High speed spoken term detection by combination of n-gram array of a syllable lattice and LVCSR result for NTCIR-SpokenDoc

Keisuke Iwami
Toyohashi University of Technology
1-1 Hibarigaoka
Toyohashi-shi
Aichi,440-8580
iwami@slp.cs.tut.ac.jp

Seiichi Nakagawa
Toyohashi University of Technology
1-1 Hibarigaoka
Toyohashi-shi
Aichi,440-8580
nakagawa@slp.cs.tut.ac.jp

ABSTRACT

For spoken document retrieval, it is very important to consider Out-of-Vocabulary (OOV) and mis-recognition of spoken words. Therefore, sub-word unit based recognition and retrieval methods have been proposed. This paper describes a Japanese spoken document retrieval system that is robust for considering OOV words and mis-recognition of subunits. Additionally our system combines Large Vocabulary Continuous Speech Recognizer(LVCSR) and a sub-word unit recognition system for In-Vocabulary(IV) words. We used individual syllables as sub-word unit in continuous speech recognition and an n-gram sequence of syllables in a recognized syllable-based lattice. We propose an n-gram indexing/retrieval method with distance in the syllable lattice for attacking OOV, recognition errors, and high speed retrieval. We applied this method to academic lecture presentation database of 44 hours, and 0.645(F-measure) of the OOV words was obtained in less than 2.5 milliseconds per query.

Keywords

Spoken Term Detection, syllable recognition, n-gram, Out-of-Vocabulary, mis-recognition , NKGW

1. INTRODUCTION

Recently, with the growth of information and communication technology, multimedia data such as audio and video can be found on the Web. Information can be found with an existing textual search engine if the target data consist of textual information such as transcribed broadcast news and newspapers; however, an efficient spoken document retrieval (SDR) or spoken term detection (STD) method is still not established, because spoken documents have specific problems such as recognition errors and out-of-vocabulary(OOV) terms. The aim of this research is to develop a robust and efficient STD method. A standard STD method is using textual search to LVCSR transcripts. However, OOV terms are not registered in a dictionary of speech recognizer. Therefore it is impossible to detect the OOV term with an existing text search engine because the word is not given as an output in the recognition result of an LVCSR. The advantage of using a sub-word unit based speech recognition system is that it can ignore grammatical constraints and recognize any OOV terms.[1].

In German, the retrieval method based on the weighted Levenshtein distance between syllables (words consist of only one syllable in a ratio of half)[2] has been proposed. In Chinese, syllable-unit (440 syllables in total) has often been used as a basic unit of recognition/retrieval[3]. In addition, other retrieval methods based on elastic matching between two syllable sequences have been tried for considering recognition errors[4]. Phoneme based n-gram has been proposed for various retrieval methods, usually with bag of words or partial exact matching[5][6]. For document retrieval, Chen et al[7] used skipped (distant) bigrams such as s_1-s_3, s_2-s_4 for the syllable sequence of $s_1 s_2 s_3 s_4$. Phoneme recognition errors such as substitution errors have not been explicitly considered[8][9] for OOV term retrieval. Akbacak et al [10] used hybrid recognition systems which contained both words and subword unit to generate hybrid lattice indexes.

In Chinese, character-based on syllable-based spoken document/term retrievals performance is comparable with word-based one[11] even for IV. This finding is due to the special property of Chinese language (almost all words consist of one or two syllable). Japanese consists of only about 110 syllables, therefore the syllable unit is suitable for the spoken retrieval of OOV words.

It is necessary to prune the many detection candidates. Typically, as with the dynamic time warping (DTW) method, a string is used to elastically match candidates for pruning. However, DTW processing is more time consuming than index base search processing. Instead of DTW, we used the n-gram array with distance measure that acounts recognition errors in the syllable recognition lattice. We show a significant improvement of processing time using this method. As a result of the retrieval, we could detect about 0.645(F-measure) of OOV words in the database of 44 hours of audio in less than 2.5 milliseconds per query.

