Intensive Acoustic Models Constructed by Integrating Low-Occurrence Models for Spoken Term Detection

Shiro Narumi[1], Kazuma Konno[1], Takuya Nakano[1], Yoshiaki Itoh[1], Kazunori Kojima[1], Masaaki Ishigame[1], Kazuyo Tanaka[2] and Shi-wook Lee[3]

[1] Iwate Prefectural University, Japan

[2] Tsukuba University, Japan

[3] National Institute of Advanced Industrial Science and Technology, Japan

y-itoh@iwate-pu.ac.jp

Abstract

Triphone acoustic models are often used as subword models for detecting out-of-vocabulary query terms in Spoken Term Detection (STD) systems. Our preliminary experiments revealed that the training data for a large portion of the approximately 8,000 triphone models are insufficient. Assuming that such insufficient models deteriorate the performance of STD, this paper proposes intensive triphone models constructed by integrating low-occurrence triphone models into high-occurrence ones. Experiments conducted using an actual lecture speech corpus showed that the proposed method improves the STD performance with regard to both triphones and demiphones, demonstrating its effectiveness.

Index Terms: Spoken Term Detection, triphone, intensive triphone

1. Introduction

Spoken Term Detection (STD) has become a hot topic in speech processing research in recent years as a result of the drastic increase in the number of spoken documents, such as video content available online. Large Vocabulary Continuous Speech Recognition (LVCSR) systems are used in representative approaches to STD, where searching for query words is performed in the set of recognition results [1]. STD systems must be capable of detecting any query words, including unknown words that are not included in the LVCSR dictionary, since such unknown query words tend to be used as query words [2] [3]. However, unknown words are difficult to detect with such word-based STD systems. Using subword recognition instead of whole-word recognition is a promising approach to detecting unknown words [3] [4]. In this regard, triphones are the most popular acoustic models for LVCSR systems, and they are also used as subword models in STD systems. Thus far, we have investigated various subword models that are suitable for STD systems, and developed demiphone models that are constructed by dividing each triphone model into two demiphone models on a time axis. While the speech recognition performance of triphone acoustic models was higher than demiphone acoustic models, demiphone acoustic models showed comparatively higher performance in our STD experiments than triphone acoustic models [3]. The number of Japanese physical triphone and demiphone models is about 8,000 and 1300, respectively. In our preliminary experiments, some triphone models were found to be associated with few training data that were insufficient for constructing well-trained triphone hidden Markov models (HMMs). Such insufficient triphone models should be reduced and merged with other, acoustically similar, triphone models with sufficient training data. In this paper, we propose the concept of intensive triphone models for STD, which are constructed by clustering triphone models using occurrence information about triphone models in a corpus, where low-occurrence triphone models in training data sets are merged with high-occurrence triphone models. Through evaluation experiments using an actual lecture speech corpus, we investigate the effectiveness of intensive triphone models, and we demonstrate that the proposed method of integrating subword models is effective for STD by comparing the performance using different subword models; demiphone models. Few studies have been conducted on improving the STD performance by reconfiguring subword models such as triphones, although new subword types, such as demiphones [3], graphemes [5] and sub-phonetic segments (SPSs) [6], have been introduced into the field of STD.

The present study pursues a method for improving the STD performance by reconfiguring existing subwords, and this paper introduces evaluation experiments conducted to evaluate the effectiveness of the proposed method. First, we investigate a method for clustering triphone models into intensive triphone models, after which we investigate the appropriate number of intensive triphone models, and eventually we evaluate the performance of the method by using demiphones as a different type of subword.

In Section 2, first we analyze triphone models and describe the proposed STD method using intensive triphones. A discussion of the results of evaluation experiments is provided in Section 3, and a conclusion is presented in Section 4.

