COMBINATION OF SYLLABLE BASED N-GRAM SEARCH AND WORD SEARCH FOR SPOKEN TERM DETECTION THROUGH SPOKEN QUERIES AND IV/OOV CLASSIFICATION

Nagisa Sakamoto[1], Kazumasa Yamamoto[1] and Seiichi Nakagawa[1]

[1]Toyohashi University of Technology, Japan

sakamoto@slp.cs.tut.ac.jp, kyama@slp.cs.tut.ac.jp, nakagawa@slp.cs.tut.ac.jp

ABSTRACT

This paper presents a Japanese spoken term detection method for spoken queries using a combination of word-based search and syllable-based N-gram search with in-vocabulary/out-of-vocabulary (IV/OOV) term classification. The N-gram index in a recognized syllable-based lattice for OOV terms, which assumes recognition errors such as substitution, insertion and deletion errors, incorporates a distance metric as a confidence score. To address spoken queries, we propose an automatic method for discriminating IV and OOV terms by using the confidence scores of spoken queries through large-vocabulary/syllable continuous speech recognition. Evaluation on an academic lecture presentation database with 44 hours of data shows that the combination of word search and syllable-based N-gram search yields significant improvement and outperforms the baseline syllable-based DTW approach.

Index Terms— spoken term detection, spoken queries, syllable-based N-gram, IV/OOV classification

1. INTRODUCTION

In recent years, web content has been enhanced as a result of more Internet users and higher communication speeds. Of the available content, multimedia such as audio clips and images are particulary abundant on the Web, and an effective technique for utilizing large amounts of data is critical. Spoken term detection (STD) represents the task of identifying the utterance location of a particular keyword represented by a word or a lattice of words in audio documents; consequently, the study thereof has recently attracted much attention in the field of information retrieval. There are two categories for STD; language independent STD and language dependent STD [1]. For language independent STD, it is not possible to train target language-specific acoustic models, therefore, the system needs to rely on a language-independent acoustic modeling approach or universal acoustic models. The Spoken Web Search task of MediaEval belongs to this category [2].

However, typical language dependent STD systems use large-vocabulary continuous speech recognition (LVCSR) to transform speech data into text, which is maintained in a data structure called an index (i.e., indexing), and then use to perform matching between the search term and text data (i.e., searching). Thus, it is necessary to consider the problem of speech recognition errors of an LVCSR system and out-of-vocabulary (OOV) terms that are not present in the speech recognition dictionary. To handle OOV terms, a sub-word unit (e.g., phones, graphones, syllables, or morphs) based index is usually constructed. The SpokenQuery&Doc task at NTCIR belongs to this category [3]. In this paper, we focus on this.

In German, a retrieval method based on the weighted Levenshtein distance between syllables has been proposed [4]. In Chinese, a syllable-unit (440 syllables in total) is often used as the basic unit of recognition/retrieval [5]. In addition, other retrieval methods based on elastic matching between two syllable sequences have been used to consider recognition errors [6]. The phoneme-based N-gram has been proposed for various retrieval methods, usually with a bag of words or partial exact matching [7, 8]. For document retrieval, Chen et al. [9] used skipped (distant) bigrams such as s_1-s_3, s_2-s_4 for the syllable sequence of $s_1 s_2 s_3 s_4$. Xu et al. [10] proposed partial phoneme sequence matching and they showed that a phoneme sequence of length with six phonemes proved optimal.

Typically, as with the dynamic time warping (DTW) method, a string is used to match candidates elastically for pruning. Katsurada et al. proposed a fast DTW matching method based on a suffix array [11]. Kanda et al. [12] proposed a hierarchical DTW matching method between phoneme sequences, where the coarse matching process is followed by fine matching. However, their method still requires a great deal of computation time and memory storage. Kon' no et al. [13] researched a coarse/fast retrieval method based on various subword units that match results calculated in advance, and then followed by fine DTW matching. This method requires significant computation in advance.

In our previous work [14, 15, 16], we proposed a robust STD method using a syllable-based N-gram with distance. The N-gram assumes three types of recognition errors (substitution, insertion, and omission errors) for spoken documents, making it possible to detect search terms including OOV words and recognition errors. Distances of the N-grams are provided as confidence scores to determine whether the search terms are matched by comparing each distance with a threshold value.

For spoken queries, many works have investigated direct matching with acoustic features obtained from documents and queries, which is referred to as query-by-example STD [17, 18]. Researchers focused on features that are frame-level phone posterior probabilities [19], or the hidden Markov model (HMM) pattern configuration [20]. Because annotated speech data are not needed in these approaches, they are beneficial if the language of the data is unknown [21]. We focus on Japanese speech data with labels to train automatic speech recognition (ASR) systems, making it possible to use most of these for creating acoustic and language models. Makino et al. [22] presented a matching method using two-pass DTW for spoken queries. First they performed sub-word level matching and then more accurate matching at the state level.

