Two-stage Vocabulary-free Spoken Document Retrieval
- Subword Identification and Re-recognition of the Identified Sections -

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

[†] Faculty of Software and Information Science, Iwate Prefectural University, Iwate
[‡]University of Tsukuba, [*] National Institute of Advanced Industrial Science and Technology (AIST)
y-itoh@iwate-pu.ac.jp

Abstract

A query word for retrieval systems is liable to be a special term not included in a speech recognizer dictionary. Spoken document retrieval (SDR) systems must therefore be vocabulary-free to deal with arbitrary query words. This paper proposes a new method for vocabulary-free spoken document retrieval. The method exploits two-stage tactics. First, when a query word is submitted, the query word is transformed to a subword sequence according to conversion rules. The subword sequence is searched for spoken documents previously transcribed to a subword sequence by subword recognition. The identified sections are extracted according to the distance between the subword sequences of the query and the identified sections. Second, each identified section is re-recognized using a grammar that includes the query subword sequence. Retrieval experiments were conducted with an actual TV program and the results demonstrated that the proposed method improved SDR performance without long delays in retrieval.

1. Introduction

Information retrieval for video data is essential in today's multi-media environment and will become more crucial with the increased use of DVD and HDD video recorders. For information retrieval of spoken documents, approaches based on speech recognition results are representative of spoken document retrieval (SDR) [1-4]. The recent progress of speech recognition technology improves SDR's desirability. If a query word is found in a speech recognizer dictionary, its recognition results can be utilized for SDR. On the other hand, a query word is liable to be a special term not included in speech recognizer dictionaries. Persons' names, place names, and special terms often characterize spoken documents and are thus suitable as query words. Therefore, spoken document retrieval systems must be vocabulary-free if they are to deal with arbitrary query words. This paper proposes a new method for vocabulary-free SDR. The method exploits two-stage tactics. First, when a query word is submitted, it is transformed to a subword sequence according to conversion rules. The subword sequence is searched for spoken documents previously transcribed to a subword sequence by subword recognition. The identified sections are extracted according to the distance between the subword sequences of the query and the identified sections. The system is characterized by the introduction of phonetic similarity between subword models, and new subword models. Phonetic similarity is derived from HMM statistics. References [5][6] have used confusion matrix for subword similarity. Sub-phonetic Segment (SPS) [7] was introduced as a new subword model considered to be more precise than triphone models. Second, each identified section is re-recognized using a grammar that includes the query subword sequence. Retrieval experiments were conducted, which applied the proposed method to an actual TV program and results indicated better SDR performance without long retrieval delays.

In the paper, the outline of the proposed retrieval system and re-recognizing methods in detail are explained first. Then the performance of the proposed method is evaluated for SDR with an actual TV program.

2. Proposed SDR system outline

Figure 1 outlines the proposed system. Speech data sets from programs such as TV news are transcribed to word sequences and subword sequences based on general speech recognition and subword recognition respectively. When more than one text query words are submitted, the system initially searches them in a speech recognizer dictionary. If the query words are known and included in the dictionary, query words sections can be obtained from recognized word sequences. If they are not included in the dictionary, that is are out-of-vocabulary (OOV), an approach using two-stage vocabulary-free SDR is proposed. Of course, this approach could also be used with known words. This paper mainly examines the gray rectangle in Figure 1 and focuses on the re-recognition of the identified sections.

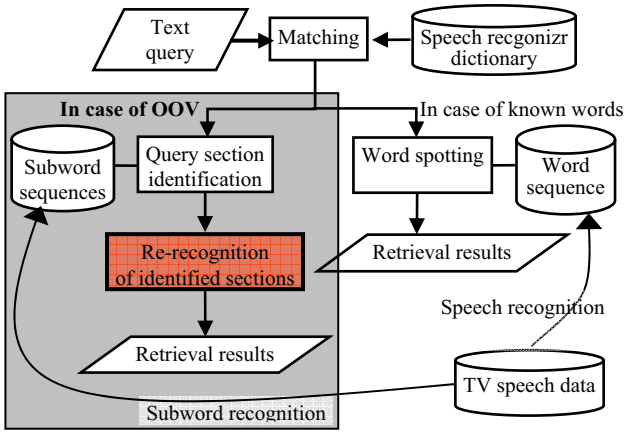

Figure 1 Proposed system outline.

