



An Integration Method of Retrieval Results using Plural Subword Models for Vocabulary-free Spoken Document Retrieval

Yoshiaki Itoh¹, Kohei Iwata¹, Kazunori Kojima¹, Masaaki Ishigame¹,
Kazuyo Tanaka², and Shi-wook Lee³

¹ Faculty of Software and Information Science, Iwate Prefectural University, Iwate
² University of Tsukuba ³ National Institute of Advanced Industrial Science and Technology (AIST)
y-itoh@iwate-pu.ac.jp

Abstract

Spoken document retrieval (SDR) systems must be vocabulary-free in order to deal with arbitrary query words because a user often searches the section where a query word is spoken, and query words are liable to be special terms that are not included in a speech recognizer's dictionary. We have previously proposed new subword models, such as the 1/2 phone model, the 1/3 phone model, and the sub-phonetic segment (SPS) model, and have confirmed the effectiveness of these models for SDR [1]. These models are more sophisticated on the time axis than phoneme models such as the triphone model. The present paper proposes an integration method of plural retrieval results that are obtained from each subword model and demonstrates the performance improvement through experiments using an actual presentation speech corpus.

Index Terms: spoken document retrieval, subword model, plural model integration

1. Introduction

The demand for information retrieval of video data will become increasingly necessary with the enormous amount of data in HDD video recorders. For spoken document retrieval (SDR), approaches based on speech recognition results are representative [2-5]. If a query word is included in a speech recognizer's dictionary, its recognition results can be utilized. In an investigation of keyword retrieval of Yahoo Japan the top 100 keywords retrieved in 2004 and 2005, almost half of the keywords are not included in the dictionary of a general speech recognizer because proper nouns and recently generated words were included. Therefore, spoken document retrieval systems must be vocabulary-free if they are to deal with arbitrary query words.

We have therefore proposed a vocabulary-free SDR system that exploits subword models such as the monophone model, the triphone model, and the proposed models. This approach is advantageous because any word can be a query word. In a previous study [1], we proposed two new subword models instead of a triphone model, which is a typical phone model in speech recognition. These new models are called the half-phone model and one-third (1/3) phone model. The half-phone model is generated by dividing a triphone into two subword models, and the 1/3 phone is generated by dividing a triphone into three subword models. Each subword acoustic model is composed by a Hidden Markov Model (HMM) with the same number of states. These subword models are therefore context-dependent models, like the triphone model and are more sophisticated on the time axis than the triphone model. The experimental results revealed that the SDR

performance using the proposed models was better than that using the monophone model or the triphone model.

We confirmed that the average retrieval performances of the 1/2 phone, the 1/3 phone, and the sub-phonetic segment (SPS) [6] models were better than that of the triphone model. The retrieval performance for each query word does not always show the same tendency. Therefore, we propose a method of integrating plural retrieval results of more sophisticated subword models on the time axis in an attempt to improve the retrieval performance. A brief outline of the proposed method is described as follows. First, each subword model m generates the distance $D_m(i, j)$ between a query Q_i and a spoken document or speech section S_j . Second, a modified distance is obtained by combining the distances $D_m(i, j)$ (e.g., $m = 3$). The present paper investigates the performance using plural subword models by combining the distance linearly and, through experiments, demonstrates that the performance can be improved using an actual presentation speech corpus.

In the present paper, an outline of the proposed retrieval system and the integration method of plural retrieval results are explained in detail. The performance of the proposed method is evaluated for SDR with a presentation corpus.

2. The Proposed SDR System and Subword Models

2.1. Outline of the Proposed SDR System

In the proposed system, we have to prepare subword acoustic models, their language models, a subword distance matrix, and subword recognition results of spoken documents. Subword acoustic models composed of HMMs and language models are obtained using the speech corpus. All statistical phonetic distances between any two subword models are computed and stored in the distance matrix. This distance matrix represents the subword similarity. All of the audio data in the video data sets are transformed beforehand into subword sequences by means of subword recognition.

