Combination of DTW-based and CRF-based Spoken Term Detection on the NTCIR-11 SpokenQuery&Doc SQ-STD Subtask

Hiromitsu Nishizaki
University of Yamanashi
4-3-11 Takeda, Kofu
Yamanashi, 400-8511, Japan
hnishi@yamanashi.ac.jp

Naoki Sawada
University of Yamanashi
4-3-11 Takeda, Kofu
Yamanashi, 400-8511, Japan
sawada@alps-lab.org

Satoshi Natori
University of Yamanashi
4-3-11 Takeda, Kofu
Yamanashi, 400-8511, Japan
natori@alps-lab.org

Kentaro Domoto
University of Tsukuba
1-1-1 Tennodai, Tsukuba-shi
Ibaraki, 305-0006, Japan

Takehito Utsuro
University of Tsukuba
1-1-1 Tennodai, Tsukuba-shi
Ibaraki, 305-0006, Japan

ABSTRACT

Conventional spoken term detection (STD) techniques, which use a text-based matching approach based on automatic speech recognition (ASR) systems, are not robust for speech recognition errors. This paper proposes a conditional random fields (CRF)-based combination (re-ranking) approach, which recomputes detection scores produced by a phoneme-based dynamic time warping (DTW) STD approach. In the re-ranking approach, we tackle STD as a sequence labeling problem. We use CRF-based triphone detection models based on features generated from multiple types of phoneme-based transcriptions. They train recognition error patterns such as phoneme-to-phoneme confusions on the CRF framework. Therefore, the models can detect a triphone, which is one of triphones composing a query term, with detection probability. In the experimental evaluation on the NTCIR-11 SpokenQuery&Doc SQ-STD test collection, the CRF-based approach and the combination approach of the two STD systems could not outperform the conventional DTW-based approach we have already proposed.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Performance

Keywords

Multiple recognizers, phoneme transition network, spoken conetent retrieval, spoken term detection, spoken document segmentation, CRF

Term name: [**ALPS**]
Subtask: [**SQ-STD (text query only)**]
Language: [**Japanese**]

1. INTRODUCTION

Spoken term detection (STD) is designed to determine whether or not a given utterance includes a query term consisting of a word or phrase. STD research has become a hot topic in the spoken document processing research field, and the number of STD research reports is increasing in the wake of the 2006 STD evaluation organized by National Institute of Standards and Technology [1].

The difficulty in STD lies in the search for terms under a vocabulary-free framework because search terms are not known prior to a large vocabulary continuous speech recognition (LVCSR) system. Many studies tackling STD have already been proposed [2, 3]. In the past, most STD studies focused on out-of-vocabulary (OOV) and speech recognition error problems. For example, STD techniques using subword (syllable or phoneme)-based lattices or confusion networks (CN) have been proposed [3]. In recent works, we also proposed a CN-based indexing and a dynamic time warping (DTW)-based search engine [4]. The CN-based index, which we call "Phoneme Transition Network (PTN)-formed index [4]," was made of 10 types of transcriptions generated by the 10 different automatic speech recognition (ASR) systems, including an LVCSR system and a phoneme recognition system. We have shown that our proposed method could outperform other STD technologies that participated in the ninth National institute of informatics Testbeds and Community for Information access Research (NTCIR-9) project STD evaluation framework [5]. A DTW-based matching between a subword sequence of a query term and a transcription of speech is weak for speech recognition errors. Therefore, the STD performance of the DTW-based technique depends on the accuracy of subword-based transcriptions.

Our DTW-based approach using a PTN-formed index for STD was very robust for ASR errors. However, this approach output many false detections because the structure of PTN was complex [6]. These false detections degraded the STD performance. In this paper, we focus on controlling false detections in a second-pass stage using a machine learning approach. Figure 1 shows our STD framework.

We explore triphone detection modeling by using a conditional random fields (CRF)-based framework for detecting

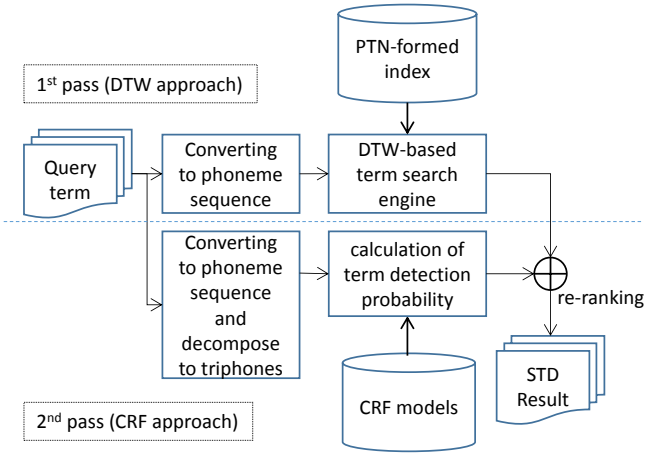

Figure 1: Overview of the two-pass STD framework using CRF-based triphone detection modeling.