In this paper, we propose methods for combining word-based results from the LVCSR system and syllable-based results for in-

vocabulary (IV) words, and for automatically determining IV and OOV words for spoken queries. When handling text queries, we are able to determine IV or OOV using the ASR dictionary. In the case of spoken queries, however, we must consider the method of classification owing to mis-recognition. We can identify spoken queries by providing a threshold for confidence scores of spoken queries through LVCSR/syllable recognition or by using a classifier[23].

In our experiments, our best IV/OOV classification accuracy was 0.938, while the retrieval results for automatically identified IV words outperformed a method with perfect classification. However, retrieval performance for automatically identified OOV words did not decrease nearly as much compared with a perfectly correct classification. The syllable-based N-gram search exceeded the baseline method (DTW) in terms of retrieval performance.

The remainder of this paper is organized as follows: In Section 2, we describe our retrieval system. Evaluation results and our conclusion are presented in Sections 3 and 4, respectively.

2. PROPOSED METHOD

In this section, we describe a method for spoken term detection that handles IV terms, OOV terms, and mis-recognition. We obtain a word sequence from an LVCSR system and N-grams from syllable-based lattices through a syllable-based ASR system that can detect IV and OOV words. To address mis-recognition errors, we construct syllable-based N-grams with consideration for the recognition error. A query is represented by a sequence of words/syllables and a spoken query is recognized as a word sequence or a syllable sequence using ASR. Details of our method are explained below.

2.1. System overview

A flowchart of the search process is illustrated in Fig. 1; this is similar to the retrieval system using text-based queries [24]. First we transform spoken queries into text queries through an ASR system. Spoken documents and spoken queries are recognized by an LVCSR for IV words and a continuous syllable recognition system for dealing with OOV and mis-recognized words. Fig. 2 illustrates the representation of our syllable lattice, in which the syllable boundaries correspond to the boundary of the best syllable sequence. Finally, indexing is applied to the lattice. Searching for OOV terms or mis-recognized words using N-grams in the syllable lattice is explained below.

A query consisting of IV words is retrieved using a standard text search of the LVCSR results. To handle mis-recognition errors of LVCSR, the system searches spoken terms in IV using the same syllable-based method as the OOV term detection and combines the results. The "OR" and "AND" operations in Fig. 1 increase the *recall* and *precision* rates, respectively. However, due to the mis-recognition of spoken queries, it may be difficult to correctly classify the IV terms and OOV terms.

In the indexing process, N-gram information of syllables, which consists of index and syllable distance information (or substitution/insertion error penalty) for each N-gram, is maintained in a data structure called the N-gram array. Fig. 3 illustrates how a trigram array is arranged. First, the appearance positions of the syllables in a recognized syllable lattice for a spoken document are located. Then, an N-gram of the syllable is constructed for each appearance position. Next, the N-gram is sorted in lexical order so that it can be searched quickly using a binary search algorithm. In earlier studies, we used only a trigram array [14, 15]. Later, we extended the method to use trigram, bigram, and unigram arrays [24].

We also construct syllable-based N-grams of the query, retrieved from the N-gram array. A query consisting of more than four syllables is retrieved using a combination of N-grams; for example, one consisting of more than four, but fewer than six syllables, is separated into a trigram and bigram or unigram for the first and second halves, respectively as shown in Fig. 4. The common method appears to use overlapping trigrams to provide some sequencing constraints such as N-gram index for documents. However, we do not allow the overlap to reduce search processing time. Thus, the query is retrieved from the trigram array and bigram or unigram array. The retrieved results are merged by considering whether the positions at which the detection occurred in the first and second halves are the same.

The query term is detected if the following distance is lower than a pre-determined threshold. Strictly speaking, the threshold depends on the query length.

$$\frac{\alpha \times \sum d_S + \beta \times \sum d_I + \gamma \times \sum d_D}{number\ of\ syllables} \tag{1}$$

where d_S, d_I, and d_D denote the distances for substitution, insertion, and deletion errors, respectively.

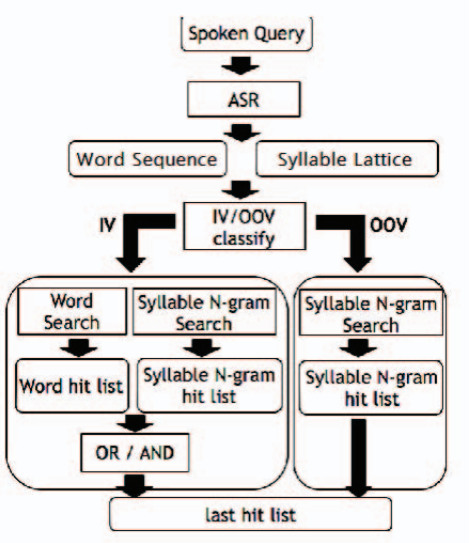


Fig. 1. Flowchart of proposed method.

