An STD system for OOV query terms using various subword units

Hiroyuki Saito
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g231j018@s.iwate-pu.ac.jp

Takuya Nakano
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
g031e133@s.iwate-pu.ac.jp

Toshiaki Chiba
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g031g110@s.iwate-pu.ac.jp

Kazuma Kon'no
Iwate Prefectural University
Sugo 152-52, Takizawa,
Iwate, Japan
+81-19-694-2556
g031g062@s.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 using various subword units, such as monophone, triphone, demiphone, one third phone, and Sub-phonetic segment (SPS) models. In the proposed method, subword-based ASR 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 in Continuous Dynamic Programming. Demiphone and one-third phone models were newly developed for an STD task. We have also proposed the method integrating plural STD results obtained using each subword models. Each candidate segment has a distance, the segment number and the document number. These plural distances are integrated linearly using weighting factors. In STD tasks of IR for Spoken Documents in NTCIR-9, we apply various subword models to the STD tasks and integrate plural 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

Keywords

[IWAPU] [Japanese] spoken term detection, subword model, plural 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. 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 plural STD results to improve the STD performance [6]. We apply the most of the methods mentioned above to the STD tasks of IR for Spoken Documents in NTCIR-9. We use various subword models such as monophone, triphone, syllable, demiphone, and SPS. Phonetic distances between subword models are applied at a CDP process. Plural 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 single subword model and phonetic distances for a local distance of CDP. In Chapter 3, the performance of the proposed method is evaluated for the test collection of NTCIR-9. 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 [7].

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 speech queries, the system performs subword recognition and transforms the speech into a subword sequence in the same manner as spoken documents (4). 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 (5). 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 plural 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 new distance is computed by integrating the of plural subword models for all candidates, and candidate segments are re-ranked (5).

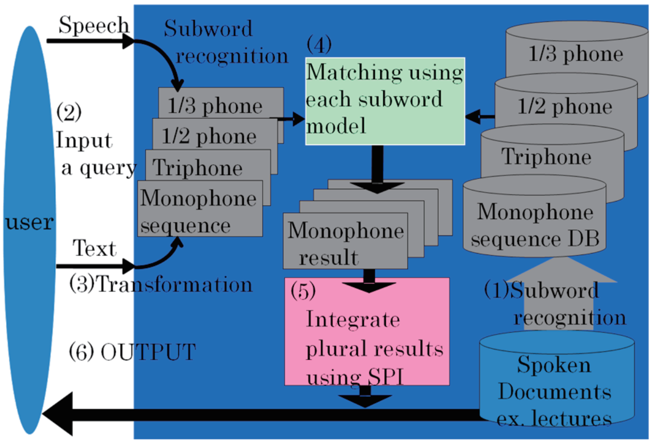

Figure 1: Outline of the STD method using plural subword recognition results

In the following section, the proposed integration method is 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 [8] and the sub-phonetic segment (SPS) [9] 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 [8].

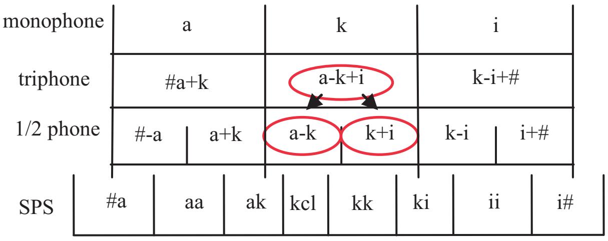

Figure 2: Subword models and “h a t” expressions.

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) [10].

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[11] was used as a training tool.

We use three types of recognition results for triphone models: one is obtained using our triphone-based language model and the others are obtained using two syllable-based language models. One is supplied by the organizer and the other is generated by ourselves, which is Intensive triphone models described in the next section.

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

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

2.3 Intensive Triphone Acoustic Model[12]

The number of physical triphone models is approximately 8,000 in Japanese. We are seeking a more suitable number of triphone models for STD tasks, and developing intensive triphone acoustic models that are to improve the STD performance. Figure 3 illustrates the image. When matching a subword sequence of query terms and subword sequences of spoken documents, some miss-match parts appear at the section of the query term in spoken documents utterance due to subword recognition errors. As a result, the detection of OOV query terms fails. Acoustically similar subword models, therefore, such as /m/ and /n/ are put together into a cluster so that some recognition errors can be recovered. In the cluster, we choose a representative subword model such as /m/ and /d/ in the figure.

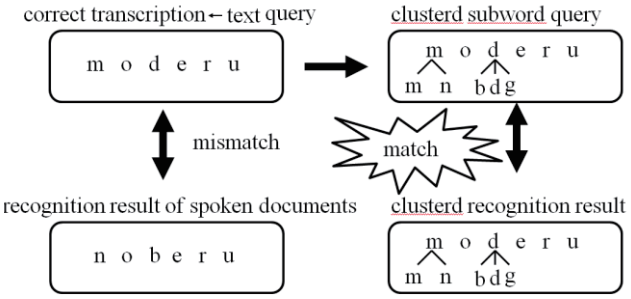
correct transcription←text query
m o d e r u
clusterd subword query
m o d e r u
m n bdg
mismatch
match
recognition result of spoken documents
n o b e r u
clusterd recognition result
m o d e r u
m n bdg

Figure3: Image of matching based on representative phones

On the other hand, when the number of subword models becomes too small, such too intensive models show the performance deterioration because of the poor discriminative ability. We investigate the appropriate number of subword models for STD tasks.

