



Japanese Spoken Term Detection Using Syllable Transition Network Derived from Multiple Speech Recognizers' Outputs

Satoshi Natori¹, Hiromitsu Nishizaki², and Yoshihiro Sekiguchi²

¹Department of Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Japan

natori@alps.cs.yamanashi.ac.jp¹, {hnishi, sekiguti}@yamanashi.ac.jp²

Abstract

This paper proposes a spoken term detection using syllable transition network (STN) derived from multiple speech recognizers. An STN is similar to a sub-word based confusion network, which is derived from the output of a speech recognizer. The one we proposed is derived from the outputs of multiple speech recognition systems, which is well known to be robust to certain recognition errors and the out-of-vocabulary problem. Therefore, the STN should also be robust to recognition errors on the STD. This experiment showed that the STN was very effective at detecting out-of-vocabulary terms, improving detection rate to 83%, which was as high as the in-vocabulary term detection performance.

Index Terms: spoken term detection, confusion network, multiple recognizers' outputs

1. Introduction

Recently, environments have evolved in which a large number of audio and multimedia archives, such as video archives, and digital libraries can be utilized easily. In particular, a rapidly increasing number of spoken documents, such as broadcast programs, spoken lectures, and recordings of meetings are archived, and some of these can be accessed through the Internet. The need to retrieve such spoken information has been growing, while an effective retrieval technique is definitely lacking at present; thus, the development of technology for retrieving such information has become increasingly important.

In the TREC Spoken Document Retrieval (SDR) track hosted by NIST and DARPA in the second half of the 1990s, a number of studies of SDR were presented on the subject of English and Mandarin broadcast news documents. Meanwhile, the NIST began a Spoken Term Detection (STD) project with the pilot evaluation and workshop in 2006; it is different from SDR. The aim of STD is to find the position of spoken terms selected for evaluation in audio archives.

The difficulty in STD lies in the search for terms under a vocabulary-free framework, because the search terms are not known a priori to a speech recognizer. Many studies tackling an STD task have already been proposed [1, 2]. Most of STD studies focused on the out-of-vocabulary (OOV) and speech recognition error problems. For example, STD techniques that use entities such as sub-word lattice and confusion network (CN) have been proposed.

This paper describes STD from spoken lectures using syllable transition networks (STNs) derived from multiple speech recognizers' outputs. Our main idea, which is different from the typical STD technique, is to use multiple speech recognizers, where syllable sequences from these recognizers' outputs

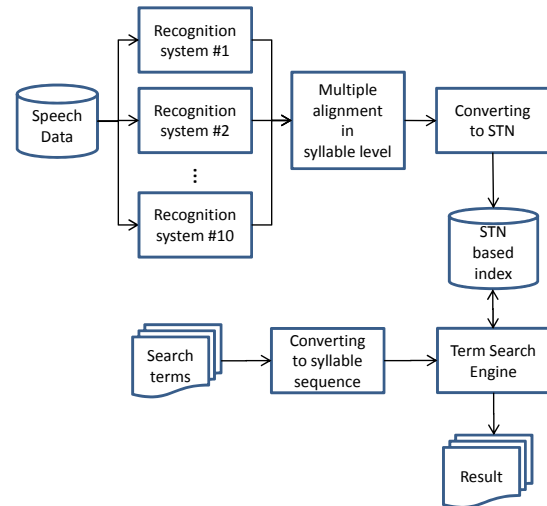


Figure 1: Overview of STD framework based on STN.

are combined and converted into an STN. This study uses ten types of speech recognition systems with the same decoder for all. Two types of acoustic models (triphone based or syllable based HMMs) and five types of language models (word based or sub-word based) were prepared.

The use of multiple recognizers and their outputs is very effective in improving speech recognition performance. For example, Fiscus [3] proposed the ROVER method which adopts a word voting scheme. Utsuro et al. [4] developed a technique for combining multiple recognizers' outputs by using a support vector machine (SVM) to improve speech recognition performance. Application of characteristics of word (or sub-word) sequences output by recognizers possibly makes good STD performance, because these characteristics are different for each speech recognizer. The STN based on outputs of multiple speech recognizers can cover more sub-word sequences of spoken terms.