2.2. Substitution error

To handle substitution errors, we use an N-gram array constructed from the m-best in the syllable lattice [14]. The N-gram array is constructed using a combination of syllables in the m-best syllable lattice. Thus, for a single position in the lattice, there are m^n N-grams. For example, even if the recognition result of the 1-best is "fu u i e he N ga N" with recognition errors, we can search for the query "fu u ri e he N ka N ("Fourier Transform" in English)" if a correct syllable is included in the m-best. We used the HMM-based Bhattacharyya distance [15] as the local distance between the first and the other candidates. The "fu u ri" distance is calculated as the distance between "fu u ri" in the target trigram and "fu u i" in the 1-best trigram, where the distance is $d_s(ri, i)$.

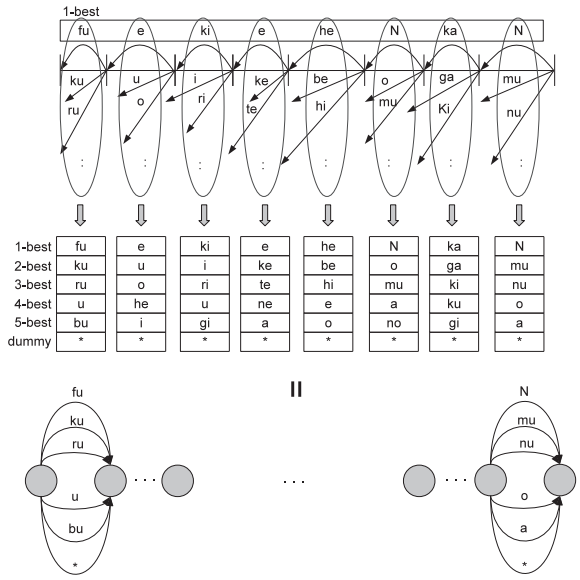
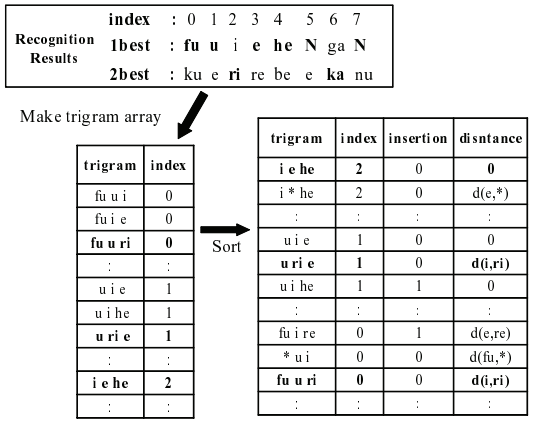


Fig. 2. Construction example of syllable lattice (5-best + dummy). The "lattice" used in this paper corresponds to a sausage-type confusion network.



trigram	index
fu u i	0
fu i e	0
fu u ri	**0**
:	:
u i e	1
u i he	1
u ri e	**1**
:	:
i e he	**2**
:	:

trigram	index	insertion	disntance
i e he	**2**	0	**0**
i * he	2	0	d(e,*)
:	:	:	:
u i e	1	0	0
u ri e	**1**	0	**d(i,ri)**
u i he	1	1	0
:	:	:	:
fu i re	0	1	d(e,re)
* u i	0	0	d(fu,*)
fu u ri	**0**	0	**d(i,ri)**
:	:	:	:

Fig. 3. Procedure for creating a trigram array.

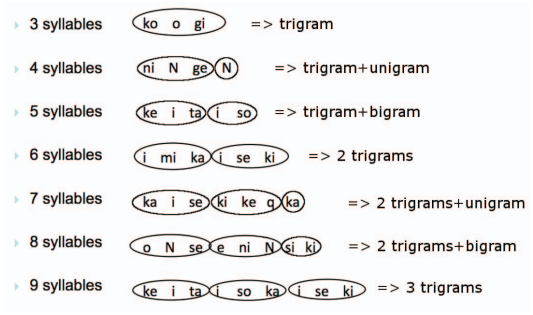


Fig. 4. Example of query division into trigram/bigram/unigram.