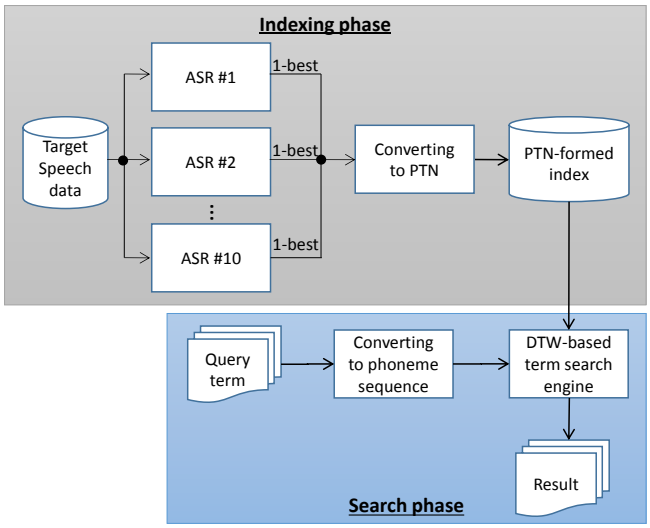

Figure 2: Overview of the first-pass stage using DTW-based matching.

query terms. A triphone detection model for each possible triphone is trained by using features generated from 10 types of phoneme-based transcriptions; all the trained models train recognition error patterns such as phoneme confusion. This approach is sensible because the features for CRF models are prepared for making a PTN-formed index, which is also derived from 10 types of transcriptions, used in the first-pass of the entire STD framework. In the STD re-ranking process, first a query term is decomposed to triphones, and for each triphone, whether or not a given utterance includes that triphone is determined using the corresponding CRF-based triphone model. Next, we calculate the probability of the product of the outputs of all the models. It is a detection probability of the query term of the given utterance. Finally, the probability is used to recompute the score of detection by the DTW-based approach. Naturally the CRF-based approach can work alone. In the experiment, we will show the STD performance of the CRF-based approach only.

Our CRF-based approach is similar to the previous researches [7, 8]. In these approaches, a phoneme sequence of a target speech is estimated by CRF models trained using ASR hypothesis-based features. This idea is close to the acoustic modeling framework using CRF [9]. The Chaudhari's technique [7, 8] was effective for the OOV detection task because the CRF models well learned the confusions of phonemes.

Our approach is positioned as an extension study of [7, 8], and solves STD as a triphone sequence labeling problem for speech data. The DTW-based approach using multiple ASR systems' outputs we have proposed improved an STD performance [4]. Therefore, we are worth trying a CRF-based triphone detection approach based on features from different types of transcriptions from ASR system's outputs.

Another machine learning approaches for STD have been recently increasing. For example, Prabhavalkar et al. [10] proposed articulatory models by discriminative training for STD under the low-resource settings. They challenged an STD framework without any LVCSR system, and their models could directly detect a query term from acoustic feature vectors. On the other hand, a multiple linear regression, support vector machines, and multi layer perceptions were also used to estimate confidence of the detected candidates in a decision process [11, 12] or in a re-ranking process [13]. Our CRF-based models train phone-to-phone confusion patterns on the basis of multiple types of transcriptions, which is different from these previous works. In addition, our study tries to investigate the effectiveness of the combination of outputs of multiple ASR systems. This is a new "cherry-picking" approach based on machine learning. The novelty of this study is that CRF is extended to provide the detection probability of a query term, and it is also used in the decision process on the second pass of our STD framework by combining the DTW-based STD score with the CRF-based probability.

The rest of the paper is organized as follows: we first shortly present the baseline system (the DTW-based approach) in Section 2, and then introduce the CRF-based triphone detection modeling and how to detect a query term from speech in Section 3. Section 4 explains about the re-ranking process using the CRF-based STD. The experimental settings and results are presented in Section 5, and some conclusions in Section 6.

