Designing text corpus using phone-error distribution for acoustic modeling

Hiroko Murakami #1, Koichi Shinoda #2, Sadaoki Furui #3

# *Dept. Computer Science, Tokyo Institute of Technology*
2-12-1 Ookayama, Meguro-ku, 152-8552, Japan
1 murakami@ks.cs.titech.jp
2 shinoda@cs.titech.ac.jp
3 furui@cs.titech.ac.jp

***Abstract*—It is expensive to prepare a sufficient amount of training data for acoustic modeling for developing large vocabulary continuous speech recognition systems. This is a serious problem especially for resource-deficient languages. We propose an active learning method that effectively reduces the amount of training data without any degradation in recognition performance. It is used to design a text corpus for read speech collection. It first estimates phone-error distribution using a small amount of fully transcribed speech data. Second, it constructs a sentence set whose phone-occurrence distribution is close to the phone-error distribution and collects its speech data. It then extends this process to diphones and triphones and collects more speech data. We evaluated our method with simulation experiments using the Corpus of Spontaneous Japanese. It required only 76 h of speech data to achieve word accuracy of 74.7%, while the conventional training method required 152 h of data to achieve the same rate.**

I. Introduction

The development of large vocabulary continuous speech recognition (LVCSR) systems requires a large amount of speech data with transcription for acoustic model training. More than 100 hours of data are needed to have sufficient recognition accuracy, but collecting such a large speech database is very expensive. This is a serious problem, especially when developing an LVCSR system for resource-deficient languages, because their markets may be too small to afford such high costs. We develop an active learning method that reduces the costs by reducing the size of training data for acoustic modeling without any degradation in recognition performance. In other words, we aim to construct a training data set that exhibits higher recognition accuracy than other sets with the same size. There are two approaches to construct a large speech database. One is to transcribe unlabeled speech data and the other is to collect read speech data of provided texts.

In the former approach to transcribe unlabeled speech data, active learning has been extensively studied [1], [2], [3], [4], [5]. In most of these studies, it has been used to select utterances from speech data for acoustic model training to decrease the annotation costs. The focus of these studies has been to provide an effective uncertainty measure for each utterance; those utterances whose transcriptions seem to be highly uncertain are preferred as training data.

There have been also several studies for the latter approach to provide a good text corpus to collect read speech data [6], [7], [8], [9]. Shen *et al.* [7] proposed to design a phonetically balanced sentence set. While this approach is useful to avoid the data sparseness problem, it does not directly increase the recognition performance. Huo *et al.* [9] selected vocabulary consisting of words expected to be highly confusable in a given task. This method is indeed effective, but may not be significantly effective in general LVCSR. We focus on this latter approach by constructing a sentence set that involves more confusable *recognition units* (such as triphones) to construct an acoustic model for LVCSR.

Recently, we developed a speaker adaptation techniques using two-step active learning [10]. This method is based on the assumption that the recognition accuracies of confusable recognition units can be improved by increasing adaptation data for such units. In this method, the initial adaptation data is first collected to obtain a phone error distribution. Then, in the second step, those sentences whose phone distributions are close to the error distribution are selected, and their utterances are collected as the additional adaptation data. In our evaluation, we found that this method was significantly better than a method using randomly chosen sentences for adaptation. We expect that this approach is also effective for acoustic model training.

In this paper, we propose an active learning method for collecting read speech data for acoustic modeling. As in the successful speaker adaptation method [10], this method designs a sentence set including more recognition units with low recognition accuracy than other units. It first estimates the phone-error distribution using a small amount of fully transcribed speech data and constructs a sentence set whose phone-occurrence distribution is close to the phone-error distribution. It extends this process to diphones and triphones and collects more speech data. Since it is difficult to evaluate this active learning method online, we conducted simulation experiments using a fully transcribed database, Corpus of Spontaneous Japanese (CSJ), with which we constructed the sentence set by selecting sentences from its transcribed texts.

Additionally, in order to apply our method to the former approach for data collection, in which utterances are selected from untranscribed speech data, we examine a semi-supervised

978-1-4673-0367-5/11/$26.00 ©2011 IEEE

utterance selection framework, where the hypothesis transcription obtained from automatic speech recognition is used instead of manual transcription. We also report the results of its evaluation.

This paper is organized as follows. Section 2 outlines our active learning method, and Section 3 explains the sentence selection algorithm. Section 4 explains our semi-supervised utterance selection framework. Section 5 reports our evaluation experiments using CSJ, and Section 6 concludes the paper.

II. Active learning scheme

Our method can be used both for *selecting* sentences from a text corpus and *generating* sentences from scratch. For simplicity, we explain our method for sentence selection. Figure 1 is a flowchart of our method.

First, we prepare a small amount of fully transcribed speech data to estimate the distribution of error occurrences. Half of the data, Data A, are used for training an initial acoustic model and the other half, Data B, are used for measuring the phone recognition accuracy. Let U be a set of phones. The phone-error distribution $P(u)$ over phones $u \in U$ is defined as:

$$P(u) = \frac{r(u)}{\sum_{u \in U} r(u)}, \qquad (1)$$

where $r(u)$ is the number of recognition errors for phone u. We count not only the number of u being misrecognized as another unit, but also that of the other units being misrecognized as u.