An experimental result for the STD showed that the STN was effective in improving STD performance; in particular, it was very robust for OOV term detection.

2. Syllable Transition Network on STD task

An STN is almost the same as sub-word based confusion network (CN) derived by a single speech recognizer. Gao et al. [5] have proposed using a sub-word based CN for STD. This CN was built from a single speech recognizer, whereas the STN we propose is built from syllable sequences converted from the outputs of multiple speech recognition systems. Using multiple

10.21437/Interspeech.2010-259

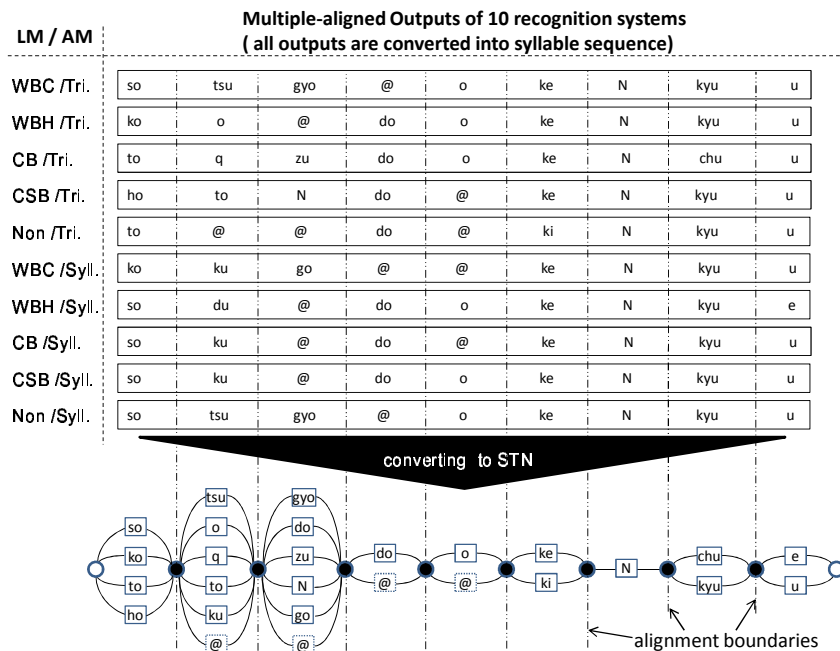


Figure 2: Example of converting outputs of ten recognition systems into STN by multiple alignment.

outputs contributes well to speech recognition, as shown in [3] and [4]. Therefore, it may also be effective for improving STD performance because of its robustness to recognition errors.

Figure 1 represents an outline of the STD framework in this paper. First, speech data is recognized by ten speech recognition systems and the outputs from these systems are converted into syllable sequences. Then, multiple alignment is performed by dynamic programming and the aligned data is translated into an STN. This section first explains the ten speech recognition systems, and describes how the recognizer outputs are converted. Then, an STD engine for an STN based index is explained.

2.1. Speech Recognition by 10 Speech Recognition Systems

As shown in Figure 1, the speech data is recognized by ten speech recognizers. Julius rev. 4.1.3, an open source decoder for LVCSR, is used in all the systems. Julius can output N-best hypotheses.

We prepared two types of acoustic models (AMs) and five types of language models (LMs) for constructing an STN. The AMs were triphone based and syllable based HMMs, consisting of 43 phonemes and 124 syllables, respectively, where both types of HMMs were trained from the 797 spoken lectures in the Corpus of Spontaneous Japanese (CSJ) [6]. Feature vectors consist of 38 dimensions: 12 dimensional Mel-frequency cepstrum coefficients (MFCCs), the cepstrum difference coefficients (delta MFCCs), its acceleration (delta delta MFCCs), delta power, and delta delta power, and they were calculated every 10 msec. The distribution of the acoustic features was modeled using 32 mixtures of diagonal covariance Gaussians for the HMMs.

