An STD system for OOV query terms integrating multiple STD results of various subword units

Kazuma Kon'no
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g231k015@s.iwate-pu.ac.jp

Hiroyuki Saito
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g231i027@s.iwate-pu.ac.jp

Shirou Narumi
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g231k030@s.iwate-pu.ac.jp

Kenta Sugawara
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g031h079@s.iwate-pu.ac.jp

Kesuke Kamata
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g031h038@s.iwate-pu.ac.jp

Manabu Kon'no
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g031g063@iwate-pu.ac.jp

Jinki Takahashi
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g031i107@iwate-pu.ac.jp

Yoshiaki Itoh
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
y-itoh@iwate-pu.ac.jp

ABSTRACT

We have been proposing a Spoken Term Detection (STD) method for Out-Of-Vocabulary (OOV) query terms integrating various subword recognition results using monophone, triphone, demiphone, one third phone, and Sub-phonetic segment (SPS) models. In the proposed method, subword-based ASR (Automatic Speech Recognition) is performed for all spoken documents and subword recognition results are generated using subword acoustic models and subword language models. When a query term is given, the subword sequence of the query term is searched for all subword sequences of subword recognition results of spoken documents. Here, we use acoustical distances between subwords when matching the two subword sequences by Continuous Dynamic Programming. We have also proposed the method re-scoring and integrating multiple STD results obtained using various subword units. Each candidate segment has a distance, the segment number and the document number. Re-scoring is performed using distances each of high ranked candidate segments, and the last distance is obtained by integrating then linearly using weighting factors. In STD tasks (SDPWS) of IR for Spoken Documents in NTCIR-10, we apply various subword models to the STD tasks and integrate multiple STD results obtained from these subword models.

Categories and Subject Descriptors

I.2.7 [**ARTIFICIAL INTELLIGENCE**]: Natural Language Processing – Speech recognition and synthesis, *Text analysis*.

General Terms

Algorithm

Team name

IWAPU

Subtasks

Spoken Term Detection (moderate-size task)

Keywords

[IWAPU], [Japanese], [Spoken Term Detection], [SDPWS], subword model, multiple model integration.

1. INTRODUCTION

According to the rapid progress of information technology and the increase of the capacity of the recording mediums such as a hard disk or an optics disk in these years, every user comes to have much opportunity to deal with multimedia data such as video data that are available on such hard disk video recorders or the Internet. Recently, SDR (Spoken Document Retrieval) and STD (Spoken Term Detection) have been hot topics among speech processing researchers to deal with such enormous amount of data that are regarded as spoken documents [1]-[3]. In case of a common STD system, it generates a transcription of speech data using a large vocabulary continuous speech recognition (LVCSR) system, and finds query terms in the transcription. Although the method is advantageous in finding In-Vocabulary (IV) query terms at high speed, it has a difficulty in detecting Out-Of-Vocabulary (OOV) query terms that are not included in a dictionary of the LVCSR system, because OOV terms in spoken documents are inevitably substituted to other words in the dictionary. STD systems must be able 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 the detection of OOV query terms, a method using subword such as monophone and triphone is representative [4][5], and we have proposed STD methods for OOV query terms using various subword units, such as monophone, triphone, demiphone, one third phone, and SPS models. For each subword model, the system compares a query subword sequence with all of the subword sequences in the spoken documents and retrieves the target segments using Continuous Dynamic Programming (CDP) algorithm. Here, we introduce a phonetic distance between any two subword models for a local distance in CDP. Re-scoring is performed to improve the retrieval performance after CDP. We performed again CDP using pseudo query term that is sequence of high ranked candidate segments that has reliable information. Detail is shown reference [6]. Though we have confirmed new subword models worked well, the retrieval performance for each query word does not always show the same tendency as their average performance. Therefore we have also proposed the method integrating these multiple STD results to improve the STD performance [7]. We apply the most of the methods mentioned above to the STD tasks of IR for Spoken Documents in NTCIR-10 [8]. We use various subword models such as monophone, triphone, syllable, demiphone, and SPS. Phonetic distances between subword models are applied at a CDP process. Multiple STD results obtained from these subword models are integrated. Furthermore, we improve the performance by applying a longer N-gram language model. The performance is evaluated according to the criteria that the organizer provided.