Then, from a large text corpus prepared beforehand, we select those sentences whose distribution of phone occurrences is close to the phone-error distribution P. The phone-occurrence distribution Q_T of a sentence set T is defined as:

$$Q_T(u) = \frac{c_T(u)}{\sum_{u \in U} c_T(u)}, \qquad (2)$$

where $c_T(u)$ is the number of occurrences of phone u in T. Kullback-Leibler divergence (KLD) [8] between them, $D(P||Q_T)$, is used as a distance measure. We will explain the sentence selection procedure in the next section.

We collect the read speech data for the selected sentences and use them as additional training data. We apply this method first to monophones, and then, repeat the same process for diphones and triphones to collect more data.

III. Sentence selection algorithm

We employ a suboptimal greedy algorithm for the sentence selection. Let S be a sentence set of the prepared text corpus, T be a set of selected sentences. Initially T consists of the transcribed texts of the utterances in Data A and B. For every sentence s in S, we calculate $D(P||Q_{T\cup\{s\}})$, KLD between P and the phone-occurrence distribution $Q_{T\cup\{s\}}$ of the set $T \cup \{s\}$.

$$D(P||Q_{T\cup\{s\}}) = \sum_{u \in U} P(u) \log \frac{P(u)}{Q_{T\cup\{s\}}(u)}. \qquad (3)$$

Then, we select the sentence whose KLD is the smallest, and move it from S to T. We repeat this selection process until when $D(P||Q_T)$ stop decreasing.

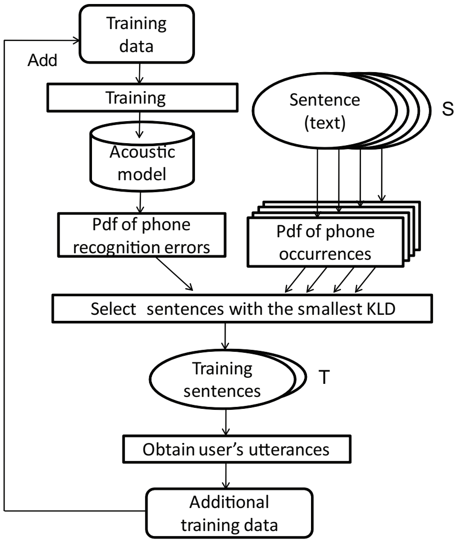

Fig. 1. Flow of proposed method

Since the number of diphones and triphones are large, it takes relatively high computational costs to calculate KLD in Eq. (3) for every sentence s in S. To reduce the costs, we approximate the difference Δ between the present KLD $D(P||Q_{T\cup\{s\}})$ and the KLD $D(P||Q_T)$ in the previous step by using Taylor expansion, utilizing the fact that the total number of occurrences of all phones in $\{s\}$, $M_{\{s\}}$, is much smaller than that in T, M_T:

$$\begin{aligned} \Delta &= D(P||Q_{T\cup\{s\}}) - D(P||Q_T) \\ &\sim \frac{M_{\{s\}}}{M_T}\left(1 - \sum_{u \in U} P(u)\frac{Q_{\{s\}}(u)}{Q_T(u)}\right), \end{aligned} \qquad (4)$$

where $Q_{\{s\}}$ is the phone-occurrence distribution in sentence s.

In the sentence selection, we ignore phones (monophones, diphones, triphones) that rarely appear since their effect on the overall recognition accuracy is very small. We use the set of phones U, each phone of which occurs over a threshold δ in the original S.

IV. Selection from untranscribed speech data

In the previous two sections, we have explained our active learning method for designing a text corpus to collect read speech data. As explained in Section 1, there is an alternative active learning scheme for data collection, in which we select speech utterances from untranscribed speech data and transcribe them. This scheme is more suitable to collect data of spontaneous speech, such as conversational speech. Here we explain our semi-supervised method in this scheme.

Basically, we apply the similar approach as in the previous sections. The difference is that the unlabeled speech data are used as the training data. We obtain their hypothesis transcription by recognizing them with a triphone acoustic model using the training data available. Since it is desirable that the accuracy of these hypothesis transcriptions be as high as possible, we use the phoneme sequences obtained from LVCSR as the hypothesis transcription. Then, we select utterances using the same algorithm, as discussed in Sections 2 and 3.

V. Experiment

A. Experimental conditions

We evaluated our method with a simulation experiment using lecture-speech data obtained from male speakers in CSJ [12]. We used 198,807 utterances (152 h) from 666 speakers as training data, and 2328 utterances (1.95 h) from ten speakers as test data. We randomly selected 10 h (13,028 utterances) of data from the training data and half were used as Data A, and the rest were used as Data B. The other data from the training data (185,779 utterances, 142 h) were used as a text corpus S.