In the first stage, sections of spoken query words are identified (spotted) and ranked according to a similarity measure. In the second stage, the identified sections are re-recognized with a grammar that includes the query words in order of similarity, and determines whether the section is acceptable or not. The two-stages are described in detail in the following section.

2.1. Stage 1: Subword based query identification

Figure 2 shows the outline of the proposed two-stage approach. First, query words are transformed to a subword sequence according to the predefined conversion rules. Then the subword sequence is identified (spotted) in the database (DB) of subword sequences by continuous Dynamic Programming. Phonetic distance between any two subwords, and new SPS subword models are introduced [8]. The phonetic distance is defined between the i-th and j-th subword models as follows, when each model is composed of N states, M mixtures Hidden Markov Models. All the distances are computed between any two distributions in the same k-th states of the i-th and j-th models, and the distance between the same k-th states of the i-th and j-th models is determined with Eq. (1). A Bhattacharya distance is used for the distance between two distributions, as in Eq. (2), the minimum distance being regarded as the distance between the two states. In the equations, m_p denotes the p-th distribution in the k-th state, and each distribution is represented by an L-dimensional uncorrelated feature vector, as in Eq. (2). The distance between subwords model 'i' and 'j' is obtained by averaging all distances between states from Eq. (3).

This first stage produces candidate sections that are ranked according to the cumulative DP distance shown in Figure 2.

$$d_s(i,j,k) = \min_{1 \le m_p, m_q \le M} d\left(g_{i,k}^{m_p}, g_{j,k}^{m_q}\right) \tag{1}$$

$$\begin{aligned} d\left(g_{i,k}^{m_p}, g_{j,k}^{m_q}\right) &= -\log \int \sqrt{g_1(x) g_2(x)}\, dx \\ &= \frac{1}{4} \sum_{\ell=1}^{I} \left\{ \frac{(\mu_{1\ell} - \mu_{2\ell})^2}{\sigma_{1\ell}^2 + \sigma_{2\ell}^2} + \log \frac{\left(\sigma_{1\ell}^2 + \sigma_{2\ell}^2\right)^2}{4\sigma_{1\ell}^2 \sigma_{2\ell}^2} \right\} \end{aligned} \tag{2}$$

$$d(i,j) = \frac{1}{N} \sum_{k=1}^{N} d_s(i,j,k) \tag{3}$$

The effectiveness of introducing phonetic subword distances and SPS models has already been confirmed [8].

2.2. Stage 2: Re-recognition of identified sections

In the first stage, phonetic distance is introduced during the matching process of two subword sequences to improve performance. As all HMM statistics are not utilized then, better performance in identified section re-recognition is expected when all the HMM parameters are introduced in the re-recognition process. Second stage processes are shown in the grey rectangle of Figure 2.

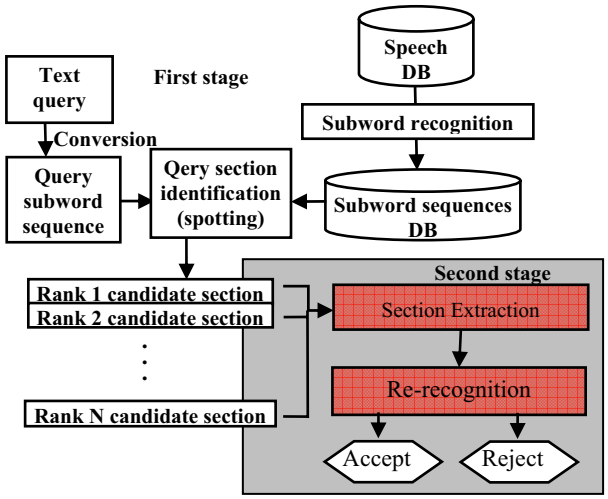

Figure 2 The two-stages SDR approach.