2.4 Matching using Single Subword and Local Distances [13]

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. Two methods are developed. One is the method referring to adjacent states, and the other is the method referring all states. We call the former one as “adjacent states reference" and the latter as "all states reference", as shown in figure 4 and 5, respectively. After computing both distances of adjacent states reference and all states reference, we integrate the both distances for a local distance linearly in this experiment.

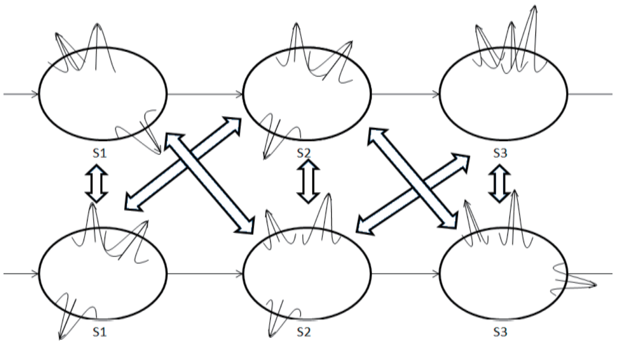

Figure4: The reference between adjacent states

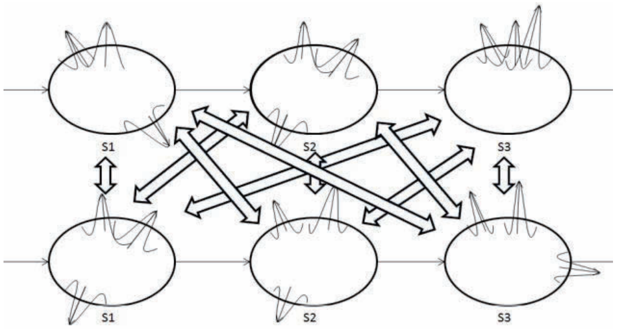

Figure5: The reference between all states

2.5 Integration of Plural STD Results Obtained from Plural Subword Models[6][7]

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 plural retrieval results obtained from plural 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.

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

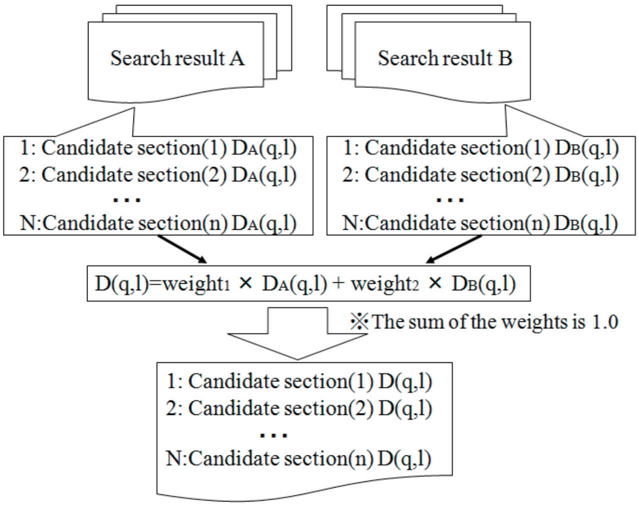
Search result A
Search result B
1: Candidate section(1) DA(q,l)
2: Candidate section(2) DA(q,l)
・・・
N:Candidate section(n) DA(q,l)
1: Candidate section(1) DB(q,l)
2: Candidate section(2) DB(q,l)
・・・
N:Candidate section(n) DB(q,l)
D(q,l)=weight1 × DA(q,l) + weight2 × DB(q,l)
※The sum of the weights is 1.0
1: Candidate section(1) D(q,l)
2: Candidate section(2) D(q,l)
・・・
N:Candidate section(n) D(q,l)

Figure 6: A image of integrating two STD results

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.

2.6 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 supplied. 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 CORE data in the formal run. The test data included 177 presentation speeches that total approximately 44 hours. Each presentation is spoken by a different speaker. The number of the queries is 50 that the organizer of NTCIR-9 provided in the test data. We used triphone, demiphone, and SPS for subword models. In addition, we also used triphone recognition results that were provided by the organizer. One best recognition result is only used through the experiments.

In the paper, we use MAP (Mean Average Precision) and F-measure for the evaluation measurements [14].

3.2 Performance using a Single Subword

Table 2 shows our STD performance using simple subword. The center column depicts the performances that are submitted at the formal run. SPS showed the best performance among four subwords. The performance of our triphone did not reach to that of the organizer's triphone. Because the language model of our triphone was triphone trigram and that of the organizer's was syllable trigram, the constraint of the organizer's language model was stronger than that of our language model. Although the language models for SPS and demiphone are not strong either, the performance of SPS and demiphone was better than that of the organizer's triphone.

