Spoken Term Detection Results using Plural Subword Models by Estimating Detection Performance for Each Query

Yoshiaki Itoh [1], Kohei Iwata [1], Masaaki Ishigame [1], Kazuyo Tanaka [2], and Shi-wook Lee [3]

[1] Faculty of Software and Information Science, Iwate Prefectural University, Iwate
[2] University of Tsukuba [3] National Institute of AIST

y-itoh@iwate-pu.ac.jp, g231d002@edu.soft.iwate-pu.ac.jp, ishigame@iwate-pu.ac.jp,
ktanaka@slis.tsukuba.ac.jp, s.lee@aist.go.jp

Abstract

The present paper proposes a new integration method of plural spoken term detection (STD) results obtained from plural subword models that we previously proposed. We confirmed that these new subword models, which are the 1/2 phone model, the 1/3 phone model, and the sub-phonetic segment (SPS) model, are effective for STD systems, which must be vocabulary-free in order to process arbitrary query words. In addition, these models are more sophisticated on the time axis than conventional phone models, such as the triphone model. In the present study, we utilize the results of the subword models explicitly when integrating the plural results. For this purpose, we introduce an STD performance index that expresses the degree of detection difficulty for each query word. The index is approximated by the recognition accuracy of the query subword sequence. We demonstrate improved performance through experiments using an actual presentation speech corpus.

Index Terms: information retrieval, spoken term detection, speech recognition, speech processing.

1. Introduction

For spoken document retrieval (SDR) and spoken term detection (STD), there exist some representative approaches based on speech recognition results [3][4]. However, query words (queries) are often special terms that are not included in the dictionary of a general speech recognizer [2]. Therefore, SDR and STD systems must be vocabulary-free in order to process arbitrary query words. We have therefore proposed a vocabulary-free STD system that exploits subword models such as the monophone model, the triphone model, and the newly proposed models. This approach is advantageous because any word can be a query word, and many studies for STD have recently investigated subword models [5][6]. In a previous study [1][2], we proposed two new subword models instead of the triphone model, which is a conventional phone model used in speech recognition. These new models are called the half (1/2) phone model and the one-third (1/3) phone model. The half-phone model is generated by dividing a triphone into two subword models, and the 1/3 phone is generated by dividing a triphone into three subword models. Each subword acoustic model is composed by Hidden Markov Models with the same number of states. Therefore, these subword models are context-dependent models that are more sophisticated on the time axis than the triphone model.

We confirmed that the average STD performances of the 1/2 phone, 1/3 phone, and sub-phonetic segment (SPS) [7] models are better than that of the triphone model. By linearly integrating the plural results obtained from the plural subword models, we confirmed the improved STD performance [2]. The STD performance for each query word, however, does not always correspond to the average performance. For example, the performance of the 1/3 phone model for a certain query word is sometimes better than that of the 1/2 phone model, although the average performance of the 1/2 phone model is better than that of the 1/3 phone model. We can estimate the advantage or disadvantage of each subword model for given query words and utilize the result of the most advantageous subword model explicitly. Therefore, in the present paper, we propose a method of integrating these plural STD results automatically by estimating the STD performance of each subword model for each query to achieve better performance than the linear integration method. The estimated STD performance is denoted by SPI (STD Performance Index), which represents the difficulty of STD for each query. SPI is approximated by the average accuracy of a query subword sequence. We propose this automatic integration method by exploiting the SPI directly for the linear weighting factor.

In the present paper, an outline of the proposed STD system and the integration methods of plural STD results using SPI are explained in detail. The performance of the proposed method is evaluated through experiments with an actual presentation corpus.

2. Proposed STD Method based on Plural Subword Models

2.1. Outline of the Proposed STD System

In the proposed system, subword acoustic models, their language models, a subword distance matrix, and subword recognition results of spoken documents are prepared beforehand [2].

First, subword recognition is performed for all of the spoken documents and a subword sequence database is prepared beforehand. Here, subword language models, such as subword bigrams and trigrams, are used. The system allows both text and speech queries. When a user inputs a text query, the text is automatically converted to a subword sequence according to conversion rules. For speech queries, the system performs subword recognition and transforms the speech into a subword sequence in the same manner as spoken documents. For each subword model, the system then retrieves the target section using Continuous DP algorithms by comparing a query subword sequence to all of the subword sequences in the spoken documents. The local distance refers to the distance matrix that represents the subword similarity and contains the statistical distance between any two subword models. The system outputs plural candidate sections that show a high degree of similarity to the query word for each subword model. Each candidate section has a distance and a sentence number of spoken documents. A new distance is computed by integrating the distances of plural subword models for all sentences, and candidate sections are re-ranked.