2. DTW-BASED APPROACH USING MULTIPLE ASR SYSTEMS' OUTPUTS

The DTW-based STD approach using PTN-formed index [4] is performed in the first-pass stage in the entire STD framework. It is also the baseline approach. Figure 2 shows an overview of the baseline method. In the indexing phase, speech data is performed by ASR, and the recognition outputs (words or sub-word sequences) are converted into the PTN-formed index for STD. Figure 3 shows an example of a PTN-formed index.

In the search phase, the word-formed query is converted into a phoneme sequence. Then, the phoneme-formed query is input to the term search engine. The term search engine searches the query term from the index at the phoneme level using the DTW framework. Unlike combination techniques of multiple STD systems like [20], the baseline system combines the transcriptions produced by multiple ASR systems.

Figure 4 represents an example of the DTW framework

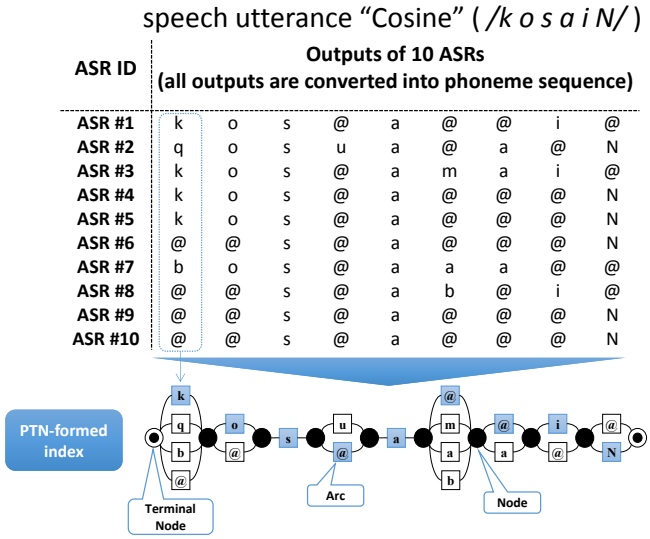

Figure 3: Generating a PTN-formed index by performing alignment using DP and converting to a PTN.

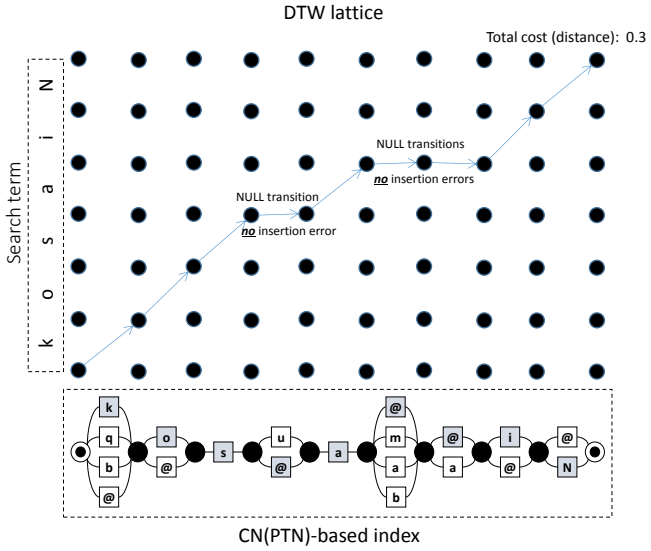

Figure 4: Example of term search on network-formed index.

between the search term "k o s a i N" (cosine) and the PTN-formed index. The PTN has multiple arcs between two adjoining nodes. These arcs are compared to one of the phoneme labels of a query term. We use edit distance as cost on the DTW paths, and the cost value for substitution, insertion, and deletion errors is commonly set to 1.0. The details of this approach is discussed on our previous paper [4].

3. CRF-BASED TRIPHONE DETECTION MODELING

Figure 5 shows an overview of the STD process using CRF-based triphone detection modeling in the second-pass stage in the entire STD framework. In this study, we use just 10 types of phoneme-based transcriptions generated by 10 different ASR systems for training CRF-based models. A query term is translated into a phoneme sequence and decomposed into triphones. Then, a CRF-based triphone model calculates the existence probability of the triphone corresponding to that model in an utterance. The final term detection probability (or score) is based on the product of the outputs of all the models. In this research, we prepared two types of acoustic models (AMs), five types of language models (LMs), and a decoder. The combinations of AMs and LMs produced 10 ASR systems. The model details will be explained in Section 5.1.