During the first stage, sections are identified and ranked. In the second stage, each candidate section is processed in order of rank. In the first stage, each candidate corresponds to an identified section that is obtained from start-time and end-time information during subword recognition. The query words (their subword sequence) are not always found within an identified section because the correct start or end section of the query words might be absent. An equal length of redundant section is therefore added before and after the identified section, extracted as a candidate section for re-recognition.

A candidate section is assumed to include redundant sections before and after the query words, as described above. Therefore, the grammar needs to contain the extracted candidate section (Figure 3). If the candidate section includes the query words, re-recognition takes the lower pass that includes the query subword sequence. The section is accepted and provided to a user. On the other hand, if the candidate section does not include the query words, re-recognition takes the upper pass using an arbitrary subword sequence, and the section is rejected. An insertion penalty precludes inclusion of the query words during the upper pass.

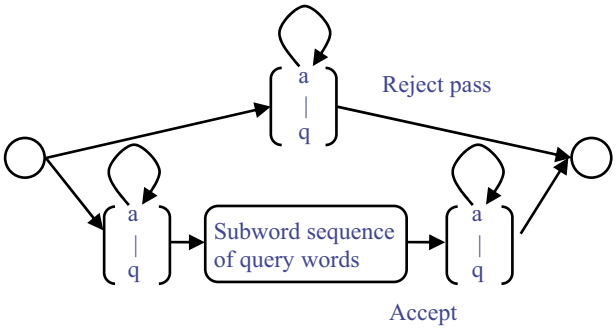

Figure 3 Grammar for candidate section re-recognition.

3. Evaluation experiments

3.1. Evaluation data and conditions

Experiments were performed to evaluate the performance in SDR of the proposed method, described in 2.2. The object data in the experiments were spoken data taken from actual broadcast TV news. The news was broadcast during the major Japanese earthquake in October 2004 - the Niigata Chuetsu Earthquake. Family and friends were anxious about the disaster victims' safety. When relatives tried to confirm their safety, it was very hard to connect to permanent and cellar phones from outside. The Japanese broadcasting company continuously broadcast messages such as "Is Mr. Ichiro Yamada of Yamakoshi village safe? Please call Mrs. Yoko Suzuki living in Tokyo." It broadcast continuously, with each message sent only once. Thus, victims had to watch the program continuously to get their relatives' or friends' messages. A retrieval system was needed for this TV program once it had been recorded as video data. The retrieval method proposed here is a valid application in this case because the messages include persons' names and place names, which can be submitted as query words.

For the experiments, 866 messages amounting to 3 hours spoken by a single announcer were extracted manually to simplify evaluation. The data was first recorded on video tape, and then converted to PC data in mpeg-3 format, assuming a real application. The sampling frequency was 16 kHz, and the frame interval was 10 ms. 12-dimensional MFCC, delta MFCC, power, and delta power were selected as feature parameters. Each subword model is trained using a Hidden Markov Model Toolkit (HTK) for HMM training, and the Continuous Speech Corpus Japanese Newspaper Article Sentences (JNAS) [9] is used for training data. The JNAS contains speech data sets for approximately 150 speech sentences spoken by 306 speakers.

One hundred victims' full names were used as query words. Each name occurred once in the 866 messages. The query words were submitted as text and automatically transcribed in subword sequences. SPS was used as subword model in the first stage, and subword recognition was performed for each message beforehand. Re-recognition was accomplished by using triphone models and Julian [10] in the grammar during the second stage (Figure 3).

In the experiments, performance was compared with a system using a general speech recognizer, Julius 3.4.2 [10] with a 50k vocabulary. To deal with unknown words in Julius, such as the victims' names, we add each name to the Julius dictionary in the category of unknown words, after the name is submitted. Julius performed speech recognition for all the 866 messages. Retrieval is successful if the names are detected in the recognition results.

3.2. Results and discussion

First, performance of the proposed initial subword matching method was compared with a system using the Julius speech recognizer (Figure 4). The speech recognizer's result is plotted as one point because its recognition results might or might not include the query word. The result of the subword matching starts (right side) when rank 1 is re-recognized for each query name. This point demonstrates performance for 100 queries as well as the top candidate - 63 were correct and 37 incorrect. The recall rate is improved by including lower rank candidates. The performance using speech was better than that of subword matching during the first stage. The results demonstrate that speech recognition works well when compared with subword matching if the query word is included in its dictionary.