Figure 1 shows an outline of the proposed SDR system based on subword models. (1) The system performs subword recognition for all of the spoken documents and prepares a subword sequence database beforehand. Here, we use subword language models such as subword bigrams and trigrams that are trained from the speech corpus. (2) The system allows both text and speech queries. (3) When a user inputs a text query, the text is automatically converted to a subword sequence according to conversion rules. For speech queries, the system performs subword recognition and transforms the speech into a subword sequence in the same

manner as spoken documents in (1). (4) The system then retrieves the target section using Continuous Dynamic Programming (CDP) algorithms by comparing a query subword sequence to all of the subword sequences in the spoken documents. The local distance in the CDP algorithm refers to the distance matrix that contains the statistical distance between any two subword models. (5) The system outputs plural target sections that show a high degree of similarity to the query word.

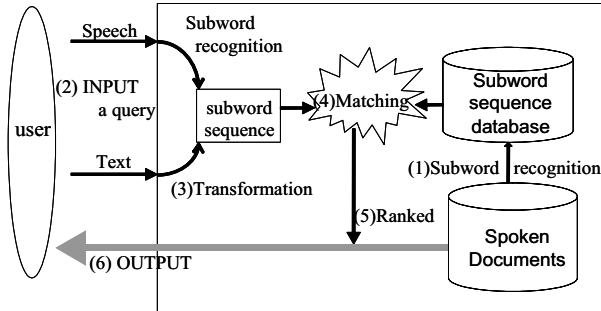


Figure 1: Outline of the proposed SDR system.

2.2. Subword Models

We previously proposed new subword models that are context-dependent models, such as triphone models, and are more sophisticated models on the time axis than triphone models. Two new subword models were proposed, and these models were confirmed to have better SDR performance than triphone models [1]. The first model is the 1/2 (demi) phone model. Each triphone model is divided into two 1/2 phone models: a model of the front part and a model of the rear part. The second model is the 1/3 (one-third) phone model. A triphone model is divided into three 1/3 phone models. In a previous study [6], the sub-phonetic segment (SPS) model based on XSAMPA was shown to have better retrieval performance than the triphone model. The SPS model is also used for plural subword integration.

Figure 2 represents each subword expression and the conceptions of the subword boundary of the three-phone sequence “k a p”. For example a context-dependent triphone “k-a+p” of a phone “a” is divided into two 1/2 phones, as in the figure. The triphone model “k-a+p” is divided into “k-a” and “a+p”. The 1/2 phone “k-a” is the front half phone model of phone “a”, which follows phone “k”, and “a+p” is the rear half phone model of phone “a” that precedes phone “p”. In the same way, the triphone model “k-a+p” is divided into “k-a”, “aa”, and “a+p”, as shown in Figure 2. Each SPS model expresses an acoustic segment according to phonetic features. The “ka” model expresses a transit part from phone “k” to phone “a”, and “kk”, “aa”, and “pp” express the center segment of each phone “k”, “a”, and “p”, respectively. Here, kcl and pcl express the silent part just before an explosive. Each subword is composed of HMMs with N states. As shown in Figure 2, the 1/2 phone, 1/3 phone, and SPS models are more sophisticated on the time axis than the triphone and monophone models. In subword recognition, subword substitutions and subword deletions often occur, and such an error has a bigger influence on a phone model, such as the triphone model than on the proposed models, which have redundantly more subword models on the time axis. Such redundancy of subword models is thought to work well for SDR.

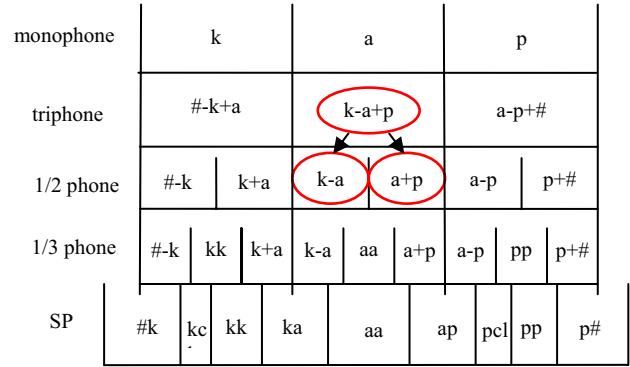


Figure 2: New subword models for the SDR system.