The present paper describes the outline of our system first, and then our subword models, their acoustic models and language models are explained. Next, the integration method of multiple STD results is explained in detail after the explanation of subword based STD process using a single subword model and phonetic distances for a local distance of CDP and re-scoring. In Chapter 3, the performance of the proposed method is evaluated for the test collection of NTCIR-10. Lastly, conclusion is presented.

2. PROPOSED METHODS

In the proposed system, subword acoustic models, their language models, a subword distance matrix, and subword recognition results of spoken documents are prepared beforehand [9].

First, subword recognition is performed for all of the spoken documents and a subword sequence database is prepared beforehand (1). Here, subword language models are used, such as subword bigrams and trigrams and so on. The system allows both text and speech queries (2). When a user inputs a text query, the text is automatically converted to a subword sequence according to conversion rules (3). In case of Japanese, the phone sequence to be pronounce of a query term is automatically obtained when a user input a query term. For each subword model, the system then retrieves the target segment using Continuous DP algorithms by comparing a query subword sequence to all of the subword sequences in the spoken documents (4).

Conference'10, Month 1–2, 2010, City, State, Country.

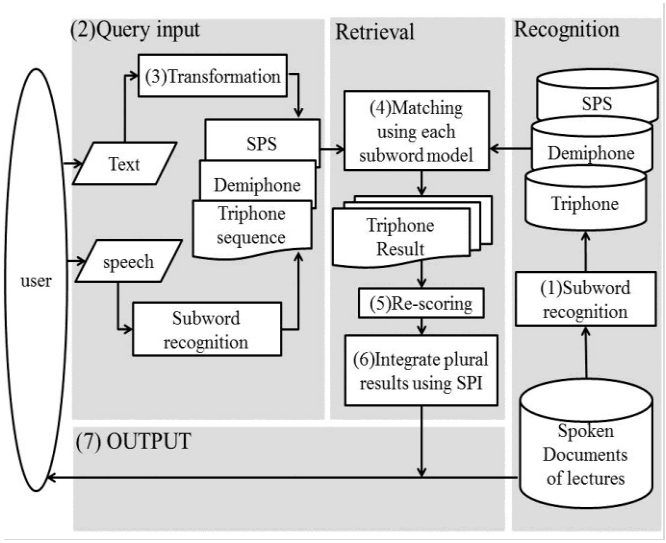

Figure 1: Outline of the STD method using multiple subword recognition results

The local distance refers to the distance matrix that represents the subword dissimilarity and contains the statistical distance between any two subword models. The system outputs multiple candidate segments that show a high degree of similarity to the query word for each subword model. Each candidate segment has a distance and a segment number of spoken documents. A distance obtained by CDP is re-rescoring in each lecture (5). A distance is computed by integrating the of multiple subword models for all candidates segments, and candidate segments are re-ranked (6).

In the following section, the proposed integration method and re-scoring method are described in detail after briefly explaining the subword models used in the present paper.

2.1 Subword Models

This section describes subword models used in the paper. Four kinds of subword models, that is, monophone, triphone, 1/2 phone [10] and the sub-phonetic segment (SPS) [11] are used for subwords in the paper. These subword models and their sample descriptions of a monophone sequence " a k i" " for each subword are shown in Figure 2. Triphone is divided into two demiphone models: a model of the front part and a model of the rear part, as shown in Figure 2. SPS models are designed so that they represent physical characteristics of pronunciation of consecutive phonemes. Demiphone and SPS models are regarded to be more sophisticate models in the time axis, because they have more models to represent the same word than monophone and triphone models. These subword models were confirmed to work well for STD [10].

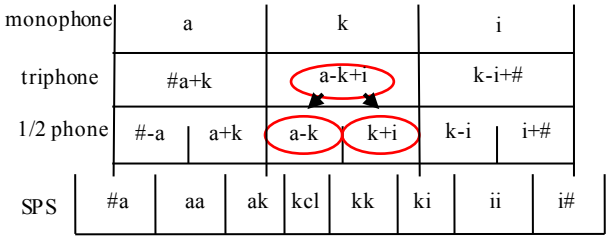

Figure 2: Subword models and "a k i" expressions.