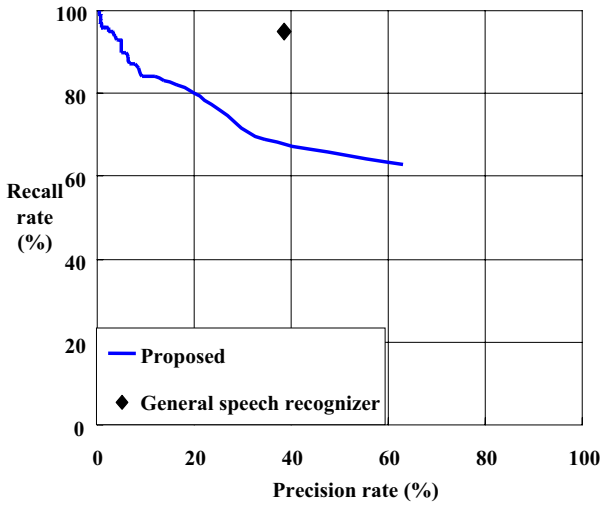

Figure 4 Performance comparison between the proposed system's first stage and a system using a general speech recognizer.

Second, performance was evaluated for re-recognition of a candidate section identified in the first stage (Figure 5). Performance was much improved with re-recognition. Few sections were rejected in spite of including the correct query section. Few messages included query word sections outside the section identified during the first stage. The proposed system's performance is comparable to that using a speech recognizer.

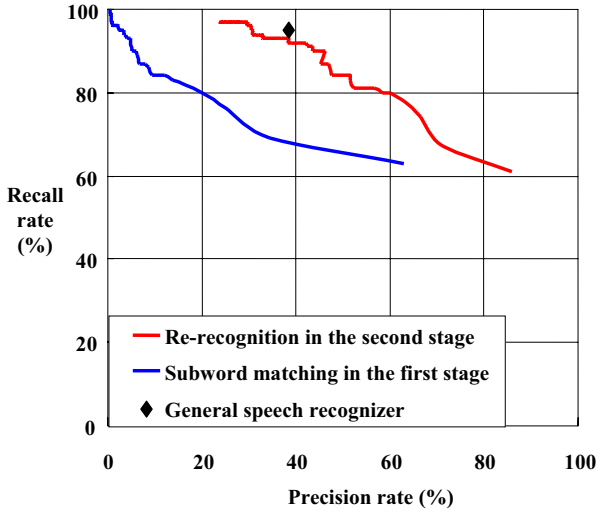

Figure 5 Performance comparison between the first stage and the second stage of the proposed system, and a system using a general speech recognizer.

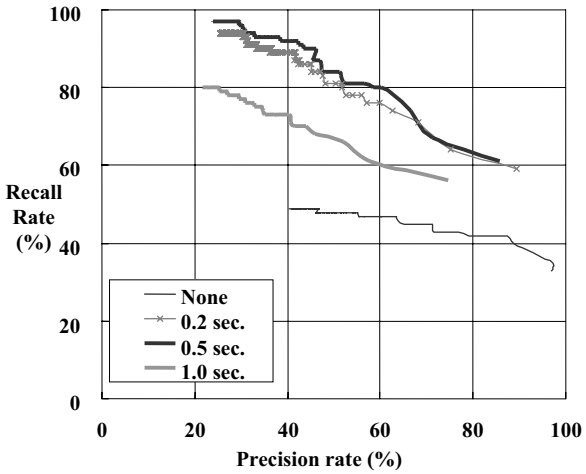

Figure 6 Performance according to the length of redundant section before and after the identified section.