2. Proposed method

2.1. Analyzing triphone models

This section analyzes triphone models that are trained by using the Corpus of Spontaneous Japanese (CSJ). CSJ is the most

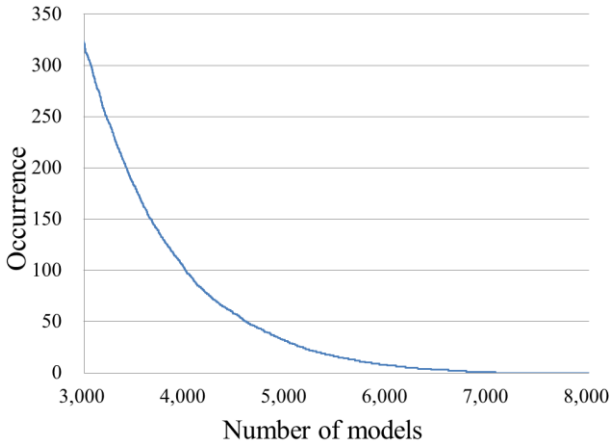

Figure 1: *Occurrence of triphones in CSJ.*

commonly used speech corpus by the speech processing community in Japan. CSJ includes 2702 actual and simulated lecture speeches that amount to a total of 604 h of spoken data, and are regarded as sufficient for training triphone models.

Figure 1 shows triphone occurrences in CSJ, omitting 3000 triphone models that occur more than 320 times. From among about 8,000 triphone models, 3,500 models occur more than 200 times, and the remaining 4,500 models occur less than 200 times. Since there are 32 or 64 Gaussian components per triphone HMM state, the data for 4,000 triphone models are insufficient for training. Therefore, we assume that insufficient models with few training data degrade the performance of STD. Thus, we considered that the STD performance can be improved by reducing the number of such insufficient models and integrating them into sufficient models.

2.2. Reducing the number of models by clustering

Since the phone sequence is inputted first and converted into Japanese characters for any Japanese word, it is easy to obtain the phone sequences for query words. For example, the query "Tokyo" in a Japanese character sequence is automatically transformed into the phone sequence "t o ky o"[1]. Let the output of phone recognition for the utterance "Tokyo" be "t o py o" (i.e., with substitution errors), as shown on the lower left in Figure 2. While the utterance is difficult to retrieve by distinguishing between "t o ky o" and "t o py o", if the "py" model is reduced and integrated into the "ky" model, both sequences become identical to "t o ky o", which allows for retrieving the utterance. In the Figure 2, "ky", "py" and "ty" are integrated into the same category ("ky"), and the "py" label in the recognition results is replaced by "ky", as shown on the lower right in the Figure 2. The "ky" model including "py" and "ty" is referred to as an intensive model.

The STD performance is considered to degrade if the number of subword models is too small, in which there might be an optimal number of intensive subword models. Therefore, we investigated the STD performance with regard to the number of intensive subword models and compared it with that of some other clustering algorithms suitable for STD systems.

2.3. Proposed STD method using intensive triphone models

This section describes the proposed STD method, which uses intensive triphone models. The method for constructing intensive triphone models consists of three steps in addition to clustering, namely training of intensive triphone models, speech recognition using the intensive triphone models, and searching query words. The details of each step are provided below, where, clustering, training and recognition are conducted in advance.

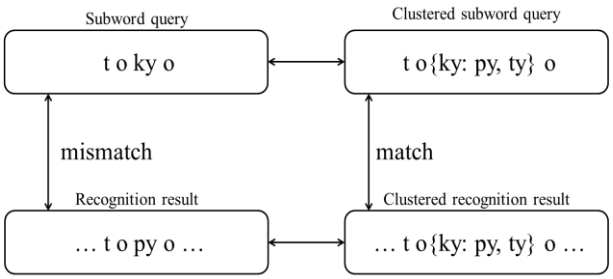

Figure 2 *An example of matching between two phones sequences by clustering them into intensive models.*

[1]Here, "t", "o","ky", "py", and "ty" correspond to Japanese phones

2.3.1. Clustering

About 8,000 triphone models are integrated into *N* intensive triphone models by the clustering method described in detail in the next section. A conversion table representing the transformation from original to intensive triphone models is constructed in this step.

