Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Integrating Plural Results of Spoken Term Detection Using Plural Language Models for Subword-based Speech Recognition

Itoh, Yoshiaki (1), Onodera, Yuji (1), Kojima, Kazunori (1), Ishigame, Masaaki (1), Tanaka, Kazuyo (2), Lee, Shi-wook (3)

(1) Iwate Prefectural University, Iwate, Japan
(2) Tsukuba University, Ibaraki, Japan
(3) AIST, Ibaraki, Japan

PACS: 43.55.CS

ABSTRACT

Recently, Spoken Term Detection (STD) that identifies the target section of user's interest in spoken documents has been one of the hottest topics in spoken document processing. We have proposed a subword-based STD method to deal with out of vocabulary query terms, and have demonstrated newly proposed subwords such as 1/2 phone and Sub-phonetic Segment worked well for STD. The paper improves the STD performance by integrating plural STD results obtained by using plural language models. We prepare three different types of language models for each subword such as monophone, triphone, 1/2 phone and SPS using three different speech corpora that consist of a JNAS (Japanese Newspaper Article Sentences) corpus that includes read speeches of newspaper articles with their pronunciations, CSJ (Corpus of Spontaneous Japanese) that includes actual presentation speeches, and our WEB dictionary whose entries were collected by searching keywords with their pronunciation for WWW texts. We used 50 presentation speeches in CSJ for test data and the rest of about 2600 presentation speeches in CSJ for training subword language models. Subword based speech recognition using each subword language model is performed for spoken documents. Three subword recognition results are obtained, and a DP matching process is performed between a sequence of subword models of a query and the three results of subword model sequences of spoken documents. Here we use subword phonetic distances between any two subword models to recover subword recognition errors. Three cumulative distances are computed for each candidate section. These distances are integrated linearly and each candidate section is re-ranked according to the integrated distance. The proposed method mentioned above could improved the STD performance in any two and three STD results, and could confirm the effectiveness integrating plural STD results obtained by different subword language models.

INTRODUCTION

The recent progress of information technologies these days allows us to handle multimedia data easily. Hard discs have recently come into widespread use, and the medium used by a home video recorder has been changing from videotape to hard disc or blue-ray disc. Such media can store recording video data of great length (long-play video data). In association with the increasingly common use of such long-play video data, demand for retrieval of the data has been growing. Meanwhile, detailed descriptions of the content associated with correct time information are not usually attached to these data, although topic titles can be obtained from electronic TV programs and attached to the data. The function for retrieving multimedia data, therefore, is needed because of the increase of stored multimedia data. Many methods have been proposed for the spoken term detection (STD) that identifies the target section of user's interest in spoken documents using a transcription generated from a speech recognizer. Such STD methods generate a word-level transcription of speech data using a large vocabulary continuous speech recognition (LVSCR) system, and find query terms in the transcription. These methods have difficulty in detecting out-of-vocabulary (OOV) terms that are not included in a dictionary of the LVSCR system, because OOV terms are inevitably substituted to other words in the dictionary. STD systems have to detect OOV query terms because query terms are likely to be OOV terms, such as technical terms, geographical names, personal names and neologism and so on. To realize an open-vocabulary STD system, the methods based on subwords that are a smaller unit than a word such as monophone and triphone have been proposed [1-5]. Because the performance of the methods using a subword for in-vocabulary (IV) query terms tends to be lower than that using the LVCSR, some researches combined these two methods [1, 2]. The performance for OOV query terms is regarded as the most critical issue for the STD task. Therefore we have proposed more sophisticated subword units such as 1/2 phone, 1/3 phone and Sub-phonetic Segment (SPS) [9] and have demonstrated the proposed subwords worked well for STD [3]. A method of integrating plural detection results obtained from several subwords was also proposed to improve the STD performance [4]. Integration is conducted by a linear combination of matching distances for each candidate section.

The integration of plural results was shown to be effective in the STD performance.