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 and 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.

CSB : character sequence based trigram in which the unit of language modeling is a few sequences 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.

Finally, the ten combinations, comprising two AMs and five LMs, are formed.

2.2. Conversion to STN

All ten outputs from the recognizers, some of which are word (or phoneme) sequences, are converted into syllable sequences for the outputs to be aligned at the syllable level.

After finishing the conversion, syllable based alignment, called the “multiple alignment,” for all outputs is performed using a dynamic programming scheme. This is the similar algorithm to the ROVER method [3] in syllable-level. Figure 2 represents an example of multiple outputs from each recognition system and the syllable based alignment result of all the recognizers’ outputs. In Figure 2, “**Tri.**” indicates the triphone based HMMs, and “**Syll.**” represents syllable based HMMs.

Finally, the aligned data is converted into an STN. “@” in Figure 2 indicates a null transition. Arcs between nodes in the STN have syllables and null transitions with an occurrence probability. However, in this paper, we do not use any syllable occurrence probabilities in the STD task.

2.3. Term Search Engine

A search term is converted to a syllable sequence using a recognition dictionary and input to a term search engine.

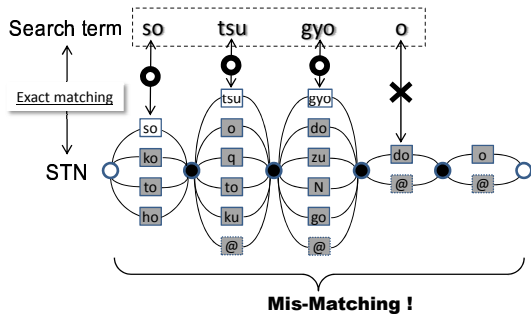


Figure 3: An example of exact matching between a search term and the STN.

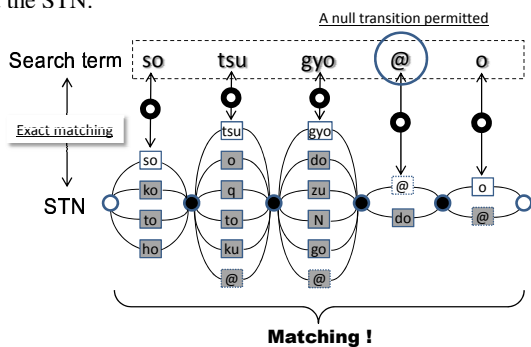


Figure 4: An example of exact matching with null transition (“Match”) between a search term and the STN.

We adopted a basic term search method because the study focuses on investigating the advantages of using multiple recognizers. The search engine tries to find syllable sequences that match exactly a syllable sequence in a search (query) term in the STN. The engine does not use any scores, such as entropy and occurrence probability, but simply decides whether the search term exists in the STN. This is because we investigate the effectiveness of using multiple speech recognizers. Figure 3 shows an example of exact matching between a search term “*so tsu gyo o*” (it means “graduation” in Japanese) and an STN based index.

However, this matching method is limited; therefore, most of the syllable sequences may not match a search term. Consequently, null transition matching is allowed for search terms. This is indicated by inserting special symbol(s), “@” between syllables converted from a search term. As shown in Figure 4, the search term will return the appropriate syllable sequence in the STN. This search method is denoted by “**Match**”.

On the other hand, **Match** may be less powerful on the index which is made from errorful transcriptions. So, we introduce another matching method, denoted by “**Tol.1/4**”, between a query and STN. It allows less than two mismatch for four syllables of a query term. For example, as shown in Figure 5, if a query term “*ko sa i N*” consists of 4 syllables, “Tol.1/4” allows under two mismatches between the query and the STN.