CRFs [14] have been successfully used in numerous text processing tasks, such as named-entity extraction [15] and phrase chunking [16]. In the speech processing area, CRFs are used for sentence boundary detection [17] and OOV detection in speech [18].

The conditional probability of an input sequence $\mathbf{x}$, given an output label sequence $\mathbf{y}$, is calculated as:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\exp(\sum_k \lambda_k F_k(\mathbf{y},\mathbf{x})) \quad (1)$$

where $F_k(\mathbf{y},\mathbf{x})$ is a feature vector for input sequence $\mathbf{x}$ and label sequence $\mathbf{y}$, and λ_k is a weight parameter for $F_k(\mathbf{y},\mathbf{x})$. $Z(\mathbf{x})$ is a normalization factor given by:

$$Z(\mathbf{x}) = \sum_y \exp(\sum_k \lambda_k F_k(\mathbf{y},\mathbf{x})) \quad (2)$$

To train phoneme-to-phoneme confusions as the error pattern training, we utilized features on the basis of the phoneme-based transcriptions by 10 different ASRs shown in Figure 6, representing examples of features derived from phoneme-based transcriptions and the beginning, inside, outside (BIO) encoding. A dynamic programming (DP)-based alignment procedure [19] was performed on phoneme-based transcriptions to make an alignment for each transcription. We utilized the BIO encoding in the CRF-based triphone detection modeling. Therefore, each CRF-based model finds BI tag sequence(s) from an utterance. A CRF-based model was trained for each possible triphone generated by pronunciation of the words. As shown in Figure 6, phoneme-based unigram, bigram, and trigram features were used for CRF-based model training. The point of this modeling is to use cross-ASR-based features. The cross-ASR bigram features enable a CRF-based model to capture phoneme-to-phoneme confusion error patterns. The model can then robustly detect triphones from erroneous transcriptions.

The detection probability $P(T|\mathbf{x}_i)$ of a query term T consisting of N triphones in utterance i is calculated by the following equation:

$$P(T|\mathbf{x}_i) = (\prod_{j=1}^{N} P_{t_j}(\mathbf{y}|\mathbf{x}_i))^{\frac{1}{N}}, \ (l_{t_1} < l_{t_j} < l_{t_N}) \quad (3)$$

where t_j is the j-th triphone of T, $\mathbf{x}_i$ is the input sequence of utterance i, and $\mathbf{y}$ is a part of output labe sequence. l_{t_j} means the location (position) of the beginning phoneme of triphone t_j. $P_{t_j}(\mathbf{y}|\mathbf{x}_i)$ is not calculated by using the conditional probability of the whole label sequence for utterance i but calculated based on the product of probability of each tag: B and I tag. A probability of O tag output is not considered. This idea is similar to maximum entropy modeling.

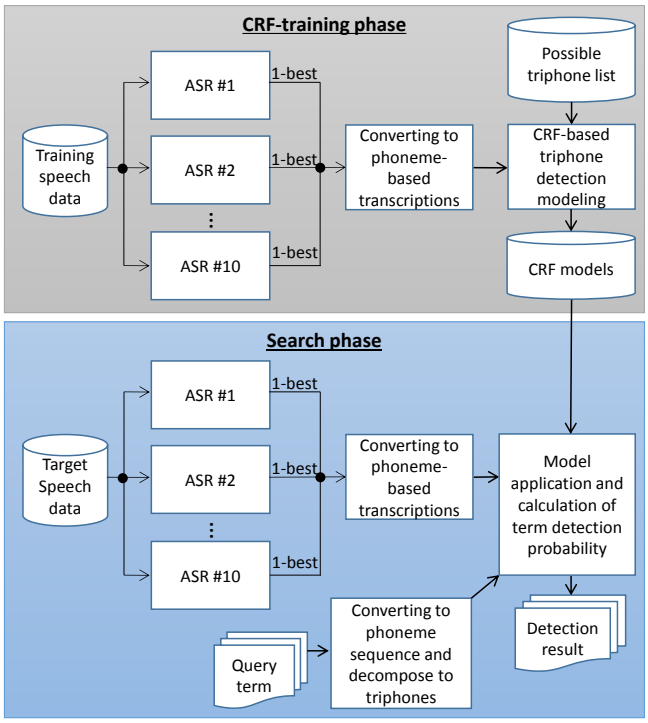

Figure 5: Overview of the STD framework using CRF-based triphone detection modeling.