Based on the above-mentioned idea of integrating plural STD results, the present paper proposes integrating plural results obtained by using plural language models for subword-based speech recognition. One of the ways to integrate plural results is to use N-best results obtained by subword recognition. The way was not much effective in our previous experiments because the diversity of the subword models in the N-best results was not abundant. Therefore, we prepare different subword language models for each subword such as monophone, triphone, 1/2 phone and SPS using different speech corpora, and obtain an abundant diversity in subword recognition results. We use three types of training data sets for subword language models; JNAS (Japanese News Article Sentences) [6], CSJ (Corpus of Spontaneous Japanese) [7], and a WEB dictionary that we constructed for another research purpose and it includes 1.2 million words. Each training data set has different characteristics. The JNAS includes read speeches of news articles with their pronunciations. The CSJ and WEB dictionary include academic presentation speeches and all kinds of words on WWW, respectively. In the present paper, we improve the STD performance by integrating three 1-best results obtained by subword recognition using the three language models. Evaluation experiments using 50 actual presentation speeches in CSJ demonstrate the proposed method improves the STD performance by integrating plural STD results obtained by different subword language models.

The present paper first describes the proposed method in detail in the next chapter. Evaluation experiments using a speech corpus and discussions are described in chapter 3. Finally, our conclusions are presented in chapter 4.

PROPOSED METHOD

Outline of STD system for OOV query term

If query terms are known words in the dictionary of a speech recognizer, the target section can be easily identified by searching the same word sequence as the query terms for a word database. The word database is constructed by performing word speech recognition for spoken documents. The paper focuses on and describes the function for out-of-vocabulary (OOV) query terms.

The outline of our STD system for OOV query terms is shown in Figure 1. The system is composed of two steps. The first step is so called subword recognition, shown in the dotted line of the figure. All utterances in spoken documents are recognized using HMM acoustic models, a subword dictionary and a subword language model that is commonly an N-gram model beforehand. The results of the subword recognition consist of a transcription of subword models (subword model sequences). The information is stored in a subword database in advance. The second step is the detection process of query terms. When query terms are given by a user, they are converted to a subword model sequence according to conversion rules. The optimal path and the cumulative distance are computed for each utterance by comparing the subword model sequence of query terms with all subword model sequences in the subword database, using continuous dynamic programming (CDP). To recover subword recognition errors, an acoustic subword distance between two subword models is used for a local distance in DP matching [4].

The system is able to deal with any query terms, and output candidate sections according to the cumulative distance of query terms.

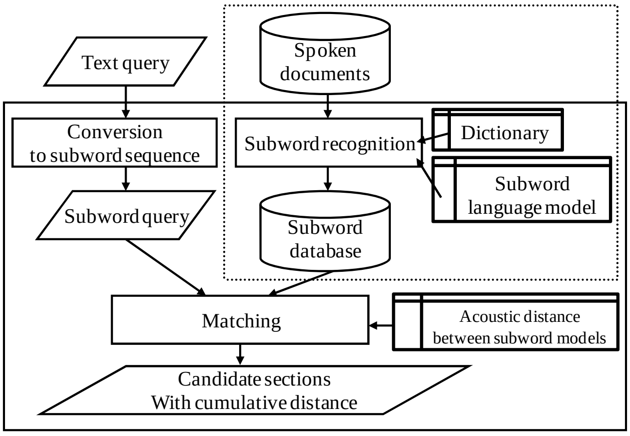


Figure 1. Outline of STD system for OOV query term.

Proposed method to integrate plural STD results using plural subword language models

Figure 2 illustrates the proposed method to integrate plural STD results using different subword language models. Subword speech recognition is performed using plural subword language models in parallel, as shown in the figure using three language models. Plural subword recognition results and their STD results including candidate sections with their cumulative distances are obtained. The plural distances for each section (for each utterance) are integrated linearly into an integrated distance. Each candidate section is re-ranked according to the integrated distance.