3. STD Experiment

3.1. Speech Data

We used a subset of the Japanese test collection [7] for spoken term detection for evaluating our method. This test collection has been constructed by a working group of Special Interest Group Spoken Language Processing (SIG-SLP) of the Infor-

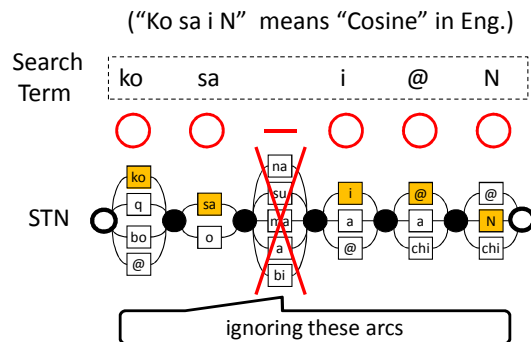


Figure 5: An example of lax matching with null transition (“Tol.1/4”) between a search term and the STN.

Table 1: Syllable-based speech recognition performance.[%]

LM/AM	Corr.	Acc.
WBC/Tri	86.5	83.0
WBH/Tri	86.3	81.4
CB/Tri	81.8	77.4
CSB/Tri	85.7	81.0
Non/Tri	71.0	51.2
WBC/Syl	79.1	76.4
WBH/Syl	79.3	75.8
CB/Syl	73.8	71.2
CSB/Syl	78.6	75.4
Non/Syl	63.7	45.4

mation Processing Society of Japan.

The CSJ is used as a spoken document test set. It includes 2,702 speeches including actual academic presentations and simulated public speech. In this paper, however, we used only 177 speeches, specially called “CORE”, from the whole data set. They are not included in the training data set of the acoustic and language models.

Table 1 shows syllable-based correct and accuracy rates of each speech recognizer are represented by “Corr.” and “Acc.,”. The best combination of AM and LM on syllable recognition performance is the word trigram-based language model and the triphone-based acoustic model. A total of about 44 hours of speech materials was used for the STD.

3.2. Search Terms

A wide variety of search terms, which includes Japanese single-word, multi-word terms, and common terms and rare terms, are prepared in the test set. All terms are uttered in the test speech data.

A total of 858 terms (78 types of terms) were prepared for a search experiment; 50 types of terms are in-vocabulary (IV) terms registered in the WBC LM; the other 28 types are OOV terms.

3.3. STD Evaluation Metric

The evaluation metrics we used are recall, precision rate, and F-measure. They are defined as follows:

$$Recall(t) = \frac{N_{corr}(t)}{N_{true}(t)} \quad (1)$$

$$Precision(t) = \frac{N_{corr}(t)}{N_{corr}(t) + N_{spurious}(t)} \quad (2)$$

$$F - measure(t) = \frac{2 \cdot Recall(t) \cdot Precision(t)}{Recall(t) + Precision(t)} \quad (3)$$

where N_{corr} and $N_{spurious}$ are the total number of correct and spurious (incorrect) term detections, and N_{true} is the total number of true term occurrences in the speech data. The STD performances for each IV set and OOV set are obtained by averaging $Recall(t)$, $Precision(t)$, and $F - measure$ for 50 and 28 search terms, respectively.

3.4. Experimental Result

Table 2 and Table 3 show STD performance on the STD test-set for IV and OOV terms, respectively. In each table, (a) and (b) represent the results when “Match” and “Tol.1/4” search engines are used, respectively.

We prepared four types of indices for the STD in this experiment as follows:

- **1-Best** index is derived from only 1-best hypothesis of the WBC/Tri. (LM/AM) recognizer, which shows the best performance among all the recognizers. This is a word-based index.
- **10-Best** includes ten types of sub-indices, each of which is derived from 10-best hypotheses of the WBC/Tri. recognizer. This is a word-based index.
- **10 Outputs** includes ten types of sub-indices, which are derived from each recognizer’s output (1-best hypothesis). These are syllable-based indices.
- **STN** represents our proposed STN based index.

For the 10-Best and 10 Outputs index, it is acceptable that a search term is matched on any sub-index.