2. SYSTEM OVERVIEW

2.1 Spoken term detection procedure for IV and OOV

We propose a Spoken Term Detection (STD) system for OOV words using a syllable lattice with mis-recognized syllables. In this study, we use an n-gram of syllables for the STD, in particular for OOV terms. A flow chart of the proposed method is illustrated in Fig. 1 . The spoken document

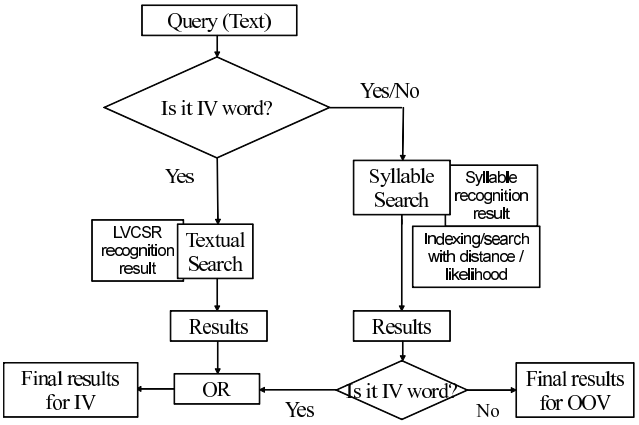

Figure 1: Flow chart of proposed technique

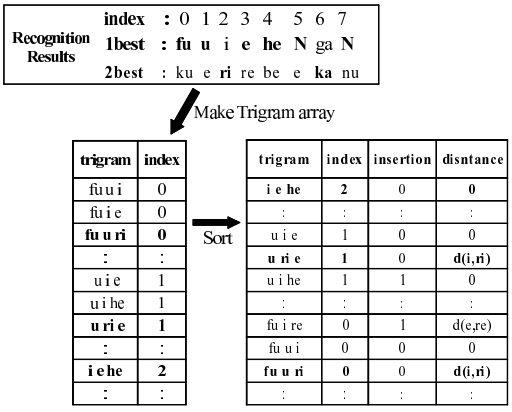

Figure 2: Procedure for making tri-gram array

is recognized by an LVCSR for IV words and by a continuous syllable recognition system for dealing with OOV words and mis-recognized words, and then the indexing is applied to the lattice. A query consisting of IV words is retrieved using a standard text search technique from the LVCSR results (confusion network). A search for OOV terms or mis-recognized words is conducted using the top m-best in the syllable lattice.

To handle mis-recognition errors of the LVCSR, the system also searches spoken terms in IV using the same syllable-based method as for OOV term detection. Finally, we combine the retrieval results of the text search and the syllable sequence search. Recently, various retrieval systems based on system combinations have been proposed and these have reported improved performance[12][13]. Nevertheless, we focus on a retrieval method based on a single system in this paper.

2.2 OOV word retrieval by DTW for subword sequence -baseline-

There are substitution, insertion and deletion errors for a sub-word automatic speech recognizer. The word retriveal has to find candidate positions from the recognized sub-word sequence. We usually call this word spotting. For word spotting, the recurrence equation of DTW (between an input sub-word sequence $A = a_1 a_2 \cdots a_I$ and a query sub-word sequence $B = b_1 b_2 \cdots b_J$) is shown in [14]. We used syllable as a sub-word unit, which consists of a consonant and a vowel, or a single vowel in Japanese. The position where the query appears can be obtained as a retrieval result.

2.3 High-speed OOV word retrieval method by n-gram with distance[14]

We propose a SDR system for OOV words using a syllable lattice with mis-recognized syllables. The spoken document is recognized by an LVCSR for IV words and by a continuous syllable recognition system for OOV words and mis-recognized words in IV words, and then the indexing is applied to the lattice. The lattice consists of plural candidates at every best candidate. In this study, we use an n-gram of syllables for Spoken Term Detection (STD), in particular for OOV terms. N-gram information of syllables is maintained by a data structure called an n-gram array that consists of index and syllable distance information within each n-gram.

Fig. 2 illustrates how a trigram array is arranged. First, the appearance position of syllables in a spoken document is allocated. Then an n-gram of the syllable at every appearance position is constructed. Next, the n-gram is sorted in a lexical order so as to search it quickly using a binary search algorithm. The column of "index" in Fig. 2 represents a position of trigram in spoken document, "insertion" represents newly defined measure, which means the number of insertion errors in an indexed trigram. The column of "distance" is will be described in section 2.4.

