

tive separability measures expressing distance between two Gaussian distributions. In the equation, $c_k(x)$ and $g_k(x)$ represent the k-th mixture in a state of HMM and the distributions function, respectively. Each distribution is represented by an L-dimensional uncorrelated feature vector, and $g_k(x)$ is represented by the l-th mean vector $\mu_{k\ell}$ and the l-th variance $\sigma^2_{k\ell}$. The first term of Eq.(1) is class separability due to the difference between class means, while the second term is class separability due to the difference between class covariance matrices.

$$\begin{aligned} d_B(c_1(x), c_2(x)) &= -\log \int \sqrt{g_1(x) g_2(x)} dx \\ &= \tfrac{1}{4} \textstyle\sum_{\ell=1}^{L} \left\{ \frac{(\mu_{1\ell}-\mu_{2\ell})^2}{\sigma^2_{1\ell}+\sigma^2_{2\ell}} + \log \frac{(\sigma^2_{1\ell}+\sigma^2_{2\ell})^2}{4\sigma^2_{1\ell}\sigma^2_{2\ell}} \right\} \end{aligned} \quad (1)$$

Second, distances between two states are obtained, as seen in Eq.(2). We define the distance between two states as the nearest distance between any two Gaussian distributions arbitrarily extracted from each state. Therefore, all the distances of $M \times M$ combined mixtures are computed for M mixtures in a state, and the minimum distance is regarded as the distance between the states s_{pj} and s_{qj}. In this paper, each subword model is composed of 16 mixtures: hence, the minimum distance between two states is selected from among 256 distances. In the equation, $c_{pjm}(x)$ corresponds to the m-th mixture in the j-th state s_{pj} of the subword model p.

$$d_s(s_{pj}, s_{qj}) = \min_{1 \le m,n \le M} d_B(c_{pjm}(x), c_{qjn}(x)) \quad (2)$$

The distance between subword models p and q is obtained according to Eq.(3). The distances of all states are then extracted and summed for the distance between two subword models. subword distances are defined as an average value of the distances of N states, which are the components of subwords p and q. These distances $d(p,q)$ are computed and stored in a distance matrix, to which the system refers in the subsequent retrieval process.

$$d(p,q) = \frac{1}{N} \sum_{j=1}^{N} d_s(s_{pj}, s_{qj}) \quad (3)$$

2.4. subword recognition

subword recognition is performed to compile a subword database and convert speech queries to subword sequences. For subword recognition, we use Julius 3.4.2, an open-speech recognizer [8].

2.5. Retrieving the target sections

The system searches a subword sequence of a spoken document query using a CDP algorithm [9]. The CDP algorithm frame synchronously detects segments with a high degree of correlation to the query. The system first identifies the most similar sections in spoken documents, then the next closest candidates, and so on. subword similarity is used as the local distance of CDP.

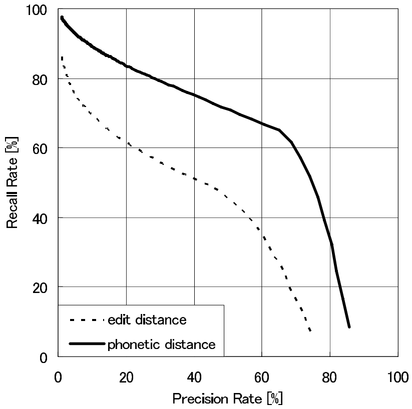

Figure 3: Performance comparison between subword distances

3. Evaluation Experiments

3.1. Evaluation data and conditions

A number of experiments were performed in order to evaluate the effectiveness and flexibility of the proposed SDR system. The object data in the experiments is a word database derived from ETL-DB [7]. The database includes 1,542 words, each spoken by 10 speakers; it therefore contains 10 instances of each word. For any given text query, the target database contains 15,420 (1,542 × 10) words, and 10 correct elements. For any speech query, the target database contains 13,878 (1,542 × 9) words, and only 9 correct elements, as the elements spoken by the subject are excluded. We used Precision-Recall rates as an evaluation measurement, as in Eq.(4). In the equation, 'Num. of relevant words retrieved' corresponds to the number of correct words retrieved within a certain rank, R, and 'Total num. of relevant words' is the number of correct words in the database, or 15,420 (1,542 × 10) for experimental text queries. The 'Total num. of words retrieved' is proportional to rank, that is 1,542 × R. The DP algorithm is used as the experimental matching algorithm. Throughout experimentation, the suitability of each subword model for the proposed SDR system was discussed.

$$\begin{aligned} \text{Precision} &= \frac{\text{Num. of relevant words retrieved}}{\text{Total num.of words retrieved}} \\ \text{Recall} &= \frac{\text{Num. of relevant words retrieved}}{\text{Total num.of relevant words}} \end{aligned} \quad (4)$$

3.2. Results and discussions

Figure 3 compares performance between subword distances in the triphone model. In the figure, the dotted line represents performance when using an edit distance for the local distance, and the solid line represents performance when using phonetic similarity, as described in section 2.3. It indicates that the introduction of subword similarity performs better, obtaining the same results for every subword model.