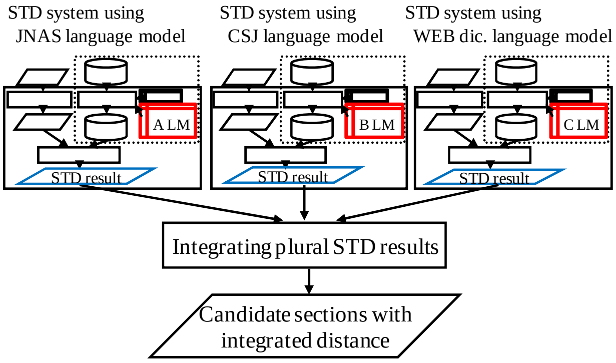

STD system using JNAS language model
STD system using CSJ language model
STD system using WEB dic. language model
A LM
B LM
C LM
STD result
STD result
STD result
Integrating plural STD results
Candidate sections with integrated distance

Figure 2. Proposed method for integrating plural STD results using different subword language models (LM).

Because several subwords are available for STD, we have proposed to a linear integration method of plural results obtained from plural subwords to improve the STD performance [5]. An integrated distance D_l for the l-th candidate section in spoken documents is computed by summing up the weighted minimum cumulative distance for each k-th subword, as follows:

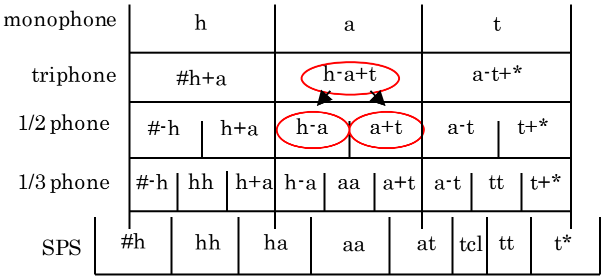
$$D_l = \sum_{k=1}^{N} weight_l(k) \times distance_l(k)$$

$$\sum_{k=1}^{N} weight_l(k) = 1 \qquad \textbf{(y} \qquad (1)$$

where N represents the number of subwords. For the proposed method, k and N represents the k-th language model and the number of language models. The first, second and third result are obtained by using the JNAS, CSJ, and WEB dictionary language models, respectively. After computing the integration distance for all candidate sections, they are ranked by D_l, and the results are presented to a user in the ranked order.

Subword models

Monophone, triphone, 1/2phone, 1/3phone and SPS [9] are used for subwords in the paper. Figure 3 represents each subword expression and the conceptions of the subword model boundary for the three monophone model sequence “h a t”. Each triphone model is divided into two 1/2 phone models: a model of the front part and a model of the rear part, as shown in Figure 3. A triphone model is divided into three 1/3 phone models. Because 1/2 phone, 1/3 phone, and SPS models have more models to represent the same word, they are considered to be more sophisticated models in the time axis, and these subwords were confirmed to work well for STD than monophone and triphone [4]. The number of models is also shown in Table 1.



(“#” and “*” represent the previous and next phone label, respectively)

Figure 3. Each subword expression and the conceptions of the subword model boundary for the three monophone model sequence “h a t”

Table1: The number of subword models in each subword

Subword	**Number of models**
monophone	**43**
triphone	**7956**
1/2phone	**1333**
1/3phone	**1374**
SPS	**423**

Subword language models

Three subword language models are trained using subword model sequences included following data sets.

(1) JNAS

The JNAS (Japanese News Article Sentences) [6] is open speech corpus including read speeches of news articles with their pronunciations. JNAS consists of newspaper article sentences: 155 text sets (about 100 sentences per set), 16176 sentences in total and ATR phonetically balanced sentences: 10 text sets (about 50 sentences per set), 503 sentences in total.

(2) CSJ

The CSJ (Corpus of Spontaneous Japanese) is one of the largest spoken language databases in the world. It contains 2702 presentation speeches that amount to 658 hours, approximately 7.5 million words spoken by more than 1,400 speakers. The CSJ consists of mainly spontaneous monologues, such as academic presentations and public speaking and spontaneous dialogues. CSJ includes not only transcriptions but also a rich set of annotations, including parts of speech, labels of phonetic segmentation and intonation.