The search process on an n-gram array is divided into 3 steps. First, a query is converted into a syllable sequence. Second, an n-gram of the query is constructed. Finally, the query is retrieved from the n-gram array. A query consisting of more than n+1 syllables is retrieved by a combination of n-grams. A query consisting of less than 2n syllables and more than n+1 syllables is separated into two n-grams in the first half and the latter half (refer to Fig.3). Thus, the query is retrieved from the n-gram array twice. The retrieved results are merged by considering whether the position where the detection results occurred in the first half and the latter half appeared within one position or not. Similarly, a query of less than 3n syllables and more than 2n+1 syllables is retrieved by the sequence of syllables by dividing it into three parts (refer to Fig.3). The purpose of considering the gap between the appearance positions is to deal with mis-recognitions. This is described in details in the next section.

2.4 How to solve mis-recognized sub-word problem for OOV detection[14]

2.4.1 Substitution error

To handle substitutions errors, we use an n-gram array constructed from the m-best of the syllable lattice[14]. An n-gram array is constructed by using the combination of syllables in the m-best syllable lattice. Thus, for one position in the lattice, there are m^n kinds of n-gram. For example, even if the recognition result of the 1-best is "fu e ki e he N ka N" having recognition errors, we can search the query "fu u ri e he N ka N("Fourie Transform" in English)", if a correct syllable is included in the m-best. We used HMM based Bhattacharrya distance [14] as the local distance between the 1st candidate and other candidate. Fig. 4 illustrates an

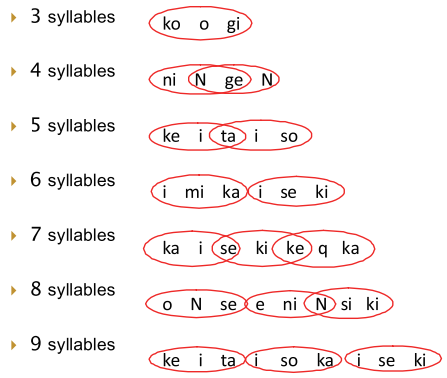

Figure 3: Examples of query division where n=3

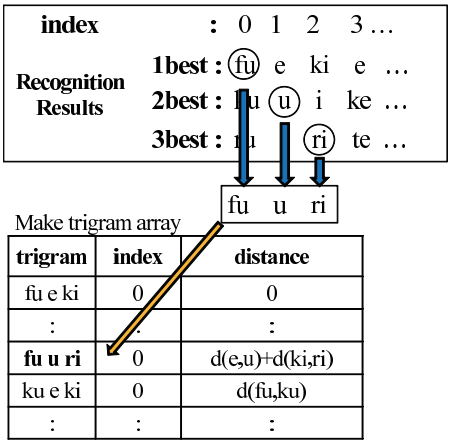

Figure 4: Example of substitution error action where n=m=3

example of substitution error action. The "fu u ri" distance is calculated as distance between "fu u ri" of target trigram and "fu e ki" of the 1-best trigram. Thus, where the distance is $d(e,u) + d(ki,ri)$.

2.4.2 *Insertion error*

To address the insertion errors, we make an n-gram array that permits a one-distant n-gram. Considering the gap between appearance positions deals with the error. Even if the recognition result is "fu ku u ri e he N ka N" having an insertion errors, we can search the query "fu u ri e he N ka N", if the n-gram array that considers a one-distant n-gram is allowed. Therefore, it is possible to deal with one insertion error within every n-gram. Fig.5 illustrates an example of insertion error action. The trigram of "fu u ri" is constructed as a skipped trigram from "fu ku u ri", when "ku" is regarded as an insertion error.

2.4.3 *Deletion error*

To handle the deletion errors, we search the query as above while allowing for the case where one syllable in the query is deleted. Even if the recognition result is "fu u e he N ka N" having a deletion error, we can search the query "fu u ri e he N ka N", if a syllable ('e') in the query is deleted (refer to Fig. 6). The deleted syllable is indicated by the