2.3 Integration of Retrieval Results from Plural Subword Models

The retrieval performance can be evaluated with the precision-recall performance or the F-value. The proposed models and the SPS model showed better average performance than the triphone and monophone models. When the performance is evaluated for each query word, the triphone model sometimes had a better performance than the 1/2 model, and the performances of the 1/2 phone, 1/3 phone, and SPS models are not always same. Therefore, we propose a method that integrates plural retrieval results of these more sophisticated subword models on the time axis in an attempt to improve the retrieval performance.

The method integrating plural retrieval results is now described. First, each subword model m ($1 \leq m \leq M$) generates the distance $D_m(i, j)$ between a spoken document or speech section S_j ($1 \leq m \leq I$) and a query Q_i ($1 \leq j \leq J$). Here, M , I , and J denote the number of subword models, the number of documents, and the number of query words, respectively. Second, a modified distance that is a new criteria is obtained by integrating the distances $D_m(i, j)$. If integrating two subword models m and m' , we simply combine two distances linearly, according to the following equation:

$$D_2(i, j) = \alpha \cdot D_m(i, j) + (1 - \alpha_m) \cdot D_{m'}(i, j) \quad (1)$$

As we can utilize five subword models, we generalize the above equation in the following equation. Here, α_m is a weighting factor for the m -th subword model. The present paper investigates the performance using plural subword models, and demonstrates the performance improvement.

$$D_M(i, j) = \sum_{m=1}^M \alpha_m \cdot D_m(i, j) \quad (2)$$

$$\sum_{m=1}^M \alpha_m = 1 \quad (3)$$

3. Experiments

3.1. Experimental Conditions

The conditions for feature extraction are listed in Table 1. In our preliminary experiment, the performance at the 5-ms frame shift was better than that at the 10-ms frame shift when using the three new subword models, and the performance at the 10-ms frame shift was better when using the monophone and triphone models [1].

Table 1. Conditions for feature extraction

Sampling	16 kHz 16bit
Feature parameter	12-dimensional MFCC + 12-dimensional Δ MFCC + Energy
Window length	16 ms.
Frame shift	10 ms. for monophone and triphone 5 ms. for 1/2 phone, 1/3 phone and SPS

We constructed five subword acoustic models and subword language models, which are the monophone, triphone, 1/2 phone, 1/3 phone, and SPS acoustic and language models. All of the acoustic and language models were trained by the JNAS database, and the Hidden Markov Model Toolkit (HTK) was used as a training tool. The JNAS contains sentence speeches by 306 speakers (153 males, 153 females), each of whom spoke approximately 150 sentences. There were 43 monophones, approximately 8,000 triphones, 1,300 1/2 phones, 1,400 1/3 phones, and 400 SPSs, respectively.

We changed the weighting factor α_m at every 0.1 from 0.0 to 1.0 in equations (2) and (3). We evaluated the performance for all combination of the weighting factor α_m . For example, in the case of two subword models, there are eleven combinations, from $\alpha_1 = 0.0$ to $\alpha_1 = 1.0$, and in the case of three subword models, all combinations such as $\alpha_1 \geq 0.1$, $\alpha_2 \geq 0.1$, and $\alpha_3 \geq 0.1$ are added to the combinations for the case of two subword models.

We used the precision-recall rates for evaluation measurement. The continuous DP algorithm was used as the matching algorithm between a query subword sequence and a subword sequence of a spoken document in the experiment.

3.2. Test Data and Evaluation Measurement

The test data in the experiments are an actual presentation corpus of CSJ [7]. The test data include 49 presentation speeches that total approximately twelve hours. Each presentation is spoken by a different speaker. We extracted 50 query words that characterize the speech. Each query has three to 50 corresponding sections in the test data.