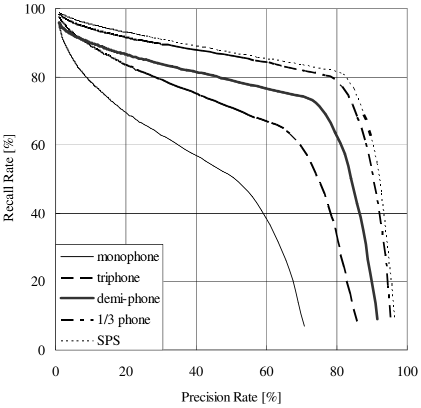

Figure 4: Performance comparison between subword models (text query)

Figure 4 compares performance between each subword model when a text query is submitted. SPS performed best, though comparably to 1/3 phone. Demi-phone, triphone, and monophone performance worsened. 6-state triphone model resembles demi-phone model. However we cannot obtain good result when using the 6-state triphone model. Figure 5 compares performance between each subword model when a speech query is submitted. The figure includes the result of a speaker whose performance is medium among 10 speakers. Performance for all 10 speakers ranked the same as the text queries. In this case, proposed subword models such as demi-phone and 1/3 phone performed better than the triphone or monophone model at both text and speech query, the reason being, proposed model perplexity was lower and sophistication level higher than in traditional subword models such as monophone and triphone. 1/3 phone was lower in perplexity and higher in sophistication than demi-phone, and the performance of 1/3 phone was thought to be higher than demi-phone. On the other hand, 1/3 phone is lower in perplexity and higher in sophistication than SPS. As the number of 1/3 phone subword models is larger than that of SPS, retrieval performance of 1/3 phone was comparable to that of SPS. The results indicate suitable subword models should adhere to three features in order to ensure better retrieval performance with an open-vocabulary SDR system:

- High sophistication level in the time axis
- Small number of subword models
- Low subword perplexity

4. Conclusion

This paper proposed a new type of open-vocabulary system for spoken document retrieval, characterized by the introduction of new subword models and phonetic similarity among subword models. These models are more sophisticated than phone models in the time-axis. Phonetic similarity is obtained from subword HMM statistics. The experiments confirmed the proposed system's effectiveness and flexibility, as well as certain suitable characteristics, upon which future subword models for SDR systems will be constructed. Moreover, the system will be employed with real TV programs to confirm the feasibility of this approach to modeling.

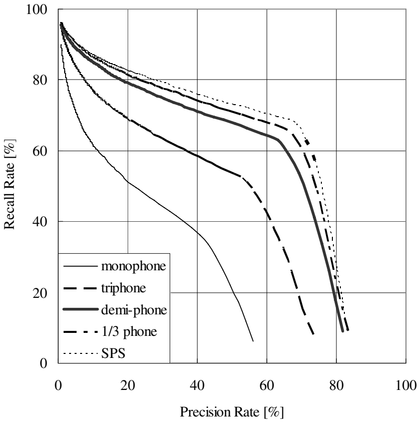

Figure 5: Performance comparison between subword models (speech query)

5. Acknowledgements

This research is based on the work supported in part by Grant-in-Aid for Scientific Research (B) (1) Project No.15300026 and (C) Project No. 17500073.

6. References

[1] F. Crestani, *Combination of similarity measures for effective spoken document retrieval*, Journal of Information Science, 29 (2), pp. 87-96, 2003.

[2] N. Moreau et al, and T.Sikora, *Phonetic confusion matrix based spoken document retrieval*, INTERSPEECH, Vol. 2, pp.1593-1596, 2004

[3] H. Raghavan, et al., *Matching Inconsistently Spelled Names in Automatic Speech Recognizer Output for Information Retrieval*, Proceedings of HLT/EMNLP, pp.451-458, 2005

[4] S. Lee, et al., *Multilayer subword units for open-vocabulary spoken document retrieval*, In INTERSPEECH-2004, 1553-1556, 2004.

[5] K. Itou, *JNAS : Japanese speech corpus for large vocabulary continuous speech recognition research*, J. Acoust. Soc. Jpn. (E), Vol. 20-3, pp.199-2006, 1999.

[6] K. Tanaka, et al., *Speech Date Retrieval System Constructed on a Universal Phonetic Code Domain*, Proceedinds of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU2001), paper a01kt080, pp.1-4, 2001.

[7] S. Hayamizu, et al., *Generation of VCV/CVC Balanced Word Sets for Speech Data Base*, Reprinted from Bul. Electrotech. Lab., Vol. 49 No. 10, 1985.

[8] A. Lee, et al., *Julius - an open source real-time large vocabulary recognition engine-*, European Conference on Speech Communication and Technology, pp.1691-1694, 2001.

[9] Satoru HAYAMIZU, Ryuichi OKA, *Experimental Studies on the Connected Words Recognition Using Continuous Dyanamic Programming*, Transactions of IECE, Vol.67-D, No.6,pp.677-684, 1984.