However, CRFs have an ability to find an optimal labeling for the entire sequence. Therefore, CRF-based models can detect triphones with high accuracy. Finally, triphone t_j detection probability is calculated by:

$$P_{t_j}(\mathbf{y}|\mathbf{x}_i) = \prod_{L=\mathrm{B}}^{\mathrm{I_{tail}}} P_{t_j}(L|\mathbf{x}_i) \tag{4}$$

where B and $\mathrm{I_{tail}}$ represent the begining tag and tailing tag of triphone t_j, respectively. In other words, the detection probability of t_j is calculated by making the product of the conditional probability of each tag between the head B and tailing $\mathrm{I_{tail}}$ tags. If $P_{t_j}(\mathbf{y}|\mathbf{x}_i)$ is less than probability ϕ, $P_{t_j}(\mathbf{y}|\mathbf{x}_i)$ is set to ϕ. This prevents the very low detection probability of T when any triphone consisting of T cannot be detected. In this study, ϕ is heuristically set to 0.01. If $P(T|\mathbf{x}_i)$ is greater than a threshold θ_C, term T seems to be in utterance i. Changing the threshold θ_C enables us to draw the recall-precision curve on the evaluation.

4. RE-RANKING OF FIRST-PASS DETECTIONS

We tried a simple combination of a DTW-based score and a CRF-based score (same as detection probability) as following equation, which is well-known as a weighted harmonic mean. The recomputed score $\mathrm{RS}(T,i)$ of the detection is calculated as follows:

$$\mathrm{RS}(T,i) = \frac{(\gamma^2+1)\cdot \mathrm{DTW}(T,i)\cdot \mathrm{CRF}(T,i)}{\gamma^2\cdot \mathrm{DTW}(T,i) + \mathrm{CRF}(T,i)} \tag{5}$$

where γ is a weight parameter that controls a balance between $\mathrm{CRF}(T,i)$ and $\mathrm{DTW}(T,i)$, $\mathrm{CRF}(T,i)$ and $\mathrm{DTW}(T,i)$

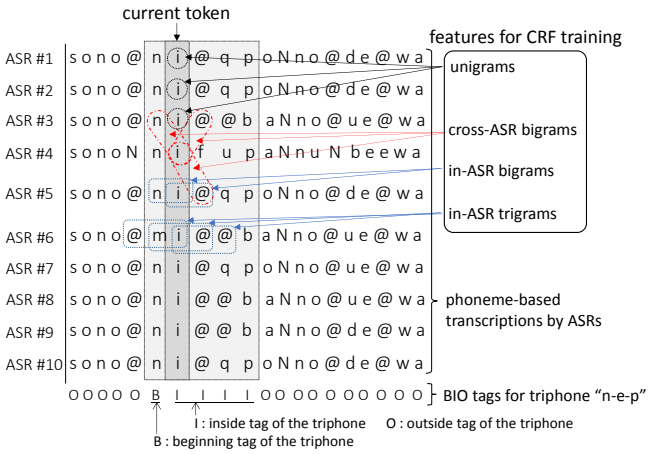

Figure 6: Example of features for CRF model training and BIO encoding.

are scores of term T in utterance i derived by the CRF-based and the DTW-based STD methods, respectively. Both of scores from the two approaches ranges from 0 to 1. γ is set to 0.08, which is determined by the moderate-size query set used in the NTCIR-10 SpokenDoc-2 [23], and common for the all query terms on the test collection.

5. STD EXPERIMENT

5.1 Target test collection

The Corpus of the 1st to 7th Spoken Document Processing Workshop (SDPWS1to7) is to be used as the document collection for evaluating the NTCIR-11 SpokenQuery&Doc SQ-STD subtask.

5.1.1 Speech Recognition

As shown in Figure 1, the SDPWS1to7 speech data is recognized by the 10 ASRs. Julius ver. 4.1.3 [22], an open source decoder for LVCSR, is used in all the systems.

We prepared two types of acoustic models (AMs) and five types of language models (LMs) for constructing the PTN. The AMs are triphone based (Tri.) and syllable based HMMs (Syl.), where both types of HMMs were trained from the spoken lectures in the Corpus of Spontaneous Japanese (CSJ) [21].

All the LMs are word and character based trigrams as follows:

WBC : word based trigram in which words are represented by a mix of Chinese characters, Japanese Hiragana and Katakana.