Table 2: Performance using a single subword

MODEL	MAP (formal run)[%]	MAP (best)[%]
Triphone	50.66	-
Intensive triphone	-	76.00
Demiphone	71.24	72.30
SPS	73.18	74.89
Triphone by organizer	69.87	73.69

The right column depicts the performances that were improved after the formal run to the current paper. Though SPS showed the best performance among the four subwords tested in the formal run, Intensive triphone that is our newly developed triphoe model outperformed SPS, and reached 76 % in MAP. The STD performances of demiphone, SPS, and triphone by the organizer were improved by the introduction of new local distances described in 2.4. We could confirm the new local distances and Intensive triphone worked well.

Table 3 shows the STD performance when integrating two or three STD results that are obtained using Demiphone, SPS and Intensive triphone. which shows the best performance in Table 2

Table 3 is STD performance when we used each subword model the shawe the "best" of Table 2.

Table 3: Performance using plural STD results by using demiphone, SPS, and Intensive triphone.

MODEL	F-measure(%)	MAP(%)
Demiphone + SPS	63.14	76.27
Demiphone + Intensive triphone	64.91	79.69
SPS + Intensive triphone	65.26	80.53
Demiphone+ SPS+ Intensive triphone	65.81	80.80

By integrating plural results obtained using different subwords, the STD performances were improved in all cases. 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.

In case of integrating two subwords, the STD performance was improved by 4.53 points at maximum in comparison with that using a simple subword when using SPS and Intensive triphone, and reached more than 80 % In case of integrating three subwords, the STD performance was improved by 4.8 points compared with that using a simple subword, and reached 80.8 % in MAP.

Table 4: Evaluation results provided by the organizer.

MODEL	F-measure (%)	MAP (%)	Term/ System
Demiphone+ SPS+ Triphone by organizer	62.8	77.2	A/1
SPS	29.7	73.3	A/2

Table 4 shows the submitted STD performance at the formal run. We submitted two results. One is the result that showed the best performance of our experiments integrating results of three subwords: demiphone, SPS and the triphone by the organizer. This result was regarded as priority one. The other is the result using a single subword, or SPS, which showed the best performance using single subword with one-best recognition result. SPS showed more than 70% at MAP and a better performance than the triphone by the organizer.

The organizer evaluated our results, and MAP was 77.2% in "IWAPU-1" that corresponds to A / 1, and 73.3% in "IWAPU-2" that corresponds to A / 2. The difference of MAP values in Table3 (73.18%) and Table 4 (73.3%) lies in the difference of the oracle hit files, which was not provided at the formal run and we had to made it by ourselves.

We performed further improvement by integrating four STD results and introducing new local distances, and succeeded in improving the STD performance as shown in Table 5.

Table 5: The best performance using four results and introducing new local distances

MODEL	F-measure(%)	MAP(%)
Demiphone + SPS + Intensive triphone + Triphone by organizer	65.81	81.43

The improvement was due to the introduction of the new distance described in 2.4 and the integration of four results obtained using demiphone SPS intensive triphone and triphone of organizer's. The STD performance was improved 5.43 points compared with that using a simple subword, and 4.23% improvement in comparison with that at the formal run. The major contribution of the performance improvement was due to he new local distances, comparing with that using three subwords shown in Table 3.

4. CONCLUSIONS

We constructed the STD systems using our proposing methods that include the introduction of new subwords such as demiphone, SPS and intensive triphone, the integration plural STD results obtained using various subwords and new phonetic distances between subword models in CDP, and so on. In the experiment using the data of the formal run of STD tasks in IR for Spoken Documents of NTCIR-9, the STD performance could be improved by 4.53 points compared with that using a single subword, and a result of 77.2 % in MAP was obtained at the formal run. After improving our system that integrates four STD results and uses new local distances, the STD performance could be furthermore improved by 4.23 points in comparison with that at the formal run and by 5.43 points in comparison with that using a single subword. As a result, it reached 81.43 % in MAP.

We could not refine our language models sufficiently in the experiments due to the time limitation. We should improve and optimize our language models using a longer N-gram models because longer and stronger constraints such as syllable triphone showed the better performance when using triphone provided by the organizer. We could not use the query specific method [15] which should also be tested. Because the only even data in CSJ were used for training acoustic and language models due to the time limitation, we should use and integrate the results using the odd data.

5. ACKNOWLEDGMENTS

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.

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

[7] Yuji Onodera et al, “Spoken Term Detection by Result Integration of Plural Subwords using Confidence Measure”, WESPAC, 2009

[8] 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

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

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

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

[12] Takuya.N, et al, “A Consideration of the Number of Triphone for Spoken Term Detection Using Triphone Occurrences”, ASJvol2, pp249-252, 2011-3.

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

[14] 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.

[15] Yoshiaki Itoh, Kohei Iwata, Masaaki Ishigame, Kazuyo Tanaka, Shi-wook Lee, “Spoken Term Detection Results Using Plural Subword Models by Estimating Detection Performance for Each Query,” International Conference on Speech Communication and Technology (INTERSPEECH), pp. 2117 -2120, 2011.