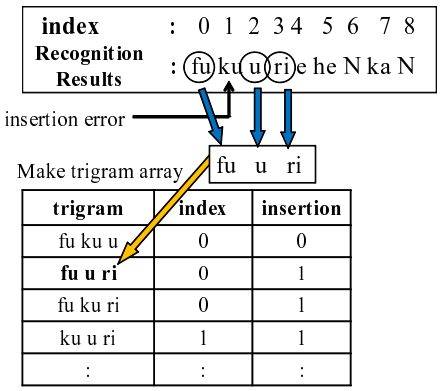

Figure 5: Example of insertion error action where n=m=3

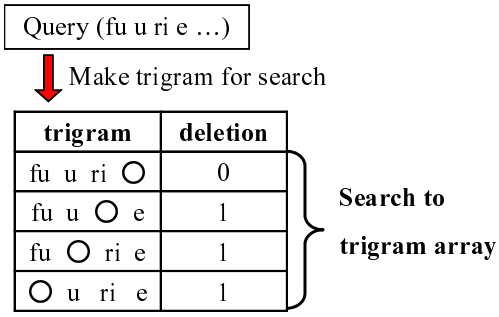

Figure 6: Example of deletion error action where n=m=3

"○" mark as shown in Fig. 6. Moreover, it is possible to deal with a substitution error by combining the action of the insertion error and the deletion error. Therefore, it is possible to tackle a substitution error even if a syllable does not exist in the m-best of the lattice. In Fig. 7, the recognition result "fu u mi e" which contains a substitution error is retrieved when both actions of insertion and deletion errors are applied. At this case, an n-gram array had trigram "fu u e" after considering an insertion error ("mi" was assumed as an insertion error). Thus, we can retrieve "fu u ri e" from n-gram array when the query "fu u ri e" is regarded as a deletion error of "ri".

When a query consisting of syllables more than 2n must consider deletions of two syllables, the errors for a long query can not be corrected simply by deleting one syllable. In such a case, the query is divided two parts, and they are made to drop out by one syllable, and retrieved. For example, for the recognition result of "fu u ri e he N ka N", it is retrieved by considering one deletion of "fu u ri e" and of "he N ka N" in the case of $n = 3$, respectively.

2.5 IV word retrieval by combination of LVCSR and syllable recognition results

We used continuous syllable recognition results for OOV queries. For IV queries, we have combined LVCSR and syllable recognition results. Our proposed method allows to search for different types of queries.

- queries containing only IV terms, where the LVCSR results and the syllable recognition results are used.

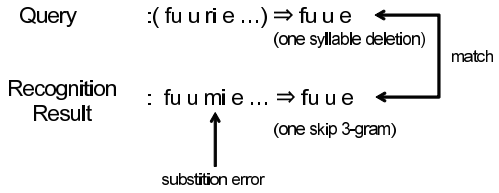

Figure 7: Example of substitution action by combining insertion and deletion error actions

- queries containing only OOV terms, where the syllable recognition results are used.

- queries containing both IV and OOV terms, where the syllable recognition results are used. For example, "名犬ラッシー" is a compound noun consisting of "名犬"(fine dog) in IV words and "ラッシー"(Lassie) in OOV words.

A search for IV term consists of 3 stages. First, a query of IV term is retrieved using an inverted index considering a word confusion network of LVCSR result. We can accurately detect the IV term using LVCSR result, if the word is correctly recognized. Second, for word mis-recognition, our approach retrieves the same IV term by using the n-gram array made from the syllable recognition results. Finally, the system combines the two results using the OR operator. Although false alarms may occur in retrieval results when using n-gram array, it is robust for mis-recognition and improves the recall rate.

2.6 Decision maker

String matching as with DTW consumes computation time linearly with the amount/length of spoken documents. To solve the problem more efficiently, we define a new distance measure to present the number of allowed errors. N-gram arrays using the distance metric will enable fast pruning of unreliable candidates without elastic string matching. The syllable distance in a substitution error is calculated by the Bhattacharyya distance between two HMM-based syllables from the 1best recognition result. In other words, the syllable in the 1best result is always zero distance (refer to the column of "distance" in Fig. 2).

Next, the syllable distance for an insertion error is represented by 1 (refer to the column of "insert" in Fig. 2.). Finally, the syllable distance for deletion errors is equal to the number of syllables that were eliminated in a query (refer to Fig. 6). In this study, the deletion error distance is 0 or 1, because we allow up to one deletion error for each n-gram.