In the following section, the proposed integration method is described in detail after briefly explaining the subword models used in the present paper.

2.2. Subword Models

In addition to triphone models and monophone models, 1/2 phone models, 1/3 phone models and sub-phonetic segment (SPS) [7] models are used for subword models in this paper. The 1/2 and 1/3 phone models have been proposed for STD in [1][2]. Figure 1 represents each subword expression and the conceptions of the subword boundary of the three-phone sequence “h a t”. Each triphone model is divided into two 1/2 phone models: a model of the front part and a model of the rear part, as shown in Figure 1. A triphone model is divided into three 1/3 phone models.

Since these models, including the SPS models, were confirmed to have better STD performance than triphone models [1], these three models are used for plural subword integration. These models are context-dependent models, such as triphone models, and are more sophisticated models on the time axis than triphone models.

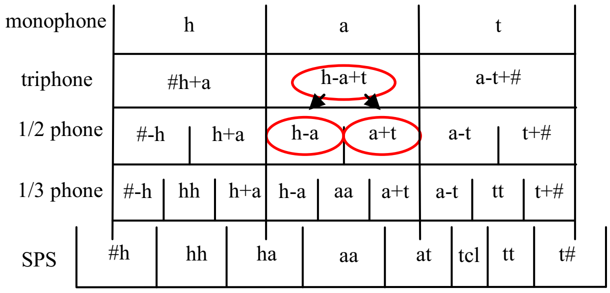

Figure 1: *Subword models and “h a t” expressions.*

2.3. Integration of Plural Results using STD Performance Index (SPI) for a Query

The average performance of the 1/2 model was better than that of the triphone model in our previous results. The triphone models, however, sometimes had a better performance than the 1/2 model for some queries when the performance is evaluated for each query. In the same way, the performances of the 1/2 phone, 1/3 phone, and SPS models for a query are not always the same as the average performance. Therefore, we propose a method that integrates plural STD results obtained from these subword models to improve the STD performance.

Each subword model m ($1 \leq m \leq M$) generates the distance $D_m(i, j)$ between a spoken document or speech section S_i ($1 \leq i \leq I$) and a query Q_i ($1 \leq j \leq J$). Here, M, I, and J denote the number of subword models, the number of documents, and the number of queries, respectively. To integrate the STD results from plural subword models, these plural distances are simply combined linearly. This modified distance $D_M(i, j)$, which is a new criteria, is obtained by integrating the distances $D_m(i, j)$, according to the following equation:

$$D_M(i,j) = \sum_{m=1}^{M} \alpha_m \cdot D_m(i,j) \tag{1}$$

$$\sum_{m=1}^{M} \alpha_m = 1 \tag{2}$$

where, α_m is a weighting factor for the m-th subword model.

The following is an explanation of three methods by which to determine the weighting factors of Eq. (1). Here, the STD performance index (SPI) for each query is introduced in the latter two methods. Because the STD performance depends on the query, as mentioned above, suitable subword models are thought to exist for a given query. A good STD performance can be expected if a query is recognized at a high accuracy. We, therefore, assume that STD performance depends on the difficulty in recognizing the given query words in spoken documents. We approximate the *SPI(m, q)* by the average recognition accuracy of the query subword sequence in the m-th subword model for a query q, which is obtained by Eq. (3). In this equation, a query q is composed of $N_{m,q}$ subwords, and $accuracy(s^{m,q}_k)$ represents the recognition accuracy of the k-th subword in the query q. *SPI(m, q)* is normalized by the average accuracy for each subword model, represented by *ave_accuracy(m)* in Eq. (4). These accuracy values are obtained from other data sets.

$$SPI(m,q) = \sqrt[Nm,q]{\prod_{k=1}^{Nm,q} accuracy(s_k^{m,q})} \tag{3}$$

$$nSPI(m,q) = \frac{SPI(m,q)}{ave\_accuracy(m)} \tag{4}$$

(1) Simple linear integration

All of the weighting factors α_m are given beforehand, and the distances are combined linearly according to Eqs. (1) and (2).

(2) Larger weighting for large *nSPI*

Given a query word *q*, *nSPI (m, q)* is computed for each subword model *m*. This method provides a larger weighting factor to the subword model that has a large *nSPI(m, q)* value. The value of the weighting factor α_m is also given beforehand. For example, when we use two subword models, α_1 and α_2 ($\alpha_1 > \alpha_2$) are given beforehand, and the larger α_1 is provided to the subword model that shows the large *nSPI* value.