Even when using syllable lattices, some substitution errors are not present in the lattice. Therefore, we introduce a dummy syllable symbol or "wild card". The dummy syllable, represented by "*", matches any syllable not present in the m-best recognition results. For example, if the recognition result of the m-best does not include "C", the original method is unable to search the query "ABCD". In this case, the query using the dummy syllable has n-gram as AB*, A*C and *BC, and we can retrieve the query "ABCD". Therefore, the recall rate is increased. However, this method has the potential to decrease the precision rate. We can adress this problem by increasing the distance between "*" and any other syllable, where only one dummy syllable is allowed in a trigram. It should be noted that this approach is different from a one distant bigram index method. We used the exact definition of $d_S(syllable\ of\ query, *)$ as follows:

$$d_S(syllableofquery, *) = \lambda \times d_S(syllable\ of\ query, best\ syllable\ for\ the\ dummy\ syllable) + \eta \quad (2)$$

,where λ and η denotes an penalty for using the dummy syllable. For example, if "query" is "i me he" and is recognized as "i e he" in Fig. 3, the distance between "me" in the query and "*" in the lattice is defined as $\lambda \times d_s(me, e) + \eta$.

2.3. Insertion/deletion errors

To handle insertion errors, we create an N-gram array that permits a one-distant N-gram [14]. By considering the gap between the appearance locations, we can deal with the error. Even if the recognition result is "fu ku u ri e he N ka N" with an insertion error "ku", we can search for the query "fu u ri e he N ka N" if an N-gram array considering a one-distant N-gram is allowed. Therefore, it is possible to deal with one insertion error within each N-gram. The trigram for "fu u ri" is constructed as a skipped trigram from "fu ku u ri", if "ku" is regarded as an insertion error. The insertion distance is defined as follows[24]:

$$d_I(C_2V_2|C_1V_1\_C_3V_3) = \min \left\{ \begin{array}{l} d_S(C_1V_1, C_2V_2) \\ d_S(V_1, C_2V_2) \\ d_S(C_2V_2, C_3V_3) \end{array} \right\} + \delta_I \quad (3)$$

where C_2V_2(C=consonant, V=vowel) denotes the insertion syllable, and C_1V_1 and C_3V_3 denote the left context and right context, respectively. "δ_I" denotes an insertion penalty. "$d_S(V_1, C_2V_2)$" means that "a part of vowel V_1" is mis-separated into the vowel and an inserted syllable.

To handle deletion errors, we search for the query as explained above while allowing for the possibility of one syllable in the query. A detailed description is given in [24].

2.4. ASR of spoken query and IV/OOV classification

If a spoken query is received by the system after ASR processing, we can treat the query as a text query by considering a word sequence or a syllable-lattice generated from ASR systems. Although the lattice for a spoken query may include insertion or deletion errors, we can attend to them by the method described in the previous sections. However, this framework is strongly dependent on the performance of ASR, consequently, the performance of STD with a spoken query is significantly lower than one with a text query[16]. In the previous research[25], we proposed a combination of candidates obtained

from multiple utterances or through different ASR systems to improve the recall score.

For a spoken query, we cannot know whether it belongs to IV vocabulary or OOV vocabulary unlike a text query. Therefore, we should confirm it, for example, by using acoustic features, linguistic features or a matching distance between the syllable sequence of a recognized word by LVCSR and the syllable sequence by continuous syllable ASR. In this paper, the query is classified using the following scores obtained during recognition of the query:

(1) *AM*: likelihood of acoustic model obtained through LVCSR, normalized by the number of frames.

(2) *LM*: prior probability of language model obtained through LVCSR.

(3) *word posterior by lattice*: word posterior likelihood calculated using the sum of scores of the top 500 candidates through LVCSR, that is, $logP(w_i|location_k) - log\{\Sigma_j P(w_j|location_k)\}$ for w_i and detected location k.

(4) *word posterior by syllable*: word posterior likelihood calculated by taking the difference of the scores from the likelihood of a recognized word by LVCSR and that by continuous syllable ASR, that is, $logP(w_i|location_k) - logP(s_{i_1}, s_{i_2}, ...s_{i_n}|location_k)$ for w_i, syllable sequence $s_{i_1}, s_{i_2}, ...s_{i_n}$ and detected location k).

(5) *DTW distance*: calculated by matching the transformed syllable sequence from a recognized word by LVCSR and the recognized syllable sequence by continuous syllable ASR.

By providing a threshold for the above score, we can discriminate IV and OOV terms. Furthermore, other features including the number of words in the LVCSR results, the number of syllables in the syllable recognition results, and the number of acoustic feature frames are combined with the above scores and used to determine whether a query contains IV or OOV words using a classifier. The experiment was conducted using a support vector machine (SVM) with a radial basis function kernel as the classifier.

Table 1. Syllable and word recognition results.