We used an average precision rate for the evaluation measurement. For each query q , an average precision is obtained according to the following equation. Therefore, the evaluation measurement becomes the average rate among the average precision rates for all of the queries. Here, k denotes the k -th rank among the extracted sections, and N denotes the last rank when all of the correct sections are extracted. D_q denotes the number of correct sections for a query q . The average precision is obtained by averaging the precision rates when each correct section is extracted.

$$\text{Average Precision} = \frac{1}{|D_q|} \sum_{1 \leq k \leq N} r_k \times \text{precision}(k) \quad (4)$$

$r_k = 1$ if the k -th extracted section is correct

$r_k = 0$ if the k -th extracted section is NOT correct

3.3. Results and Discussion

Table 2 shows the recognition rates for each subword model. The recognition rate depends on the number of

subword models, the perplexity of the subword language models, and the redundancy in the time axis [1].

Table 3 shows the retrieval performance using each subword model. These results indicate that the performance of the triphone model is not better than those of the other models. This is because the triphone has a number of models, and we assume that the triphone language model was not trained sufficiently. The performance can be improved by using a number of training data sets. The performances of the proposed models and the SPS model were better than those of the phone models, as shown in Table 3. The correct rate was high when extracting the top candidate by the 1/2 phone and SPS models.

Table 2. Basic data for each subword model

Subword model	# of models	perplexity	recognition rate
monophone	43	7.91	73.5
triphone	7,956	4.73	55.9
1/2 phone	1,333	2.97	65.1
1/3 phone	1,374	2.02	70.3
SPS	423	2.65	77.8

Table 3. Retrieval performance by each subword model

Subword model	Average precision	Correct rate for the top candidate
monophone	38.6	82
triphone	34.04	72
1/2 phone	67.49	98
1/3 phone	53.49	88
SPS	60.28	96

Figure 3 shows the results for the case of integrating two subword models. In the figure, 1/2, 1/3, monof, and tri denote the 1/2 phone, 1/3 phone, monophone, and triphone models, respectively. The horizontal axis corresponds to the weighting factor, which was varied from 0.0 to 1.0. We confirmed that the performance was improved in all cases of integrating any two subword models. In many cases, the performance was improved when the weighting factor of the subword model that shows the best performance is larger. For example, in the case of integrating the 1/2 phone and SPS models, the performance improved when the weighting factor of the 1/2 phone increased because the performance of the 1/2 phone model was better than that of the SPS model, as shown in Table 3.

Table 4 shows the best retrieval performance using plural subword models. N denotes the number of subword models. In the case of $N = 1$, the best performance was obtained using the 1/2 phone, as shown in Table 3. Figure 3 shows the case of $N = 2$. The best performance was obtained when using the 1/2 phone and the SPS models at weighting factors of 0.6 and 0.4, respectively, and the performance was improved by 3%. In the case of $N = 3$, the best performance was obtained when using the 1/2 phone SPS and the 1/3 phone models at weighting factors of 0.5, 0.3, and 0.2, respectively, and the performance was improved by approximately 1%.