(3) Automatic weighting by *nSPI*

This method exploits the *nSPI* for the weighting factors α_m directly and does not require the adjustment of the weighting factors. We simply define the weighting factor as the *nSPI(m, q)* proportion among all *nSPI(m, q)*, shown in the following equation:

$$\alpha_m = \frac{\boldsymbol{nSPI(m,q)}}{\sum_{j=1}^{M} \boldsymbol{nSPI(j,q)}} \qquad (5)$$

3. Evaluation Experiments

3.1. Experimental Conditions

The conditions for feature extraction are listed in Table 1. We constructed five subword acoustic models and subword language models, which are the monophone, triphone, 1/2 phone, 1/3 phone, and SPS acoustic and language models. All of the acoustic and language models were trained by the JNAS [8] database. There were 43 monophones, approximately 8,000 triphones, 1,300 1/2 phones, 1,400 1/3 phones, and 400 SPSs.

We changed the weighting factor α_m at every *0.1* from *0.0* to *1.0* in Eqs. (1) and (2) for the simple linear integration. We evaluated the performance for all combinations of the weighting factor α_m.

3.2. Test Data and Evaluation Measurement

The test data in the experiments were an actual presentation corpus of CSJ [9]. The test data included 49 presentation speeches that total approximately twelve hours. Each presentation is spoken by a different speaker. We used 50 query words and each query has three to 50 corresponding sections in the test data. This data set is provided by the SIG-SLP (Special Interest Group – Spoken Language Processing) of Information Processing Society of Japan [10] that is constructing Japanese standard SDR and STD test collections. We used the average precision rate for evaluation measurements [2].

3.3. Results and Discussion

Table 2 shows the recognition rates for each subword model. The recognition rate depends on the number of subword models, the perplexity of the subword language models, and the redundancy in the time axis [1]. The table also shows the STD performance using each subword model. These results indicate that the performance of the triphone model is not better than those of the other models. This is because the triphone has a number of models, and we assume that the triphone language model was not trained sufficiently. The performance can be improved by using a number of training data sets. The performances of the proposed models and the SPS model were better than those of the phone models, as shown in Table 3. The correct rate was high when extracting the top candidate by the 1/2 phone and SPS models.

Table 1. Conditions for feature extraction

Sampling	16 kHz 16 bit
Feature parameter	12-dim. MFCC+ 12-dim. Δ MFCC+ Energy
Window length	16 ms.
Frame shift	10 ms. for monophone and triphone 5 ms. for 1/2 phone, 1/3 phone, SPS

Table 2. Basic data and results for each subword model

Subword model	# of models	perplexity	recognition rate	Average precision
monophone	43	7.91	73.5	38.60
triphone	7,956	4.73	55.9	34.04
1/2 phone	1,333	2.97	65.1	67.49
1/3 phone	1,374	2.02	70.3	53.49
SPS	423	2.65	77.8	60.28

Table 3. Results of STD performance for the simple linear integration method

M	Weighting factor					Average precision
1	1/2					67.49
	1.0					
2	1/2	SPS				70.78
	0.6	0.4				
3	1/2	SPS	1/3			71.66
	0.5	0.3	0.2			
4	1/2	1/3	SPS	monof		71.76
	0.5	0.2	0.2	0.1		
5	1/2	1/3	SPS	monof	tri	71.76
	0.5	0.2	0.2	0.1	0	

Table 3 shows the best STD performance among all combinations of the weighting factor α_m, using plural subword models based on the first method of simple linear integration. *M* denotes the number of subword models. In the case of *M = 1*, the best performance was obtained using the 1/2 phone, as shown in Table 2. In the case of *M = 2, t*he best performance was obtained using the 1/2 phone and the SPS models at weighting factors of 0.6 and 0.4, respectively, and the performance was improved by 3%. In the case of *M = 3*, the best performance was obtained using the 1/2 phone SPS and the 1/3 phone models at weighting factors of *0.5*, *0.3*, and *0.2*, respectively, and the performance was improved by approximately 1%.

When using four subword models at *M = 4*, the best performance was obtained by adding the monophone

Table 4. Results of the best STD performance for larger weighting for large SPI

M	Subword models for the best performance	The best weighting factor combination					Ave. Precision
2	1/2, SPS	0.7	0.3				73.35
3	1/2, 1/3, SPS	0.7	0.2	0.1			75.03
4	monof, 1/2 1/3, SPS	0.7	0.2	0.1	0.0		74.89
5	monof, tri 1/2, 1/3, SPS	0.4	0.4	0.1	0.1	0.0	72.48

Table 5. Results of STD performance for automatic weighting by SPI

M	Subword models					Ave. Precisio
2	1/2	SPS				72.92
3	1/2	SPS	1/3			73.12
4	1/2	SPS	1/3	monof		72.86
5	1/2	SPS	1/3	monof	tri	70.95

model to the three subword models at $M = 3$. The performance improvement was small. In the case of $M = 5$, the performance was the same as that in the case of $M = 4$ because the weighting factor of the triphone was *0.0*, and the result of the triphone was not utilized.