(a) Spoken documents

Output	Measure	Del	Ins	Subs	Corr	Acc
Syllable (1-best)	SRR	3.9	3.6	12.5	83.6	80.0
Syllable (3-best)	SRR	3.9	2.2	6.9	89.1	86.9
Syllable (5-best)	SRR	4.1	1.9	4.9	91.0	89.1
Word (1-best)	WRR	5.4	4.6	22.7	71.9	67.3

(b) IV queries

Output	Measure	Del	Ins	Subs	Corr	Acc
Syllable (1-best)	SRR	2.8	3.9	18.0	79.2	75.4
Word (1-best)	WRR	2.3	6.9	22.9	74.9	67.9

(c) OOV queries

Output	Measure	Del	Ins	Subs	Corr	Acc
Syllable (1-best)	SRR	2.7	7.1	20.7	76.7	69.6

3. EVALUATION AND RESULTS

3.1. Experimental setup

We used the 44 hours of core data in the CSJ (corpus of spontaneous Japanese) as experimental data [26] and SPOJUS++ [27], developed in our laboratory, as the LVCSR. For recognition of spoken documents, context-dependent syllable-based HMMs (928 HMMs in total) were trained on 2707 lectures within the CSJ corpus excluding the core data. We used a left-to-right HMM, consisting of four states with self-loops, and four Gaussians with full covariance matrices per state. We used an IV term set of 60 queries (818 occurrences) and an OOV term set of 40 queries (185 occurrences) in the core data for LVCSR. For the recognition of spoken queries, we used an acoustic model, which is a context-dependent syllable-based GMM-HMM trained on the ASJ corpus. Both spoken documents and spoken queries were recognized using a syllable-based 4-gram language model and word-based 3-gram language model with a vocabulary size of 20000. The number of speakers for speech input queries was 6 adult males.

The syllable recognition rates (SRR) and word recognition rates (WRR) are summarized in Table 1. Table 1(b) and (c) shows the IV and OOV query ASR results, respectively. Our baseline retrieval system is a DTW method that computes the distance between the 1-best syllable sequence of a spoken query and the 5-best syllable sequences of spoken documents [24]. We compared our system and the baseline system using F-measure (max) and mean average precision (MAP) as measures of search performance.

Table 2. IV/OOV classification accuracy of spoken queries.

Score	IV	OOV	All
AM	0.967	0.083	0.613
LM	0.847	0.358	0.652
word posterior by lattice	0.722	0.833	0.767
word posterior by syllable	0.944	0.892	0.923
DTW distance	1.000	0.000	0.600
Combination (SVM)	0.917	0.971	0.938

3.2. IV/OOV classification

We performed IV/OOV binary classification on the spoken queries using the confidence scores described in Section 2.4. Results of the IV/OOV classification using individual scores and the combined score based on the SVM are given in Table 2, with the SVM performance evaluated by six-fold cross-validation (speaker independent).

As shown, "word posterior by syllable" yielded the highest discrimination rate of the single feature with a classification performance of 92.3%. Fig. 5 illustrates the histogram of IV/OOV occurrences and IV/OOV classification accuracy by "word posterior by syllable" in accordance with a normalized score/threshold. Furthermore, the classification performance by a combination of features with the SVM increased to 93.8%.

3.3. Retrieval results

(a). Text Input Queries

For text input queries, retrieval results by a combination of word-based search and syllable based N-gram search are shown in Fig. 6 for text queries. "Text query" corresponds to Word Error Rate (WER) = 0.0% (ASR Accuracy = 100%) and IV/OOV correct classification rate = 100% for "spoken query". As we can see in these results, however, the only syllable N-gram system provides low performance of F-measure and MAP because recall is very low. Therefore, in the case of IV queries where we can classify IV/OOV correctly for text queries, we combined n-gram search and word search, and obtained an improvement in total overall performance from 0.455 of F-measure to 0.692. The proposed N-gram search method outperformed the baseline DTW method.

Table 3. Retrieval results of spoken queries.

(a) IV queries

STD method	Classification method	F-value	MAP
Syllable DTW	-	0.418	0.514
Syllable N-gram	-	0.433	0.487
Word	oracle	0.486	0.360
Word	posterior by syllable	0.488	0.345
Word	SVM	0.490	0.350
Syllable DTW OR Word	oracle	0.526	0.610
Syllable DTW OR Word	posterior by syllable	0.524	0.608
Syllable DTW OR Word	SVM	0.526	0.610
Syllable N-gram OR Word	oracle	0.525	0.589
Syllable N-gram OR Word	posterior by syllable	0.525	0.588
Syllable N-gram OR Word	SVM	0.530	0.587
Syllable DTW AND Word	oracle	0.486	0.382
Syllable DTW AND Word	posterior by syllable	0.493	0.399
Syllable DTW AND Word	SVM	0.486	0.381
Syllable N-gram AND Word	oracle	0.472	0.348
Syllable N-gram AND Word	posterior by syllable	0.479	0.366
Syllable N-gram AND Word	SVM	0.413	0.380