Table 1 : Conditions of feature extraction for acoustic models.

Sampling	16 kHz 16 bit
Feature Parameter	12-dim. MFCC+ energy 12-dim. Δ MFCC+Δenergy 12-dim. ΔΔMFCC+ΔΔenergy
Window Length	25 ms
Frame Shift	10 ms for monophone and triphone 5 ms for demiphone and SPS

2.2 Acoustic Models and Language Models

This section describes subword acoustic models and subword language models used in the paper. The conditions of feature extraction for acoustic models are listed in Table 1. The speech data of an actual presentation corpus of CSJ (Corpus of Spontaneous Japanese) are used for training data. The speech data were segmented based on an XML file. The analysis window length was 25ms. The frame shift was 10 ms for monophone and triphone, and 5 ms for demiphone and SPS. A 38-dimensional MFCC feature vector is used for training acoustic models, as shown in Table 1. All of the acoustic models were trained using the Hidden Markov Model Toolkit (HTK) [12].

The training data for subword language models are the same CSJ data as those for acoustic models. Subword bigram and subword trigram are used for language models. All of the language models were trained by the Parm Kit [13] was used as a training tool.

We use two types of recognition results for triphone models: one is our triphone-based language models and the other is syllable-based language models. Both LMs are generated by using CSJ.

The CSJ includes 2702 lecture speeches in total, and is divided into three parts in the NTCIR: CORE includes 177 lecture speeches, Odd and Even include about 1265 lecture speeches except CORE respectively. We trained each subword model using Even lecture speech data, because of our time limitation.

2.3 Matching using Single Subword and Local Distances [14]

For each subword model, the distance $D(i, j)$ is computed between a query Q_i and a segment of a spoken document or speech segment S_i. Here, i, and j denote a query number and a segment number of spoken documents, respectively. We use CDP (Continuous Dynamic Programming) for matching the subword sequences of spoken document and a query subword sequence. Although an edit distance is representative for a local distance in string matching, we have proposed a phonetic distance between subwords so far [5]. A phonetic distance represents the statistical dissimilarity between subwords and the phonetic distance matrix contains all the distances between any two subword models. In the CDP process, local distances only have to refer to the matrix. The system outputs candidate segments according to the distances that show a high degree of similarity to the query word. Each candidate segment has a distance and a segment number of spoken documents.

To improve the STD performance, we modify the method computing a phonetic local distance between subwords when computing a local distance between states in Hidden Markov Models statistically. The method referring to adjacent states. We call the former as "adjacent states reference", as shown in figure 4.

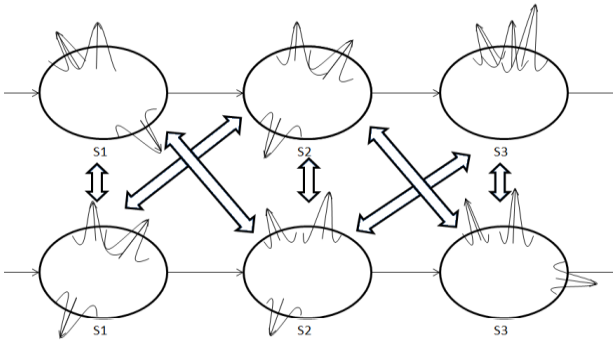

Figure4: The reference between adjacent states

2.4 STD Results Obtained from Multiple Subword Models[7][9]

Each subword model m ($1 \le m \le M$) generates the distance $D_m(i, j)$ between a query Q_i ($1 \le i \le I$) and segment of a spoken document or a speech segment S_j ($1 \le j \le J$) and Here, M, I, and J denote the number of subword models, the number of queries, and the number of spoken segments, respectively. We have proposed a linear integration method for multiple retrieval results obtained from multiple subword recognition using various subword models to improve STD performance. The modified distance $D(i, j)$, which is a new criteria, is obtained by integrating the distances $D_m(i, j)$, according to the following equation. Here, $weight_m$ is a weighting factor for the m-th subword STD result.

$$D(i,j) = \sum_{m=1}^{M} weight_m \times D_m(i,j) \qquad (2)$$

The image of the integration of two retrieval results ($M = 2$) is shown in Figure 6. The STD results A and B are obtained in parallel. Each candidate segment has the segment number and a matching distance, such as $D_A(q_i, 1)$, $D_B(q_i, 1)$ for the i-th query q_i and the first candidate segment in spoken documents. An integrated distance $D(q_i, 1)$ for the first candidate segment is computed by summing the weighted distances for each result. After computing the integrated distance for all candidate segments, the segments are re-ranked according to the integrated distance $D(q_i, 1)$, and the results are presented to a user in the ranked order.