2.7 Memory and search time

The memory required by the proposed method corresponds to the amount of memory used by the index for the trigram array. It is possible to expand to any n-gram array, however, retrieval method using only trigram has hitherto been proposed. The amount of memory in an n-gram array is estimated by Eq. (1).

Table 1: Memory size per trigram index

(a) Memory size of S_1

method	n-gram	number of entries	total
simple	4bytes	4bytes	8bytes
compressed	3bytes	4bytes	7bytes

(b) Memory size of S_2

method	position	insertion	substitution	total
simple	4bytes	4bytes	4bytes	12bytes
compressed	4bytes	1bit	7bits	5bytes

$$
\begin{aligned}
M &= M_1 \times S_1 + M_2 \times S_2 \quad (1)\\
S_1 &= \text{memory size of } \{\text{n-gram type} + \text{number of entries}\}\\
S_2 &= \text{memory size of } \{\text{position}\\
&\qquad +\text{insertion distance}\\
&\qquad +\text{substitution distance}\}
\end{aligned}
$$

Where M_1 is the number of kinds of trigram, M_2 ($>> M_1$) is the number of n-gram entries in the index, S_1 is the amount of memory for a kind of n-gram, and S_2 is the amount of memory for an n-gram entry(refer to Table 1).

Three syllables are included in a trigram. Each syllable consists of one or two characters (or phonemes). It is assumed that the quantity of spoken document generated is six syllables per second considering the 3-best in the lattice. If the above simple expression is used, a trigram array of about 2MB is created for a spoken document file, one hour in duration for the case of 3-best. For the case of 5-best, the size becomes 12MB. Additionally, the index size becomes 4 times when considering the insertion errors, for example, ABC, ABD and ACD and BCD for ABCD. In this simple expression case, the memory size per hour becomes about 40MB for 5-best. If we implement the compressed expression as shown in Table 1, the required memory size will become $17MB$ per hour. Furthermore, we can reduce the memory size to about $1/2$ by using a pruning method[14]. If 4-gram (the 4-gram for the retrieval of a word that consists of the 5-best), the number of indices increases significantly and the amount of memory required becomes enormous. When retrieval is carried out with a 4-gram, only one insertion error per four syllables can be considered. By considering the required amount of memory, speech recognition performance, retrieval performance, search speed, and so on,we can determine whether a 3-gram or 4-gram is better.

Since the proposed method uses a binary search, the computation amount is $O(\log_2 M_1)$ in a single search of a trigram. Then, the algorithm checks sequentially the corresponding trigram entries at the different positions of M_2/M_1 in average and performs the post-processing for the connectivity. Actually, our method divides a query into several parts every 3 syllables, so it searches multiple times. For example, when the query is composed of 7 syllables, our method searches the query in three parts from n-gram array(refer to Fig. 3). If the index has M_1 kinds, the computation of binary search is $k \log_2 M_1$, where k is the number of division of the query. The retrieved results are merged into the final result while considering the connectively of adjacent retrieved results. The merging process consumes much more computation than the search process. Therefore, the

Table 3: Recognition results (%)

output	Del	Ins	Subs	Corr	Acc
Syllable (1best)	3.9	3.6	12.5	83.6	80.0
Syllable (3best)	3.9	2.2	6.9	89.1	86.9
Syllable (5best)	4.1	1.9	4.9	91.0	89.1
Converted syllable from LVCSR	4.1	2.3	12.5	83.3	81.1
WORD from LVCSR (1best)	5.4	4.6	22.7	71.9	67.3

Table 4: Result of IV retrieval (LVCSR+n-gram)

	LVCSR	n-gram(5-best)	LVCSR+n-gram	DTW	LVCSR+DTW
Detect	103	96	147	88	141
Correct	98	84	130	83	131
Recall	0.51	0.44	0.68	0.43	0.69
Precision	0.95	0.86	0.88	0.94	0.93
F-measure	0.67	0.59	0.77	0.59	0.79

Table 5: Result of OOV retrieval (n-gram, 5-best)

OOV	n-gram index								
	no proces	(1)Sub	(2)Ins	(3)Del	(1)+(2)	(1)+(3)	(2)+(3)	(1)+(2)+(3)	DTW
Detect	9	31	9	15	40	54	18	72	79
Correct	9	28	9	15	38	44	18	57	52
Recall	0.05	0.17	0.05	0.09	0.23	0.26	0.11	0.34	0.31
Precision	1.0	0.90	1.0	1.0	0.95	0.81	1.0	0.79	0.66
F-measure	0.10	0.28	0.10	0.16	0.37	0.40	0.19	0.47	0.42

computation amount depends largely on a query length and amount of the retrieved results (i.e. proportional to the duration of speech document).