(3) WEB dictionary

We have been collecting words with their pronunciation on Web sites, including Wikipedia, Hatena keyword home pages and so on. The WEB dictionary includes approximately 1.2 million words so far.

EVALUATION EXPERIMENTS

Experimental data sets

For test data set, we use 49 presentation speeches spoken by 49 male speakers that amount to about 13 hours. Each utterance divided by a silence section was converted to a subword model sequence by mean of subword recognition. We used 50 query words and each query has three to 50 corresponding sections in the test data. This data set is provided by the SIG-SLP (Special Interest Group – Spoken Language Processing) of Information Processing Society of Japan [10] that is constructing Japanese standard SDR and STD test collections. The acoustic models for the subword recognition were trained by male spoken data in JNAS in this paper.

The three language models mentioned in the previous chapter were trained by (1) JNAS male data set, (2) 1054 male presentation data set, and (3) the WEB dictionary including 1.2 million words. The number of monophone included in the three training data sets is shown in Table 2.

Table 2: The number of monophones in training data sets

Training data set	The number of monophones
JNAS	1,505,847
CSJ	11,519,416
WEB dictionary	17,657,409

Evaluation methods

We used the mean average precision (MAP) rate for evaluation measurements [4]. Precision is the fraction of the sections detected that are correct for the query terms, and is obtained by Equation (2). Average precision for a query q is the average of precisions computed at the point of each correct section for q in the ranked candidates as shown in Equation (3), where k is the rank, δ_k is equal to 1 or 0 when k-th candidate is correct or incorrect, respectively, R the rank of the last correct section, N the number of correct sections. Lastly, MAP is obtained by computing the mean of average precision over all Q queries. It gives an overall evaluation of a retrieval algorithm.

$$precision(k) = \frac{\#\ of\ \ correct\ in\ k - th\ candidate}{k} \times 100 \qquad (2)$$

$$average\ \ precision(q) = \frac{1}{N}\sum_{k=1}^{R}\delta_k \times precision(k) \qquad (3)$$

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q} average\ \ precision(q) \qquad (4)$$

Results and discussions

Table 3 shows the STD results using a single language model for each subword. The performance using monophone was much lower than that using other models. The number of

Table 3: STD performance using a single language model for each subword

Language model Subword	JNAS	CSJ	WEB dictionary
monophone	36.96	37.43	33.36
triphone	49.04	61.63	54.02
1/2 phone	75.51	74.85	74.55
1/3 phone	64.36	64.71	66.45
SPS	70.64	71.44	70.01

Table 4: STD performance obtained by the proposed method integrating plural results using plural language model for each subword

Subword	Single LM (best)	JNAS + CSJ	Single LM (best)	JNAS + CSJ + Web dic.
monophone	37.43	38.91	37.43	39.56
triphone	61.63	62.05	61.63	63.79
1/2 phone	75.51	76.99	75.51	76.95
1/3 phone	64.71	65.97	66.45	66.94
SPS	71.44	73.16	71.44	72.90

monophone models was too small, and monophone was inferior in a discriminating ability. On the contrary, the number of triphone was large, and the performance using the JNAS language model was thought to deteriorate. The performance improvement was remarkable for triphone using the CSJ language model. This is because the amount of the training data of JNAS was small and the triphone language model was not trained sufficiently against the large number of triphone models (about 8,000 models), as shown in Table 2.

Table 4 shows the STD performance that was obtained by the proposed method, integrating plural results using plural language model for each subword. The left side denotes the case using two language models; the JNAS and CSJ language model. “Single LM (best)” in the table corresponds to the better performance using a single subword model among these two language models. The right side denotes the case using three language models; the JNAS, CSJ, and WEB dictionary language models. “Single LM (best)” corresponds to the best performance among these three language models.