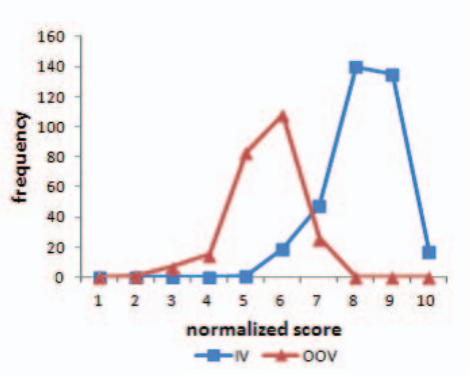
(b) OOV queries

STD method	Classification method	F-value	MAP
Syllable-DTW	-	0.350	0.498
Syllable N-gram	-	0.380	0.441
Word	oracle	-	-
Word	posterior by syllable	0.009	0.006
Word	SVM	0.009	0.006
Syllable DTW OR Word	oracle	0.350	0.498
Syllable DTW OR Word	posterior by syllable	0.315	0.498
Syllable DTW OR Word	SVM	0.349	0.500
Syllable N-gram OR Word	oracle	0.380	0.441
Syllable N-gram OR Word	posterior by syllable	0.338	0.442
Syllable N-gram OR Word	SVM	0.375	0.444
Syllable DTW AND Word	oracle	0.350	0.498
Syllable DTW AND Word	posterior by syllable	0.317	0.464
Syllable DTW AND Word	SVM	0.346	0.488
Syllable N-gram AND Word	oracle	0.380	0.441
Syllable N-gram AND Word	posterior by syllable	0.365	0.404
Syllable N-gram AND Word	SVM	0.367	0.429

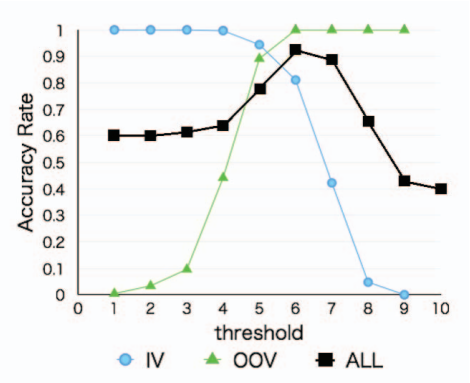
(b). Speech input query

Retrieval results obtained using a combination of word search and syllable search are given in Table 3. Fig. 7 shows the recall-precision (R-P) curve for spoken queries, with IV/OOV classification carried out using an SVM. Taking the "OR" operator in combination with the syllable N-gram search and word search yielded the highest recall rates around the precision rate of 0.1∼0.7 or the highest precision rate around the recall rate of 0.4∼0.7, while the "AND" operator achieved the highest precision rates around the recall rate of 0.1∼0.3 for IV queries as shown in Fig. 7.

We compared the performance of the baseline (Syllable DTW), syllable-based N-gram search (Syllable N-gram), word search (Word), and the combination of syllable-based N-gram search and word search with "OR" or "AND". We also show retrieval performance in the case of "oracle"; that is, we assumed that IV/OOV classification was perfect, with cases of automatic classification using "word posterior by syllable" and "SVM". The F-measure of the syllable N-gram search was better than that in the search results of the baseline DTW method. The combination of syllable-based N-gram search and word search with "OR" (Syllable N-gram OR Word) improved the retrieval performance compared with the single search methods (Syllable N-gram, Word) for IV terms, and outperformed the baseline syllable-based DTW in terms of F-value. However, for OOV terms, only the syllable-based N-gram search method showed improvement. As there is negligible loss in performance, we show that the classification of IV and OOV words does not adversely affect the retrieval performance for OOVs. Automatic IV/OOV classification using an SVM achieved the highest



(a) Histogram of IV/OOV occurances



(b) IV/OOV classification accuracy

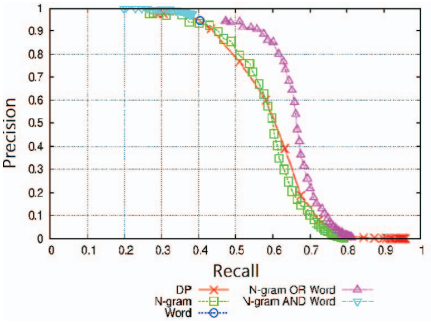
Fig. 5. IV/OOV classification results by "word posterior by syllable" score

performance with respect to F-measure. Our proposed method outperformed the "oracle" case, because difficult mis-recognized IV queries may be regarded as OOVs.