WBH : word based trigram in which all words are represented only by Japanese Hiragana. The words composed of Chinese characters and Katakana are converted into Hiragana sequences.

CB : character based trigram in which all characters are represented by Hiragana.

BM : character sequence based trigram in which the unit of language modeling is two of Hiragana characters.

Non : No LM is used. Speech recognition without any LM is equivalent to phoneme (or syllable) recognition.

Each model is trained from the many transcriptions in the CSJ under the open for the speech data of STD.

Finally, the ten combinations, comprising two AMs and five LMs, are formed. The condition is completely the same as the description in the overview paper [24].

5.1.2 *Query set of the STD subtasks*

The NTCIR-11 SpokenQuery&Doc organizers provided two types of query sets: the text query set and the spoken query set[24]. We evaluated our STD engine on the text query set only.

5.2 Training CRF-based model

CRF-based triphone detection models were trained from a part of the CSJ except the 177 lecture speeches using a CRF++ toolkit[1]. A total of 1,200 speeches were used to train models. The number of trained triphone models in this study was 10,600, derived from 48 types of Japanese monophones, and we did not adapt any clustering algorithm for grouping similar triphones together as AM training before training CRF-based models. The most rarely occurring triphone "n-e-p" among the all triphones included in the query terms existed in only 10 utterances.

5.3 Evaluation metrics

The evaluation metrics used in this study were recall, precision, F-measure, and mean average precision (MAP) values [5, 24]. These measurements are frequently used to evaluate information retrieval performance and are defined as follows:

$$Recall = \frac{N_{corr}}{N_{true}} \tag{6}$$

$$Precision = \frac{N_{corr}}{N_{corr} + N_{spurious}} \tag{7}$$

$$F - measure = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \tag{8}$$

Here N_{corr} and $N_{spurious}$ are the total number of correct and spurious (false) term detections, respectively, and N_{true} is the total number of true term occurrences in the speech data. The F-measure values for the optimal balance of *Recall* and *Precision* values are denoted by "Max. F-measure."

The STD performance for the query sets can be illustrated by a recall−precision curve, which is plotted by changing the threshold θ_C in the CRF-based STD method or θ_D in the DTW-based baseline.

MAP is the mean of the average precision values for each query term. It can be calculated as follows:

$$MAP = \frac{1}{Q} \sum_{q=1}^{Q} AveP(q) \tag{9}$$

where Q is the number of whole queries and $AveP(q)$ denotes the average precision of the q-th query term of the query set. Average precision is calculated by averaging the precision values computed for each relevant term in the list in which retrieved terms are ranked by a relevance measure.

$$AveP(q) = \frac{1}{Rel_q} \sum_{r=1}^{N_q} (\delta_r \cdot Precision_q(r)) \tag{10}$$

where r is the rank, N_q is the rank number at which all the relevance terms of query term q are detected, and Rel_q is the number of the relevance terms of the query term q. δ_r is a binary function for a given rank r.

[1]CRF++: Yet Another CRF toolkit, `https://code.google.com/p/crfpp/`

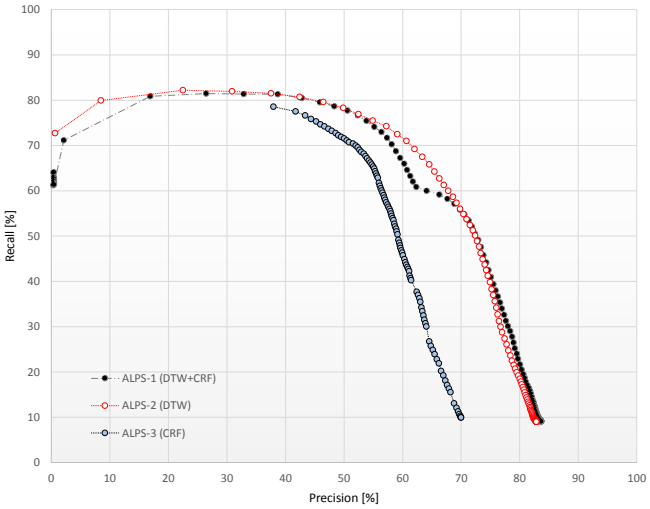

Figure 7: Recall-precision curves of the STD methods.