The weighting factor *weight*$_l$*(k)* for the *k*-th language model in Equation (1) was determined by a 5-fold cross-validation method. The test data were divided into 5 subsets. One of 5 subsets was used for the actual test data, and the rest of 4 subsets were used for determining the weighting factor. We tested each subset, and performed fair evaluations. The STD performance improved in all cases, as shown in Table 4, although the improvement was not so remarkable. The best performance at 76.99 % of MAP was obtained when using 1/2 phone, the JNAS and CSJ language models. The WEB dictionary also contributed the STD performance when using monophone, triphone and 1/3 phone. The similar tendency was observed for other combinations such as the CSJ and WEB dictionary language models. These results demonstrated the effectiveness of the proposed method that integrates plural results using plural language models for subword-based speech recognition.

When using a single language model, the processing time for a query was approximately 0.3 s for 13-hour spoken document. Because the processing time is mainly due to the matching process, it is proportional to the number of language models used in the proposed method. We think it is possible to reduce the processing time by parallelizing the matching process, as illustrated in Figure 2.

We confirmed the effectiveness of intergrating plural results obatained by using plural subword models in [5]. Therefore, we are going to integrate plural results using plural subwords plural subword language models, and plural acoustic models. When integrating plural results, the method for determining the weighting factor in Equation (1) is important. In the paper, the weighting factor was determined by a 5-fold cross validation method. We are seeking to determine it automatically by using confidence measures obtained from speech recognition or by introducing a difficulty measure for each query term.

CONCLUSIONS

The present paper proposed an integration of plural STD results that are obtained from plural language models. We used three types of language models that were trained by the JNAS, CSJ, WEB dictionary data, respectively. Experimental results demonstrated that the performance could be improved by integrating plural STD results by using plural language models, and showed the effectiveness of the proposed method. We are going to improve the STD performance by integrating plural results using plural subwords and plural subword language models et al., and seek to determine a weighting factor automatically using confidence measures obtained from speech recognition. The reduction of the processing time of the proposed system is also a future work.

ACCKNOWLEGEMENTS

This research is supported by Grand-in-Aid for Scientific Research (C) Project No. 20500096 and (C) Project No. 22500145, Japan Society for Promotion of Science.

REFERENCES

1 Jonathan Mamou, et al, “Vocabulary Independent Spoken Term Detection”, *Proc. of SIGIR’07*, pp. 615-622, 2007

2 Roy Wallace, et al, “A Phonetic Search Approach to the 2006 NIST Spoken Term Detection Evaluation”, *Proc. of INTERSPEECH*, pp. 2385-2388, 2007

3 Naoyuki Kanda, et al, “Open-Vocabulary Keyword Detection from Super-Large Scale Speech Database”, Proc. of *IEEE MMSP*, pp939-944, 2008

4 Iwata, K., et al, “Open-Vocabulary Spoken Document Retrieval based on new subword models and subword phonetic similarity,” *INTERSPEECH*, 2006

5 Yoshiaki Itoh, et al, “An Integration Method of Retrieval Results using Plural Subword Models for Vocabulary-free Spoken Document Retrieval”, *Proc. of INTERSPEECH,* pp. 2389-2392, 2007.

6 Katunobu Itou et al, “JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research”, *J. Acoust. Soc. Jpn, (E), Vol. 20-3*, pp.199-206, 1999

7 Kikuo Maekawa, "Corpus of Spontaneous Japanese: Its Design and Evaluation", *Proc. of ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, pp.7-12, 2003

8 Tatsuya Kawahara et al, “Recent Progress of Open-Source LVCSR Engine Julius and Japanese Model Repository”, *In Proc. International Conference on Spoken Language Processing*, pp. 688-691, 2004

9 Tanaka. K, et al, "Speech data retrieval system constructed on a universal phonetic code domain”, *IEEE ASRU’01*, pp.323-326, 2001

10 Akiba, Tomoyosi et al., “Developing an SDR test collection from Japanese lecture audio data,” APSIPA, 2009.