2.3.2. Training

Original triphone models attached to speech data are converted into intensive triphone models according to the conversion table. Then, subword trigram language models and intensive triphone acoustic models are trained using speech data and their intensive triphone labels, and acoustic distances used in the next step are computed. Acoustic distances are a statistic representation of the acoustic dissimilarity between any two intensive triphone models. Each triphone model is composed of an HMM with multiple Gaussian components in a state. Acoustic distances are obtained by comparing these distributions for two intensive triphone models.

2.3.3. Speech recognition using intensive triphone models

Speech recognition is performed using an intensive triphone dictionary, language models of intensive triphone trigram and intensive triphone acoustic models, and intensive triphone sequences for spoken documents are obtained and stored as a result of speech recognition.

2.3.4. Searching

When a query is given, its phone sequence is easily obtained through the input process, as mentioned above. The phone sequence of the query is converted into an original triphone sequence, and some triphone models are replaced by intensive triphone models based on the conversion table. Then, a search is performed for the intensive triphone sequence of the query in the intensive triphone sequences in spoken documents by using Continuous Dynamic Programming (CDP), where the acoustic distances mentioned in the training step are used as local distances in CDP.

2.4. Clustering using occurrence information

In this paper, we propose a method for integration of low-occurrence models into high-occurrence models. The training data of low-occurrence models are insufficient, which leads to errors in recognition. At this step, each low-occurrence model is assigned to the high-occurrence model with the shortest acoustic distance to the low-occurrence one. The process of integrating triphone models is as follows.

(1) Rank triphone models in the order of the occurrences. The top *N* triphone models in terms of occurrence are regarded as intensive triphone models, where *N* is chosen in advance (this is investigated in Section 3.3).

(2) Integrate the approximately 8,000 triphone models into *N* intensive triphone models by assigning each non-intensive triphone model to an intensive triphone model. A non-intensive triphone is assigned to the intensive triphone with the shortest acoustic distance to the non-intensive triphone.

(3) Create a conversion table from non-intensive triphone models to intensive triphone models according to the results of the procedure in (2).

3. Evaluation experiment

3.1. Data set and experimental setup

3.1.1. Training data

Speech data of an even number of lectures from CSJ described in 2.1 were used for training subword acoustic models and subword language models. The training data did not include evaluation data, and training for subword acoustic models and subword language models was conducted using the HTK and Palmkit software tools, respectively.

Table 1. Acoustic analysis conditions.

Sampling	16KHz, 16bit
Feature parameter	38dim MFCC + ΔMFCC + ΔΔMFCC + ΔPOWER + ΔΔPOWER
Analysis windows	Hamming window
Window length	25ms
Frame shift	5ms for demiphone 10ms for triphone

Table 2. *Configuration of the test sets.*

Test set	Spoken documents	Queries
1	Simple sets composed of 50 lecture speeches	50
2	CSJ core data composed of	50
3	177 lecture speeches	50

3.1.2. Evaluation setup

The feature parameters as extracted with HTK are shown in Table 1 together with the conditions for extracting the parameters.

3.1.3. Evaluation data

We prepared three test sets for the evaluation data, and the details of the test sets are shown in Table 2. Query terms are different among three test sets. The decoder used was Julius rev. 4.1.5.1.