Performance was also evaluated according to the length of a redundant section added before and after the section identified during first stage (Figure 5). Performance without the redundant section deteriorated because reliable segmentation could not be obtained with subword recognition, and many identified subword sections were missing the beginning or end sections of the query words. Performance was best when approximately 0.5 second sections were added before and after the identified section. On the other hand, long redundant sections caused performance deterioration, as shown in the case of 1.0 second sections. These results suggest word spotting does not work well under loose constraints with the grammar described in Figure 3.

3.3. Retrieval time

Delays are a critical problem in retrieval systems, so SDR's retrieval time was evaluated. For the three hours of speech data, it took 120 minutes using the Julius speech recognizer. When using the proposed system, it took approximately 1.5 seconds for subword identification in the first stage, and approximately 1.0 second for re-recognition of a single candidate section. Therefore, it takes about (N+1.5) seconds to finish the process of all candidates within the N-th rank.

Figure 7 illustrates the recall rate according to waiting time. The horizontal line at 94% indicates the recall rate using a speech recognizer, which took 120 minutes, as mentioned above. The proposed system can obtain the same recall rate with only a 150 second-delay.

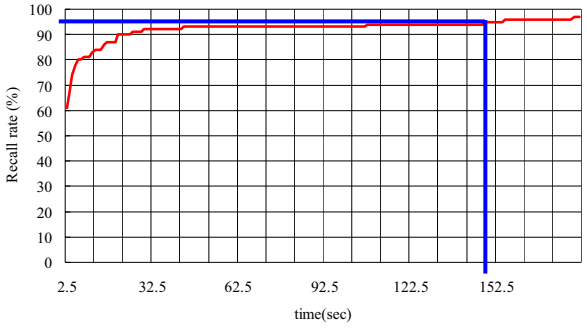

Figure 7 *Recall rates against waiting times.*

The proposed system generates one candidate in 2.5 seconds at the fastest, and generates candidates at a 75 % recall rate within 5 seconds. While the user is checking the candidate message for 12 seconds on an average, re-recognition continues to be processed. After checking the first candidate message, the system can produce candidates at a 90% recall rate and a 45% precision rate. From these points of view, the proposed system's feasibility was confirmed.

4. Conclusions

This paper proposed a new approach for spoken document retrieval. The proposed system exploits two-stage tactics. First, query words are transcribed as subword sequences, and subword identification is performed for spoken documents. Second, each identified section is re-recognized using a grammar that includes the query subword sequence. Retrieval experiments were conducted using the method with an actual TV program. Results indicated improved SDR retrieval without long delays, confirming the feasibility of the proposed system.

5. Acknowledgements

This research is supported in part by Grand-in-Aid for Scientific Research B (1) Project No. 15300026 and (C) Project No. 1750073, Japan Society for Promotion of Science.

6. References

[1] Rose R. C., Chang E. I. and Lippmann R. P., "Techniques for information retrieval from voice messages," ICASSP, Vol. I, pp.317-320, Apr.1991.

[2] Garofolo J. S., Auzanne C., Voorhees E M., "The TREC Spoken Document Retrieval Track: A Success Story," Recherche d'Informations Assiste par Ordinateur, 2000.

[3] Auzanne C., Garofolo J. S., Fiscus J. G., Fisher W. M., "Automatic Language Model Adaptation for Spoken Document Retrieval," B1, 2000 TREC-9 SDR Track, 2000.

[4] Fujii A., Itou K., "Evaluating Speech-Driven IR in the NTCIR-3Web Retrieval Task," Third NTCIR Workshop, 2003.

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

[6] Moreau N., Kim H., and Sikora T., "Phonetic confusion matrix based spoken document retrieval," INTERSPEECH, ICSLP, Vol. 2, pp.1593-1596, 2004

[7] Tanaka, K., Kojima H., "Speech recognition method with a language-independent intermediate phonetic code", ICSLP, Vol. IV, pp.191-194, 2000.

[8] Lee S., Tanaka K., and Itoh Y., "Robust spoken document retrieval based on multilingual subphonetic segment recognition," International Conference on Enterprise Information Systems, Vol. 5, pp. 134-139, 2004.

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

[10] Lee A., Kawahara T. and Shikano K., "Julius - an open source real-time large vocabulary recognition engine," EUROSPEECH, pp. 1691-1694, 2001.