Additionally, the number of n-gram search is proportion to the number of deletion actions. For example, when the query is composed of 6 syllables, the deletion actions are activated by 6 times. On the other hand, the search from word inverted index based on LVCSR results is much faster than n-gram array, because the word inverted index size is relatively small and we do not need to consider any recognition errors.

3. EVALUATION AND RESULTS

3.1 Experimental setup

For our experimental data, we used the 44 hours of core data in the CSJ (Corpus of Spontaneous Japanese) corpus as experimental data[15] and SPOJUS[16] developed in our laboratory as the LVCSR. The context-dependent syllable-based HMMs were trained on 2707 lectures within the CSJ corpus excluding the core data. Table 2 summarizes the conditions for speech analysis. We used a left-to-right HMM, consisting of four states with self loops, and has four Gaussians with full covariance matrices per state. The language model for LVCSR was trained by using 2707 lectures within the CSJ corpus excluding the core data. The vocabulary size is about 28k, and the cutoff value in the language model was set to 4. Continuous syllable recognition was performed by context-dependent 928 syllable-based HMMs and syllable-based 4 grams as a language model. In Japanese, there are only 116 kinds of syllables.

We used the query set for the formal run in NTCIR9[17]. In our experiments, the query set was divided into IV queries and OOV queries. The syllable recognition rate, word recognition rate from LVCSR and syllable recognition rate converted from recognized words are summarized in Table 3. The best result of continuous syllable recognition is slightly worse than that of the syllable recognition rate converted from the LVCSR result.

We implemented the proposed method on the following machine specification: Xeon 2.93GHz, 24core CPU, and 74GB memory (we used only a single core).

3.2 In-vocabulary term detection

Table 4 shows the results (with the maximum of F-measure) when these queries were retrieved using results of the LVCSR (confusion network) and the column of "LVCSR" represents the performance using the LVCSR results, and "n-gram" represents the performance using the n-gram array from a recognized syllable (lattice) sequence. "DTW" means the performance by syllable (lattice) sequences based on Bhattacharyya distance as a local distance. "LVCSR+DTW" dontes the combination of the LVCSR based approach and DTW based approach. Fig. 8 illustrates comparison results for IV term detection.

When considering the balance between precision and recall, there was no difference for IV retrieval between the proposed method and DTW. In the retrieval case from LVCSR results, we obtained F-measure of 0.67. Additionally, the combination of the LVCSR and n-gram array, we were able to increase Recall from 0.51 to 0.68 and F-measure from 0.67 to 0.77, respectively. We should notice that the recall rate 0.68 was lower than the word correct rate 0.719 in Table 3, because the recognition of query terms in IV was hard and some queries were the compound of multiple words. "LVCSR+DTW" shows the best performance, but the DTW-based approach consumes huge computation.

3.3 Out-of-vocabulary term detection

Table 5 shows the performance by attacking various recognition errors (with the maximum of F-measure) and Fig. 9 illustrates comparison results for OOV term detection. The mark of (1) denotes the substitution error processing, (2) is the insertion error processing, and (3) is the deletion error processing. For the measures of precision and recall, the "no processing" action means the result without error considerations, that is , exact matching using only the first candidate.

As a baseline, we used DTW that the local distance is Bhattacharyya distance. This distance was derived from the 5-best of the syllable lattice[14]. In addition, we compared varius local distances on the DTW; syllable-based edit distance and phoneme-based edit distance (1best, 5best). The best result with the baseline in these experiments was obtained from Bhattacharyya distance with the 5-best[14]. Finally, we obtained F-measure of 0.47 for OOV term detection. Although the value is lower than the case of IV term detection, our proposed method outperformed the DTW based method (F-measure of 0.42).

Fig. 10 illustrates that our proposed method was comparable with the conventional DTW in all queries. The result of "n-gram+LVCSR" was submitted to the formal run task.