5.4 Experimental results

Figure 7 shows recall-precision curves of the each STD approaches. Table 1 also represents maximum (max.) F-measure (micro and macro averages) and MAP values of our STD methods. We compared the STD performances between three STD methods in this study. The STD system "ALPS-2 (DTW)" explained in Section 2 is the baseline in this study, the system "ALSP-3 (CRF)" is the CRF-based approach only, and "ALPS-1 (DTW+CRF))" is the the proposed approach that recomputes the scores of the detections by the baseline.

As shown in Figure 7 and Table 1, the CRF-based STD alone did not work well comparing to the baseline approach. In addition, our proposed approach also did not outperformed the baseline at the best F-measure point. However, the combination method slightly improved the precisions in the area of high recall range.

On the other hand, we have already evaluated our STD engine on the different STD test collection based on the CSJ speeches[25]. On the collection, the combination of DTW and CRF-based approaches outperformed the baseline (same as the ALPS-2 system)[26], and got the best performance on max. F-measure and MAP values among all the systems because the recall-precision curve completely improves. The all queries of the test collection was composed of OOV terms. However, the most queries of NTCIR-11 SpokenQuery&Doc SQ-STD subtask consist of IV terms. Therefore, the simple method may get better performance rather then the others on the IV query set. In fact, the baseline result provided by the task organizers is the 2nd best performance on the

Table 1: STD performances for the each query run.

run	micro ave.		macro ave.			index	search
	max. F [%]	spec. F [%]	max. F [%]	spec. F [%]	MAP	size [MB]	speed [s]
ALPS-1	63.72	61.38	57.19	56.62	0.666	713	8.125
ALPS-2	65.54	53.56	58.52	50.56	0.672	591	6.770
ALPS-3	59.86	59.86	52.94	52.61	0.553	122	0.887

same text query set among all the runs[24]. In addition, the CRF-based triphone detection models were trained by the transcriptions from the CSJ speeches. Therefore, there may be some mis-matches between the training and testing data.

6. CONCLUSION

In this paper, we proposed a CRF-based re-ranking approach that recomputes the scores of the detections by the DTW-based STD engine. The CRF model finds triphones composed of a query term from an utterance. We used CRF-based triphone detection models based on features generated from multiple types of phoneme-based transcriptions that are used for making the PTN-formed index used in the DTW-based approach. The aim of this approach is to train recognition error patterns such as phoneme-to-phoneme confusions on the CRF framework and to control the false detections from the DTW approach.

In the STD experiment on the NTCIR-11 SpokenDoc&Query SQ-STD subtask, the CRF-based approach and the re-ranking method which combines the CRF-based and DTW-based approach could not outperform the DTW-based STD approach by using the outputs of multiple ASR systems.

As future work, we are going to study a triphone clustering approach to train CRF-based models. This approach may solve the training data shortage problem and improve detection accuracy of each triphone.

7. ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (B) Grant Number 26282049 and Grant-in-Aid for Scientific Research (C) Grant Number 24500225.

8. REFERENCES

[1] NIST, "The Spoken Term Detection (STD) 2006 evaluation plan," `http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf`, 2006, Accessed: 4th/7/2014.

[2] D. Vergyri, I. Shafran, A. Stolcke, R. R. Gadde, M. Akbacak, B. Roark, and W. Wang, "The SRI/OGI 2006 spoken term detection system," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH2007)*. 2007, pp. 2393–2396.

[3] S. Meng, J. Shao, R. P. Yu, J. Liu, and F. Seide, "Addressing the Out-of-Vocabulary Problem for Large-scale Chinese Soken Term Detection," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH2008)*. 2008, pp. 2146–2149.

[4] Satoshi Natori, Yuto Furuya, Hiromitsu Nishizaki, and Yoshihiro Sekiguchi, "Spoken Term Detection Using Phoneme Transition Network from Multiple Speech Recognizers' Outputs," *Journal of Information Processing*, Vol. 21, No. 2, pp. 176–185, 2013.

[5] Tomoyosi Akiba, Hiromitsu Nishizaki, Kiyoaki Aikawa, Tatsuya Kawahara, and Tomoko Matsui, "Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop," in *Proceedings of the 9th NTCIR Workshop Meeting*, 2011, pp. 223–235.

[6] Satoshi Natori, Yuto Furuya, Hiromitsu Nishizaki, and Yoshihiro Sekiguchi, "Entropy-based False Detection Filtering in Spoken Term Detection Tasks," in *Proceedings of the 5th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2013)*, 2013, pp. 1–7.