We confirmed the improvement of the STD performance by integrating the results from plural subword models. The weighting factor shown in Table 3 illustrates that large values should be given to the subword models that show better STD performance.

Table 4 shows the best STD performance and the weighting factors based on the second method of larger weighting for large *SPI*. In the case of $M = 2$, the best performance was obtained using the 1/2 phone and the SPS models at weighting factors of 0.7 and 0.3, respectively. When a query was given in this case, SPI values were computed for both models, and the weighting factor 0.7 was provided for the subword model, the SPI value of which was larger. SPS sometimes receives the weighting factor of 0.7. The performance was improved by 6% compared with that using the single model and by 3% compared with that of simple linear integration in the case of $M = 2$. In the case of $M = 2$ and $\alpha_1 = 1.0$ to $\alpha_2 = 0.0$, where only the subword model showing larger SPI was used, the average precision became 71.77, which is superior to the best performance of simple linear integration in the case of $M = 2$. In the cases of $M = 3$, 4, and 5, the performance using SPI was better than that of simple linear integration. The introduction of SPI was confirmed to work effectively.

Table 5 shows the STD performance based on the third method of automatic weighting by SPI. In the case of $M = 2$, without setting the weighting factors, the best performance was comparable to the best performance among 100 weighting factor combinations by the second method in table 4. The performance was improved by 6% compared with that using the single model in the case of $M = 3$. In this case, the performance does not reach the best performance of the second method, which is, however, the best result among 660 weighting factor combinations in table 4. In the cases of $M = 4$, and 5, the performance declined because the method had to use the subword models that showed low SPI. These results indicate the method can be improved by training a weighting method in the future.

4. Conclusions

The present paper proposed a method of integrating plural STD results that are obtained from plural subword models. The STD performance index (SPI) was introduced to improve the STD performance. The SPI expresses the difficulty of STD for each query word and is exploited for a weighting factor directly when integrating plural distances. We demonstrated experimentally that the performance could be improved by integrating plural subword models. The introduction of the SPI enabled the STD performance to be improved without setting weighting factors. In the future, the average accuracy among subword models should be used for the weighting factors.

5. Acknowledgements

This research is supported by Grand-in-Aid for Scientific Research (C) Project No. 20500096 and (C) Project No. 22500145, Japan Society for Promotion of Science.

6. References

[1] Iwata, K., Itoh, Y., Kojima, K., Ishigame, M., Tanaka, K. and Lee, S., “Open-Vocabulary Spoken Document Retrieval based on new subword models and subword phonetic similarity,” INTERSPEECH, 2006.

[2] Itoh, Y., Iwata, K., Kojima, K., Ishigame, M., Tanaka, K. and Lee, S., “An Integration Method of Retrieval Results using Plural Subword Models for Vocabulary-free Spoken Document Retrieval,” INTERSPEECH, 2007.

[3] Garofolo J. S., Auzanne C., Voorhees E M., “The TREC Spoken Document Retrieval Track: A Success Story,” Recherche d'Informations Assiste par Ordinateur, 2000.

[4] Auzanne C., Garofolo J. S., Fiscus J. G., Fisher W. M., “Automatic Language Model Adaptation for Spoken Document Retrieval,” B1, 2000TREC-9 SDR Track, 2000.

[5] Moreau N., Kim H., and Sikora T., “Phonetic confusion matrix based spoken document retrieval,” INTERSPEECH, 2004.

[6] Hori, T., Hetherington, L., Hazen, T. and Glass, J., “Open-vocabulary spoken utterance retrieval using confusion networks,” ICASSP 2007.

[7] Tanaka, K., Kojima H., "Speech recognition method with a language-independent intermediate phonetic code", ICSLP, Vol. IV, pp.191-194, 2000.

[8] Itou K., “JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research,” J. Acoust. Soc. Jpn. (E), Vol. 20-3, pp.199-2006, 1999.

[9] Maekawa, K., “Corpus of Spontaneous Japanese: Its design and evaluation.” SSPR, 2003.

[10] Akiba, Tomoyosi et al., “Developing an SDR test collection from Japanese lecture audio data,” APSIPA, 2009.