3.4 Memory size and retrieval time

The size of the index used in these experiments is about 3.5GB. However, we can reduce the size of the index to 600MB by a pruning method for the lattice without loss

Table 2: Conditions for acoustic analysis of input speech

Sampling Rate	16kHz
Preemphasis	$1 - 0.98z^{-1}$
Analysis Window	Hamming Window
Analysis Frame Length	25 ms
Analysis Frame Shift	10 ms
Feature Parameters	MFCC + △MFCC + △△MFCC + △Pow + △△Pow (38 dimensions)

of performance.

Fig. 11 shows the comparison of the two methods on the average search time per query at every number of syllables. The average search time was 500msec using DTW, whereas that the trigram array method integrated with distance was 2.5msec for each query. DTW has a drawback that the processing time is propotinal to the amount of spoken document. On the other hand, even if the number of spoken documents increases, query could be searched rapidly using a trigram array with the proposed trigram method. This difference becomes significant with the increase of spoken documents.

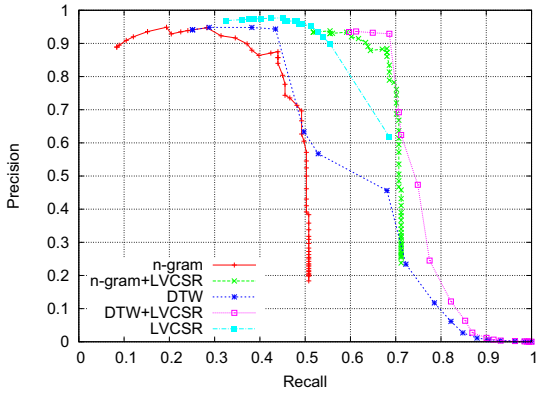

Figure 8: Retrieval results for IV (formal run: n-gram+LVCSR)

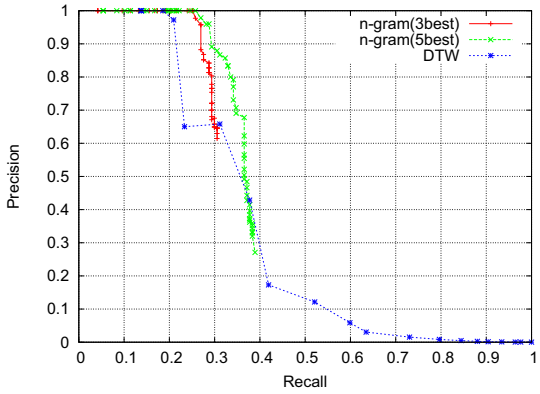

Figure 9: Retrieval results for OOV (formal run :n-gram(5-best))

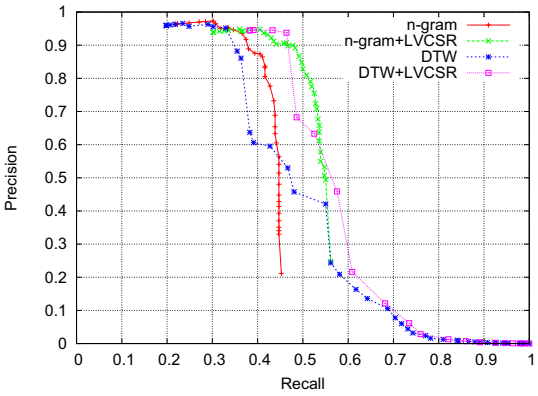

Figure 10: Retrieval results for IV+OOV (formal run: n-gram+LVCSR)

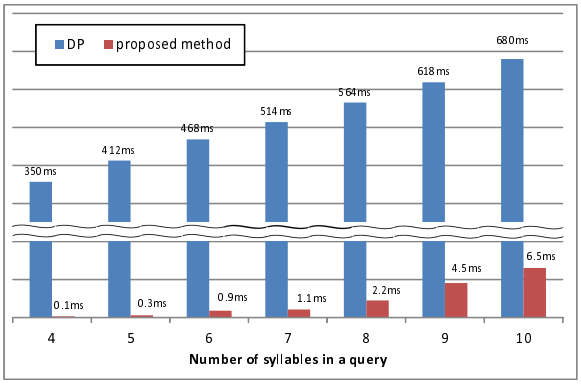

Figure 11: Comparison of retrieval time

4. CONCLUSION

In this paper, we could retrieve OOV and IV terms for spoken documents at 0.645 of F-measure. We have presented the search method that combines the syllable based n-gram array and word based inverted index. The system can deal with all kinds of queries such as containing substitution error, insertion error and deletion error.