[7] Upendra V. Chaudhari and Michael Picheny, "Improved vocabulary independent search with approximate match based on conditional random fields," in *Proceedings of the IEEE Internationl Workshop on Automatic Speech Recognition and Undersitanding (ASRU2009)*, 2009, pp. 416–420.

[8] Upendra V. Chaudhari and Michael Picheny, "Matching criteria for vocabulary-independent search," *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 20, No. 5, pp. 1633–1643, 2012.

[9] Asela Gunawardana, Milind Mahajan, Alex Acero, and John C. Platt, "Hidden conditional random fields for phone classification," in *Proceedings of the 6th Annual Conference of the International Speech Communication Association (INTERSPEECH2008)*. 2005, pp. 1117–1120.

[10] R. Prabhavalkar, K. Livescu, E. Fosler-Lussier, and J. Keshet, "Discriminative articulatory models for spoken term detection in low-resource conversational settings," in *Proceedings of The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2013)*, 2013.

[11] Dong Wang, Simon King, Joe Frankel, and Peter Bell, "Term-dependent confidence for out-of-vocabulary term detection," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH2009)*. 2009, pp. 2139–2142.

[12] J. Tejedor, A. Echeverria, and Dong Wang, "An evolutionary confidence measurement for spoken term detection," in *Proceedings of the 9th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2011, pp. 151–156.

[13] Tsung wei Tu, Hung yi Lee, and Lin shan Lee, "Improved spoken term detection using support vector machines with acoustic and context features from pseudo-relevance feedback," in *Proceedings. of the IEEE Internationl Workshop on Automatic Speech Recognition and Undersitanding (ASRU2011)*, 2011, pp. 383–388.

[14] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of The 18th International Conference on Machine Learning (ICML '01)*, 2001, pp. 282–289.

[15] K. Nongmeikapam, T. Shangkhunem, N.M. Chanu, L.N. Singh, B. Salam, and S. Bandyopadhyay, "CRF based Name Entity Recognition (NER) in Manipuri: A Highly Agglutinative Indian Language," in *Proceedings of the 2nd National Conference on Emerging Trends and Applications in Computer Science (NCETACS)*, 2011, pp. 1–6.

[16] Fei Sha and Fernando Pereira, "Shallow parsing with conditional random fields," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL '03)*, 2003, pp. 134–141.

[17] Yang Liu, Andreas Stolcke, Elizabeth Shriberg, and Mary Harper, "Using conditional random fields for sentence boundary detection in speech," in *Proceedings of The 43rd Annual Meeting on Association for Computational Linguistics (ACL '05)*, 2005, pp. 451–458.

[18] Carolina Parada, Mark Dredze, Denis Filimonov, and Frederick Jelinek, "Contextual information improves oov detection in speech," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)*, 2010, pp. 216–224.

[19] J. G. Fiscus, "A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition adn Understanding (ASRU'97)*, 1997, pp. 347–354.

[20] Murat Akbacak, Lukas Burget, Wen Wang, and Julien van Hout, "Rich system combination for keyword spotting in noisy and acoustically heterogeneous audio streams," in *Proceedings of The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2013)*, 2013, pp. 8267–8271.

[21] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*. 2003, pp. 7–12.

[22] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," in *Proceedings of the 1st Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC2009)*, 2009, pp. 131–137.

[23] Tomoyosi Akiba, et al., "Overview of the NTCIR-10 SpokenDoc-2 Task," in *Proceedings of the 10th NTCIR Conference*, 2012, pp. 573–587.

[24] T. Akiba, H. Nishizaki, H. Nanjo, and G. Jones, "Overview of the NTCIR-11 SpokenQuery&Doc Task," in *Proceedings of the 11th NTCIR Conference*, 2014.

[25] Yoshiaki Itoh, Hiromitsu Nishizaki, Xinhui Hu, Hiroaki Nanjo, Tomoyosi Akiba, Tatsuya Kawahara, Seiichi Nakagawa, Tomoko Matsui, Yoichi Yamashita, and Kiyoaki Aikawa, "Constructing japanese test collections for spoken term detection," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH2010)*. 2010, pp. 677–680.

[26] N. Sawada, S .Natori and H. Nishizaki, "Re-Ranking of Spoken Term Detections Using CRF-based Triphone Detection Models," in *Proceedings of the 1st Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC2014)*, 2014, pp. 1–4.