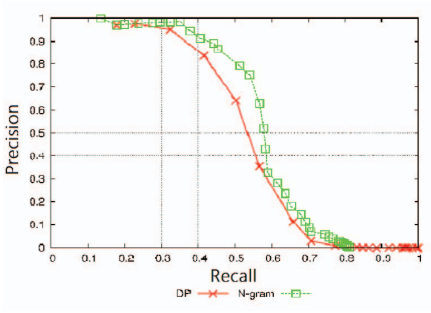
Fig. 8 illustrates the relationship between IV/OOV discrimination performance and retrieval results based on syllable N-gram OR Word. The horizontal axis denotes the IV recall rate by the IV/OOV classification method, where "0.0" in the horizontal axis corresponds to the case that all query terms are classified into OOV category. The rate is controlled by changing the classification threshold for "word posterior by syllable". The left and right vertical axes denote the OOV recall rate / IV WER (word error rate) and F-measure, respectively. The "0" for IV recall rate means "no processing" of IV/OOV classification; that is, all spoken queries are regarded as OOV. Conversely, "1.0" for IV recall rate means that all spoken queries are regarded as IV. "Syllable N-gram OR Word" in Table 3 and Fig. 7 corresponds to the case with "0.944" IV recall rate in Fig. 8 (F-measure of "ALL" queries = 0.476). As we can see, the average retrieval performance for "ALL" queries is the highest for the case with a "0.811" IV recall rate (F-measure of "ALL" queries = 0.492). This shows that only the reliable results of IV/OOV classification should be regarded as IV and all other cases should be regarded as OOV.

3.4. Retrieval time

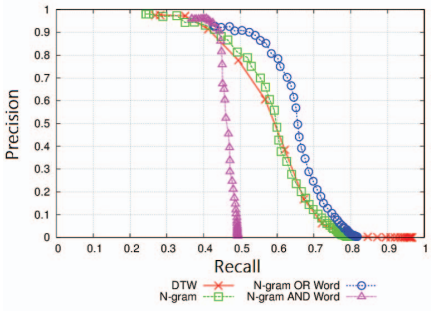
We experimentally compared the average search time per query of the DTW and 3-gram methods for the documents consisting of 44 hours. The average search time using the DTW method was about



(a) IV queries



(b) OOV queries



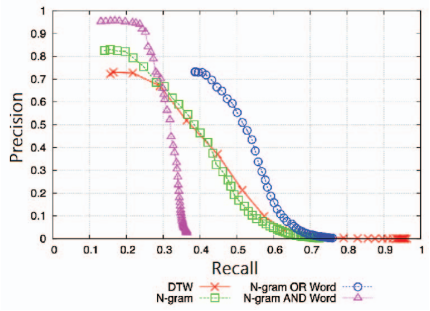
(c) ALL queries

Fig. 6. Recall-Precision Curve for text queries

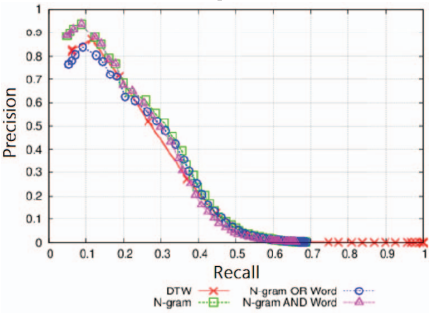
650 ms. On the other hand, the average search time using 3-gram method was about 60 ms.

4. CONCLUSION

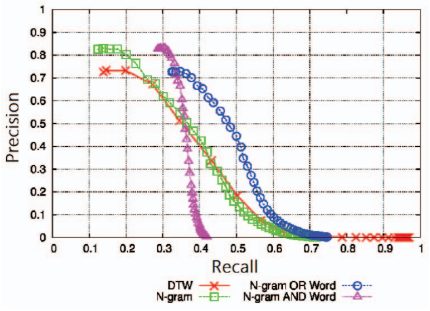
In this paper, we proposed a method for automatic determination of whether a spoken query is IV or OOV using the confidence score obtained through LVCSR and continuous syllable recognition. We achieved a discrimination accuracy of 93.8%. Compared with the case where it is perfectly determined, an improvement of up to 0.05% was observed in the F-value of the retrieval performance with respect to IV words, making it possible to search for OOV words without any loss in performance. Finally, the combination of the syllable-based N-gram search and word search with "OR" operator exceeded the baseline syllable-based DTW method in terms of retrieval performance. As future work, to improve retrieval accuracy, we would like to consider other products such as investigating the re-scoring or normalization method using contextual information and an integration method for multiple ASR scores.



(a) IV queries



(b) OOV queries



(c) ALL queries

Fig. 7. Recall-Precision curve for spoken queries.

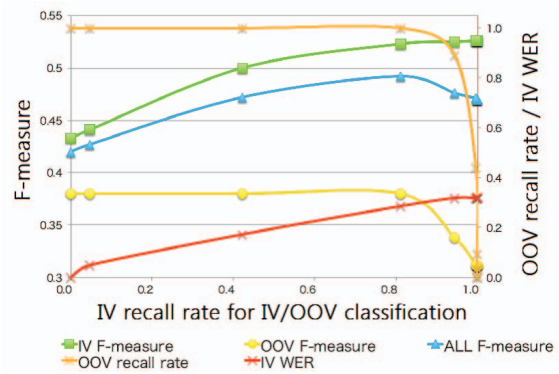


Fig. 8. Relationship between IV/OOV classification performance and retrieval performance.