The advantage of our proposed method is to be able to retrieve IV and OOV terms with high precision. On the other hand, the disadvantage is not to be able to retrieve them with high recall rate. This is caused by the hard decision for the n-gram entry registration. This relaxation is one of future work.

An other important topic for future work is to improve the retrieval performance. One way to improve the retrieval accuracy is to use only low confidence parts as OOV candidates from the results of the LVCSR.[18]. Another way

is to improve the syllable recognition rate by combining the results of several decoders[18]. Finally, we may use the syllable's likelihood obtained from the decoder, instead of the syllable distance to improve the retrieval accuracy.

5. REFERENCES

[1] K.Ng, "Towards robust methods for speech document retrieval," ICSLP, 1998, pp. 1088–1091.

[2] M.Larson and S.Eickeler, "Using syllable-based indexing features and language models to improve German spoken document retrieval," EuroSpeech, 2003, pp. 1217 – 1220.

[3] H.Wang, "Experiments in syllable-based retrieval of broadcast news speech in Mandarin Chinese," Speech Communication, 2000, vol. 32, pp. 49 – 60.

[4] M.Wechsler, E.Munteanu, and P.Schauble, "New techniques for open-vocabulary spoken document retrieval," SIGIR, 2008, pp. 20 –27.

[5] C.Allauzen, M.Mohri, and Saracla M, "General indexation of weighted automata - application to spoken utterance retrieval," Workshop on interdisciplinary approaches to speech indexing and retrieval, 2004, pp. 33–40.

[6] M.Saraclar and R.Sproat, "Lattice-based search for spoken utterance retrieval," HLT/NAACL, 2004, pp. 129–136.

[7] B.Chen, H.Wang, and L.Lee, "Retrieval of broadcast news speech in Mandarin Chinese collected in Taiwan using syllable-level statistical characteristics," ICASSP, 2000, pp. 2985–2988.

[8] C.Ng, R.Wilkinson, and J.Zobel, "Experiments in spoken document retrieval using phoneme n-grams," Speech Communication, 2000, vol. 32, pp. 61 – 77.

[9] S.Dharanipragada and S.Roukos, "A multistage algorithm for spotting new words in speech," IEEE Transactions on Speech and Audio Processing, 2002, vol. 10, pp. 542 – 550.

[10] M.Akbacak, D.Vergyri, and A.Stolcke, "Open-vocabulary spoken term detection using graphone-based hybrid recognition sytems," ICASSP, 2008, pp. 5240–5243.

[11] H.M.Meng, W.K.Lo, Y.C.Li, and P.C.Ching, "Multi-scale audio indexing for Chinese spoken document retrieval," ICSLP, 2000, vol. 4, pp. 101–104.

[12] S.Natori, H.Nishizaki, and Y.Sekiguchi, "Japanese spoken term detection using syllable transition network derived from multiple speech recognizers outputs," INTERSPEECH, 2010, pp. 681–684.

[13] C.Parada, A.Sethy, M.Dredze, and F.Jelinek, "A spoken term detection frame work for recovering out-of-vocabulary words using the web," INTERSPEECH, 2010, pp. 1269–1272.

[14] K.Iwami, Y.Fujii, K.Yamamoto, and S.Nakagawa, "Out-of-vocabulary term detection by n-gram array with distance from continuous syllable recognition results," SLT, 2010, pp. 200–205.

[15] Y.Itoh, H.Nisizaki, and et.al., "Constructing Japanese test collections for spoken term detection," INTERSPEECH, 2010, pp. 677–680.

[16] Y.Fujii, K.Yamamoto, and S.Nakagawa, "Large vocabulary speech recognition system: Spojus++," MUSP, 2011, pp. 110 – 118.

[17] T.Akiba, H.Nishizaki, K.Aikawa, T.Kawahara, and T.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.

[18] H.Nishizaki and S.Nakagawa, "Japanese spoken document retrieval considering OOV keywords using OOV detection processing and word spotting," HLT, 2002, pp. 144–151.