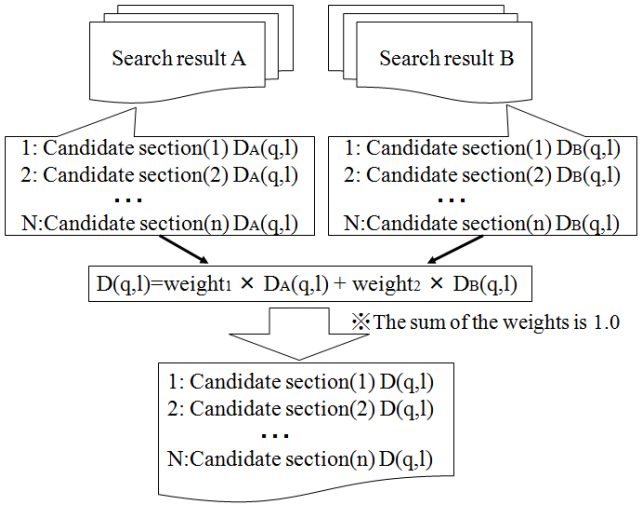

Figure 6: A image of integrating two STD results

2.5 Threshold for F-measure Optimization

For the threshold to optimize an F-measure, we simply use the integrated distance described in the previous section. If the integrated distance becomes less than a constant threshold value, the candidate segment is provided. The setting of the threshold was determined experimentally and was not optimized due to our time limitation.

3. EVALUTION EXPERIMENTS

3.1 Test Data and Evaluation Measurement

The test data in the experiments were the test collection of unknown terms for SDPWS data in the formal run. The number of the queries is 200 that the organizer of NTCIR-10 provided. We used triphone, demiphone, and SPS for subword models. One best recognition result is only used through the experiments for each subword model. We use MAP (Mean Average Precision) [15].

3.2 Performance using a Single Subword

Table 2: Performance using a single subword and applied re-scoring

MODEL	TRAIN DATA	Acoustic Distance (Mixture)	MAP (formal run)[%]	
			SINGLE	RE-SCORING
Triphone syllable LM	ALL	2	48.07	54.81
		4	48.66	54.89
	EVEN	2	49.90	55.75
		4	50.24	56.27
	MALE	2	50.43	55.81
		4	50.48	55.68
Triphone Subword LM	ALL	2	45.38	50.89
		4	45.57	52.02
	EVEN	2	48.21	53.53
		4	48.07	53.66
	MALE	2	49.97	56.69
		4	49.32	56.11
SPS	ALL	2	39.72	45.20
		4	39.09	44.09
	EVEN	2	39.15	42.65
		4	38.88	43.33
	MALE	2	44.35	48.93
		4	44.49	48.93
demiphone	ALL	2	50.08	56.78
		4	50.78	57.51
	EVEN	2	51.60	56.26
		4	52.29	58.03
	MALE	2	54.95	59.77
		4	54.02	61.17

Table 2 shows our STD performance using a simple subword. Demiphone showed the best performance among three subwords.

By integrating multiple results obtained using different subwords, the STD performances were improved in all cases as shown in Table 3. We believe it is because the different STD results worked complementally and the other subword could make up with the STD fails of one subword.

Table 3: Performance using multiple STD results.