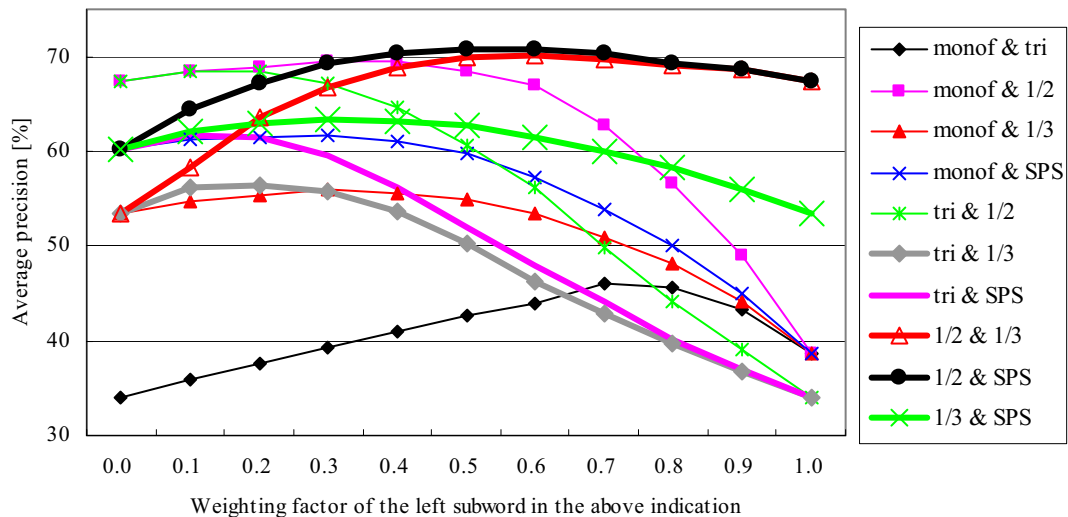


Figure 3. Retrieval performance obtained by integrating two subword models according to the weighting factor.

Table 4. Best performance and the weighting factor when integrating two to five subword models.

N	Weighting factor					Average precision
1	1/2					67.49
	1.0					
2	1/2	SPS				70.78
	0.6	0.4				
3	1/2	SPS	1/3			71.66
	0.5	0.3	0.2			
4	1/2	1/3	SPS	monof		71.76
	0.5	0.2	0.2	0.1		
5	1/2	1/3	SPS	monof	tri	71.76
	0.5	0.2	0.2	0.1	0	

When using four subword models at $N = 4$, the best performance was obtained by adding the monophone model and the previous three subword models at $N = 3$. The weighting factor of the monophone model was 0.1 , and the performance improvement was also small. In the case of $N = 5$, the performance was the same as that in the case of $N = 4$ because the weighting factor of the triphone was 0.0 , and the result of the triphone was not utilized.

We confirmed the improvement of the retrieval performance by integrating the results from plural subword models. The weighting factor shown in Figure 3 illustrates that large values should be given to the subword models that show better subword recognition performance. A method for determining the weighting factors automatically should be developed.

4. Conclusions

We have developed new subword models for SDR systems that can deal with arbitrary query words. The new subword models such as the 1/2 phone, 1/3 phone, and sub-phonetic segment models are more sophisticated on the time

axis than phone models such as the triphone model. The present paper proposed a method of integrating plural retrieval results that are obtained from these plural subword models. We demonstrated experimentally that the performance can be improved by integrating plural subword models. A method for determining the weighting factors automatically will be investigated in the future.

Acknowledgements

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

References

- [1] Iwata, K., Itoh, Y., Kojima, K., Ishigame, M., Tanaka, K. and Lee, S., "Open-Vocabulary Spoken Document Retrieval based on new subword models and subword phonetic similarity," INTERSPEECH, 2006.
- [2] 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.
- [3] Garofolo J. S., Auzanne C., Voorhees E M., "The TREC Spoken Document Retrieval Track: A Success Story," Recherche d'Informations Assiste par Ordinateur, 2000.
- [4] Auzanne C., Garofolo J. S., Fiscus J. G., Fisher W. M., "Automatic Language Model Adaptation for Spoken Document Retrieval," B1, 2000TREC-9 SDR Track, 2000.
- [5] Fujii A., Ito K., "Evaluating Speech-Driven IR in the NTCIR-3Web Retrieval Task," Third NTCIR Workshop, 2003.
- [6] Tanaka, K., Kojima H., "Speech recognition method with a language-independent intermediate phonetic code", ICSLP, Vol. IV, pp.191-194, 2000.
- [7] Ito K., "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," J. Acoust. Soc. Jpn. (E), Vol. 20-3, pp.199-2006, 1999.
- [8] Itoh, Y., Otake, T., Iwata, K., Kojima, K., Ishigame, M., Tanaka, K. and Lee, S., "Two-stage Vocabulary-free Spoken Document Retrieval -Subword Identification and Re-recognition of the Identified Sections-," INTERSPEECH, 2006.