3.1.4. Evaluation set up

We used the mean average precision (MAP), which is a standard measurement parameter in the field of STD (MAP was also used as an evaluation measure in the NTCIR-9 workshop). Furthermore, the average precision (AP) for a query is obtained from Eq. (1) by averaging the precisions at every correct occurrence, and MAP is obtained from Eq. (2) as the average of AP for each query *s*, where *T* is the total number of queries. In Eq. (1), *C* and *R* are the total number of correct sections and the lowest rank of the last correct section, respectively. Let to be *1* if the *i*-th candidate section of query *s* is correct and *0* otherwise. Therefore, Eq. (1) averages the precision when a correct section is presented.

$$AP(s) = \frac{1}{C}\sum_{i=1}^{R}\delta_i \times precision(s,t) \quad (1)$$

$$MAP = \frac{1}{T}\sum_{s=1}^{T}AP(s) \quad (2)$$

3.2. Comparison of clustering methods in terms of retrieval performance

The performance of the proposed clustering method was evaluated through comparison with the k-means method for the test set 1, and the results are shown in Table 3. The number of clusters was 3,500, which resulted in high performance. The MAP improved by 0.65% when using the proposed method, whereas in the case of the k-means method it degraded by 9.76% as compared with the original triphone models. In the k-means method, models which should have been clustered into different categories were often clustered into the same category.

Table 3. *Comparison of the MAP according to clustering method.*

Clustering method	MAP (%)
Proposed method	82.64
k-means method	72.23
Original triphone models (8,254)	81.99

Other clustering methods, such as the decision tree method, which is used for clustering triphone models when tying the states of HMMs, were tested with no observable improvement in performance, which confirmed the effectiveness of the proposed method for reducing the number of low-occurrence triphone models.

3.3. Evaluation of the optimal number of intensive triphone models

We investigate the performance according to the number of intensive triphone models (*N*), reducing *8,254* triphone models to *2,000~5,000* intensive triphone models by using the

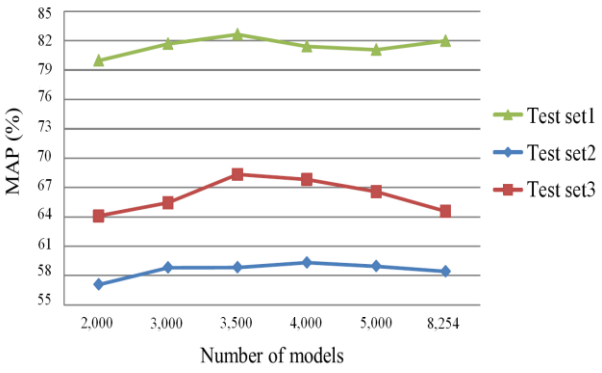

Figure 3: *Comparison of retrieval accuracy for original (N=8254) and proposed triphone models (N<8254).*

proposed method introduced in Section 2. The results for three test sets shown in Figure 3 indicate that the number of models affects the retrieval performance, which was higher than that for the original triphone models in the case of *N=3,500* for all test sets.

Figure 4 shows the cover rate of triphone models in the training data set according to the number of intensive triphone models. Most of triphone models (99.3%) could be covered when using *3,500* intensive triphone models, as shown in the figure, and we believe other triphone modes that are not covered *3,500* intensive triphone models do not have sufficient training data.

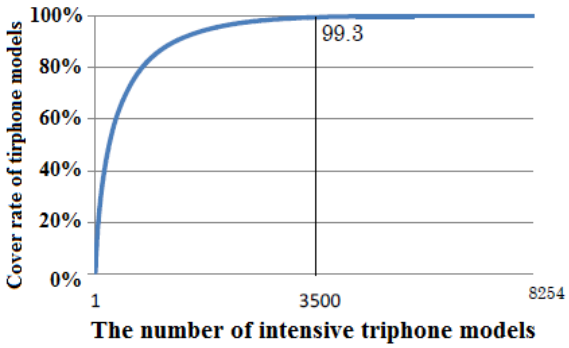

Figure 4: *Cover rate of triphone models according to the number of intensive triphone models*