No	MODEL	MAP(%)
1	Triphone syllable LM ALL (2mix + 4mix)	55.40
2	Triphone syllable LM EVEN (2mix + 4mix)	56.96
3	Triphone syllable LM MALE (2mix + 4mix)	56.39
4	Triphone subword LM ALL (2mix + 4mix)	52.19
5	Triphone subword LM EVEN (2mix + 4mix)	53.66
6	Triphone subword LM MALE (2mix + 4mix)	56.78
7	SPS ALL (2mix + 4mix)	44.89
8	SPS EVEN (2mix + 4mix)	43.33
9	SPS MALE (2mix + 4mix)	49.16
10	demiphone ALL (2mix + 4mix)	57.22
11	demiphone EVEN (2mix + 4mix)	57.58
12	demiphone MALE (2mix + 4mix)	60.92
13	2+3 Triphone syllable LM (EVEN + MALE)	60.39
14	13+1 Triphone syllable LM (EVEN + MALE+ALL)	62.28
15	7+8 SPS (ALL + EVEN)	46.78
16	15+9 SPS (ALL + EVEN + MALE)	50.40
17	10+12 demiphone (ALL + MALE)	62.51
18	17+11 demiphone (ALL + MALE + EVEN)	62.98
19	14+18 Triphone syllable LM + demiphone	67.98
20	19+16 Triphone subword LM + demiphone + SPS	67.66
21	20+6 Triphone syllable LM + demiphone + SPS+ Triphone subword LM	67.82

Table 4: Evaluation results provided by the organizer.

MODEL	MAP(%)	Term/ System
triphone syllable LM + demiphone + SPS+ Triphone subword LM	67.5	IWAPU-1

The organizer evaluated our results, and MAP was 67.5% in "IWAPU-1" show in Table 4. The difference of the best MAP values in Table3 (67.82 %) and Table 4 (67.5%) lies in the difference between the oracle hit file of the organizers and ours. The oracle hit file was not provided at the formal run.

4. CONCLUSIONS

We constructed an STD system using our proposing methods that include the introduction of new subwords such as demiphone, SPS, the integration multiple STD results obtained using various subword units and re-scoring method. In the experiment using the data of the formal run of STD tasks in IR for Spoken Documents of NTCIR-10, the STD performance could be improved by 12.55 points compared with that using a single subword, and a result of 67.5 % in MAP was achieved at the formal run.

5. ACKNOWLEDGMENTS

This research is supported by Grand-in-Aid for Scientific Research (C) Project No. 24500124.

6. REFERENCES

[1] 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.

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

[3] Petr Motlicek, Fabio Valente, Philip N, "Garner English Spoken Term Detection in Multilingual Recordings", INTERSPEECH 2010, pp.206-209, 2010..

[4] 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.

[5] Roy Wallace, et al, "A Phonetic Search Approach to the 2006 NIST Spoken Term Detection Evaluation", INTERSPEECH 2007, pp2385-2388, 2007.

[6] Kazuma.K, et al, "Re-ranking of candidates using highly ranked candidates in Spoken Term Detection", ASJ , pp.191-194, 2012-9.

[7] Yoshiaki Itoh, et al, "An Integration Method of Retrieval Results using Multiple Subword Models for Vocabulary-free Spoken Document Retrieval", Proc. of INTERSPEECH 2007, pp2389-2392, 2007.

[8] T. Akiba, et al. : Overview of the NTCIR-10 SpokenDoc-2 Task, Proceedings of the NTCIR-10 Conference, 2013

[9] Yuji Onodera et al, "Spoken Term Detection by Result Integration of Multiple Subwords using Confidence Measure", WESPAC, 2009

[10] Iwata K, et al, "An Investigation of New Subword Models and Subword Phonetic Distance for Vocabulary-free Spoken Document Retrieval System", IPSJ Journal, Vol.48, No.5, pp. 1990-2000, 2007

[11] Tanaka. K, et al, "Speech data retrieval system constructed on a universal phonetic code domain", ASRU'01 IEEE, pp.323-326, 2001.

[12] HTK, http://htk.eng.cam.ac.uk/

[13] palmkit, http://palmkit.sourceforge.net/.

[14] Fumitaka.T, et al "Improving perfomance of spoken term detection by appropriate distance between subwoed models", ASJvol2, pp.239-240, 2011-3.

[15] Tomoyosi Akiba, Hiromitsu Nishizaki, Kiyoaki Aikawa, Tatsuya Kawahara, Tomoko Matsui.Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop. Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2011.