5. REFERENCES

[1] Javier Tejedor, Michal Fapšo, Igor Szöke, Jan "Honza" Černocký, and František Grézl, "Comparison of methods for language-dependent and language-independent query-by-example spoken term detection," in *ACM Trans. Inf. Syst.*, 2012, vol. 30, pp. 18:1–18:34.

[2] Florian Metze, Xavier Anguera, Etienne Barnard, Marelie Davel, and Guillaume Gravier, "Language independent search in mediaeval's spoken web search task," in *Computer Speech & Language*, 2014, vol. 28, pp. 1066 – 1082.

[3] T. Akiba, H. Nishizaki, H. Nanjo, and G. J. F. Jones, "Overview of the ntcir-11 spokenquery&doc task," in *proceedings of NTCIR-11*, 2014, pp. 350–364.

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

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

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

[7] 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.

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

[9] 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.

[10] Haihua Xu, Peng Yang, Xiong Xiao, Lei Xie, Cheung-Chi Leung, Hongjie Chen, Jia Yu, Hang Lv, Lei Wang, Su Jun Leow, Bin Ma, Eng Siong Chng, and Haizhou Li, "Language independent query-by-example spoken term detection using n-best phone sequences and partial matching," in *proceedings of ICASSP*, 2015, pp. 5191–5195.

[11] K. Katsurada, S. Teshima, and T. Nitta, "Fast keyword detection using suffix array," Interspeech, 2009, pp. 2147–2150.

[12] N. Kanda, H. Sagawa, T. Sumiyoshi, and Y. Obuchi, "Open-vocabulary keyword detection from super-large scale speech database," MMSP, 2008, pp. 939–944.

[13] Kazuma Kon'no, Hiroyuki Saito, Shirou Narumi, Kenta Sugawara, Kesuke Kamata, Manabu Kon'no, Jinki Takahashi, and Yoshiaki Itoh, "An STD system for oov query terms integrating multiple STD results of various subword units," 10th NTCIR Workshop, 2013, pp. 592–596.

[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] K. Imami, Y. Fujii, K. Yamamoto, and S. Nakagawa, "Efficient out-of-vocabulary term detection by n-gram array indeis with distance from a syllable lattice," ICASSP, 2011, pp. 5664–5667.

[16] N. Sakamoto and S. Nakagawa, "Robust/fast out-of-vocabulary spoken term detection by n-gram index with exact distance through text/speech input," APSIPA, 2013, 4 pages.

[17] Marijn Huijbregts, Mitchell McLaren, and David van Leeuwen, "Unspervised acoustic sub-word unit detection for query-by-example spoken term detection," ICASSP, 2011, pp. 4436–4439.

[18] Haipeng Wang, Cheung-Chi Leung, Tan Lee, Bin Ma, and Haizhou Li, "An acoustic segment modeling approach to query-by-example spoken term detection," ICASSP, 2012, pp. 5157–5160.

[19] Alberto Abad, Luis Javier Rodriguez-Fuentes, Mikel Penagarikano, Amporo Varona, and German Bordel, "On the calibration and fusion of heterogeneous spoken term detection systems," Interspeech, 2013, pp. 20–24.

[20] Cheng-Tao Chung, Chun an Chan, and Lin shan Lee, "Unsuperised spoken term detection with spoken queries by multi-level acoustic patterms with varying model granularity," ICASSP, 2014, pp. 7864–7868.

[21] Chun-An Chan and Lin-Shan Lee, "Unspervised hidden markov modeling of spoken queries for spoken term detection without speech recognition," Interspeech, 2011, pp. 2141–2144.

[22] Mitsuaki Makino, Naoki Yamamoto, and Atsuhiko Kai, "Utilizing state-level distance vector representation for improved spoken term detection by text and spoken queries," Interspeech, 2014, pp. 1732–1736.

[23] N. Sakamoto, K. Yamamoto, and S. Nakagawa, "Spoken term detection based on a syllable n-gram index at the ntcir-11 spokenquery&doc task," in *proceedings of NTCIR-11*, 2014, pp. 419–424.

[24] S. Nakagawa, K. Imami, Y. Fujii, and K. Yamamoto, "A robust/fast spoken term detection method based on a syllable n-gram index with a distance metric," Speech Communication, 2012, vol. 35, pp. 470–485.

[25] N. Sakamoto and S. Nakagawa, "Spoken term detection method by using multiple recognition results of spoken query," Spoken Document Processing Workshop, 2014, 6 pages, (in Japanese).

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

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