As shown in Table 2, 10 Outputs and STN indices could improve the recall rate for the IV terms compared to 1-Best and 10-Best, although false alarm errors were increased. Using N-best hypotheses from the single recognizer is well-known as powerful technique on a wide variety of spoken language processing. But, this experimental results claim that using various transcriptions from multiple speech recognizers’ outputs is effective to improve STD performance. However, there is no difference from 10 Outputs and STN on recall, whereas STN has higher expression ability of the hypothesis than 10-Best.

Table 3 shows the results of STD experiment on the OOV query test set. Because 1-Best and 10-Best are word-based indices, no performance is obtained. Comparing with 10 Outputs and STN, STN outperforms the 10 Outputs on recall. The recall rates are improved by 8% and 13% using “Match” and “Tol.1/4” search engine, respectively. A fusion of multiple speech recognizers’ outputs enables flexible search of a query term under the case of using speech recognition technology.

The STN could make more transitions between syllables available, but it induced false alarms that degraded the precision. However, to perform a less-restricted matching scheme with score (likelihood or cost), such as an edit distance measure, on a search engine makes the STD performance much better in an STN framework.

Finally, our technique can find 83% of OOV among 44 hours speech data. This OOV query term recall (detection) rate is as high as IV query terms detection (84%).

4. Conclusions

This paper proposed an STN, essentially a syllable based confusion network, derived from multiple speech recognizers’ out-

Table 2: STD performance for IV query terms.

(a) “Match”			
Type of Index	Recall	Precision	F-measure
1-Best	0.59	0.91	0.69
10-Best	0.61	0.90	0.70
10 Outputs	0.68	0.90	0.75
STN	0.70	0.82	0.73
(b) “Tol.1/4”			
Type of Index	Recall	Precision	F-measure
10 Outputs	0.84	0.34	0.41
STN	0.84	0.17	0.23

Table 3: STD performance for OOV query terms.

(a) “Match”			
Type of Index	Recall	Precision	F-measure
1-Best	0.00	0.00	0.00
10-Best	0.00	0.00	0.00
10 Outputs	0.32	0.43	0.36
STN	0.40	0.41	0.41
(b) “Tol.1/4”			
Type of Index	Recall	Precision	F-measure
10 Outputs	0.70	0.29	0.31
STN	0.83	0.09	0.12

puts for the STD task. The main aim was to use multiple outputs of speech recognizers for constructing the STN, which is different from the sub-word based approaches proposed earlier.

Experimental results showed that the STN functioned well in improving the STD performance on OOV search terms. For IV terms, the STN and simple 10 sorts of recognizers could retrieve the most appropriate terms from the speech data compared with the basic indices. Using multiple recognizers for the STD task is effective to improve STD performance.

In future, we intend to introduce certain score such as occurrence probability, when a search term is matched on the STN. In addition, a combination method of word and sub-word based matching schemes on the STN will be adopted to improve STD performance.

5. References

- [1] D. Vergyri et al., “The SRI/OGI 2006 spoken term detection system,” *Proc. of the INTERSPEECH2007*, 2007, pp. 2393–2396.
- [2] S. Meng et al., “Addressing the out-of-vocabulary problem for large-scale Chinese spoken term detection,” *Proc. of the INTERSPEECH2008*, 2008, pp. 2146–2149.
- [3] J. G. Fiscus, “A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER),” *Proc. of the ASRU’97*, 1997, pp. 347–354.
- [4] T. Utsuro et al., “An empirical study on multiple LVCSR model combination by machine learning,” *Proc. of the HLT-NAACL2004*, 2004, pp. 13–16.
- [5] J. Gao et al., “Spoken term detection using dynamic match sub-word confusion network,” *Proc. of the 4th ICNC*, 2008, pp. 250–254.
- [6] Kikuo Maekawa, “Corpus of spontaneous Japanese: Its design and evaluation,” *Proc. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.
- [7] Y. Itoh et al., “Constructing Japanese Test Collections for Spoken Term Detection,” *Proc. of the INTERSPEECH 2010*, 2010.