3.4. Retrieval performance of intensive demiphone

We also evaluated the proposed method by applying it to subwords other than triphones. We used demiphone models, where the number of models was *1,623* and the number of intensive models was varied between *800* and *1,200*. The results are shown in Table 4, where the MAP is improved by 1.68% for *N=1000* as compared with the original demiphone models. The results for three test sets shown in Figure 5 indicate that the number of models affects the retrieval performance. Although the number of demiphone models was smaller than that of intensive triphone models (*N=3,500~4,000*), intensive demiphone models improved the retrieval performance by reducing the number of low-occurrence demiphone models. The results obtained with intensive triphone and demiphone models demonstrate that the integration of low-occurrence models into high-occurrence ones is effective for improving the STD performance. The number of models was reduced by about 40~60%, although the suitable number is not the same for all subword types. This method is applicable to other subwords such as monophone and SPS (Sub-phonetic segments[6]) to improve the retrieval performance.

Table 4. *Comparison of retrieval accuracy for original and proposed demiphone models.*

Number of models	MAP (%)
1,623(original)	76.89
1,200	78.34
1,000	78.57
800	75.53

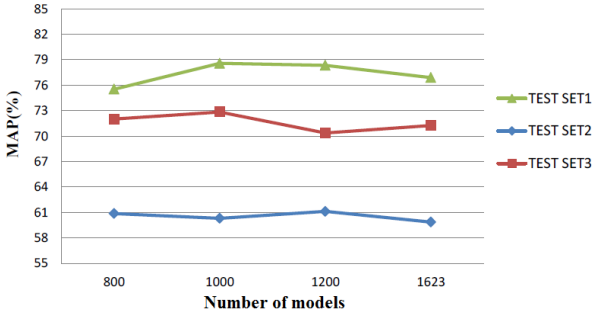

Figure 5: *Comparison of retrieval accuracy for original (N=1623) and proposed demiphone models (N<1623).*

4. Conclusions

In this paper, we proposed the concept of intensive subword models as a means to improve the STD performance, where low-occurrence subword models are integrated into high-occurrence (intensive) subword models. Evaluation experiments showed that intensive subword models improve the STD performance in the case of both triphone and demiphone models, demonstrating the effectiveness of the proposed intensive models. In future work, we plan to apply the proposed method to other subword types, as well as to improve the STD performance even further by combining the results for multiple subword types.

5. Acknowledgements

This research is supported by Grand-in-Aid for Scientific Research (C) Project No. 20500096, KAKENHI of Japan Society for Promotion of Science.

6. References

[1] John S. Garofolo, Cedric G. P. Auzanne and Ellen M. Voorhees, "The TREC spoken document retrieval track: A success story," Ninth TREC, NIST, 2000.

[2] B. Logan, JM. Van Thong, "Confusion-based query expansion for OOV words in spoken document retrieval", Proc.ICSLP, 2002.

[3] K. Iwata, Y. Itoh, K. Kojima, M. Ishigame, K. Tanaka and S. Lee, "An Investigation of New Subword Models and Subword Phonetic Distance for Vocabulary-free Spoken Document Retrieval System", Trans. IPS Japan vol.48, No.5, pp.1990-2000, 2007.

[4] Y. Onodera, Y. Itoh, K. Kojima, M. Ishigame, K. Tanaka, S. Lee, "Spoken term detection by result integration of plural subwords using confidence measure", WESPAC, 2009.

[5] D. Wang, J. Frankel, J. Tejedor, S. King, "A comparison of phone and Grapheme－based spoken term detection", ICASSP Vol.11, pp.4969-4972, 2008.

[6] K. Tanaka, H. Kojima, "A between-word distance calculation in a symbol domain and its applications to speech recognition", ICONIP, pp.11107-1111, 1997.

[7] Y. Itoh, H. Nishizaki, X. Hu, H. Hiroki, T. Akiba, K. Aikawa, T. Kawahara, S. Nakagawa, T. Matsui, Y. Yamashita "Development of Test Collection for Spoken Term Detection - Interim Report -", IPSJ SIG Technical Report, Vol.2009-SLP-78 No.4, 2009.