In this simulation experiment, we assumed that the speech data corresponding to the text corpus S were not available at the beginning of the active learning process. Every time our method selected a sentence s from S, it actually retrieved the speech data corresponding to s in CSJ, instead of recording read speech for s.

The frame period for speech analysis was 10 ms and the frame width was 25 ms. The speech feature vector was 39 dimensional, consisting of 12-order mel-frequency cepstral coefficients (MFCCs) appended with energy, delta, and delta-delta coefficients. We applied cepstral mean subtraction to all utterances.

We set the threshold δ described in Section 2 to 10000. There were 37 recognition units for monophones, 211 for diphones, and 521 for triphones. We used the left diphones as the diphones.

We used monophone hidden Markov models (HMMs) with three states in phone recognition to estimate phone-error distribution P. There were 64 mixtures. We used concatenated phone recognition using a grammar representing the Japanese syllable structure.

For evaluating recognition accuracy, we used triphone HMMs with 3000 states, each of which had a Gaussian-mixture probability density function. There were 16 mixtures in each state. We applied a two-pass search for speech recognition. A 2-gram language model was used in the first pass and a 4-gram language model was used in the second. A language model was trained with all the training data. We used word accuracies as the evaluation measures. Hidden Markov Model Toolkit (HTK) [13] was used in the experiment.

We compared our method with a random selection method, with which the training sentences are randomly selected from the text corpus, and with a phonetically-balanced selection method [6].

B. Comparison with other methods

Figure 2 shows the recognition results. We compared the proposed, random selection, and phonetically-balanced methods. We tested the random selection method three times with different seeds. Their averages are shown in Figure 2. Our proposed method performed significantly better than the other two methods. To achieve a word accuracy of 74.7%, the proposed method required only 76 h of data while the other methods required 152 h. It performed better at the termination points of selection using monophones and diphones. The accuracy of the phonetically-balanced method was almost the same as that with random selection method. The phonetically-balanced method is effective when the number of phones with low occurrence in training data is not enough. However, in our situation, the amount of training data is large and such phones occur enough in training data. Because of this, the phonetically-balanced method is not effective.

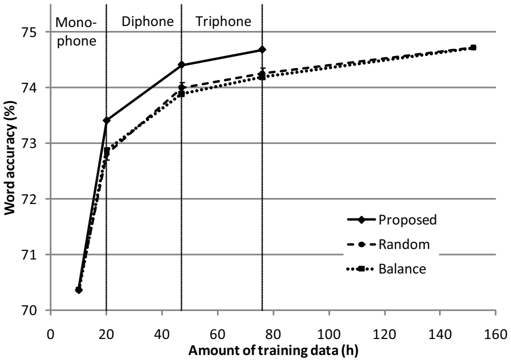

Fig. 2. Comparison of proposed method with random selection (Random) and phonetically-balanced (Balance) methods.

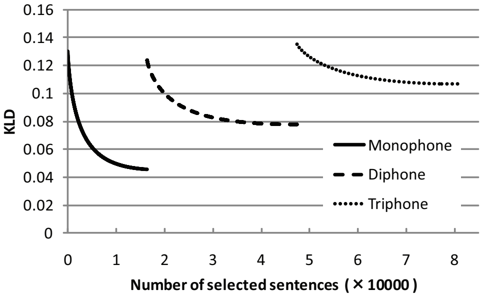

Fig. 3. Change in KLD values between the phone-error distribution and the accumulated phone-occurrence distribution in accordance with number of selected sentences.

C. KLD values

Figure 3 shows the change in KLD values between P and Q in accordance with the increase in the number of selected sentences. By changing the recognition units from monophones to diphones and triphones, the reduction rate of KLD values decreases and the number of the selected sentences increases. The final KLD value for each recognition unit class increases as the number of the recognition units increases. Accordingly, it becomes more difficult to achieve Q closer to P.

TABLE I

Comparison of proposed method using approximation. Org indicates results without approximation, App indicates results obtained from Eq. (4). The table shows recognition accuracy and time required for sentence selection. An Intel(R) Xeon(R) CPU (E5540, 2.53GHz) was used for calculation. Its memory size was 24.7 GB.

	Diphone		Triphone	
	Org	App	Org	App
Accuracy(%)	74.3	74.2	74.6	74.7
Time(h)	4.0	1.8	5.3	3.2

D. Approximation

Table 1 lists the results of the approximation using the Taylor expansion. The accuracies of the proposed method using approximation were almost the same as those without approximation. We reduced the computational time for sentence selection by 55% for diphones and by 44% for triphones. For comparison, we show the time required for the other computation processes, training an acoustic model, recognizing Data B. For diphones, 6.5 h is required for training, 1.0 h for recognition. Therefore, our method reduced the total computational costs by 16%. For triphones, 12.5 h is required for training, 1.0 h for recognition. Our method reduced the total computational costs by 9%. It should be noted that collecting speech data also requires large costs.

E. Selection from untranscribed data

Figure 4 compares the recognition results of our semi-supervised learning method for selecting utterances from untranscribed data, explained in Section 4. We compared it with the random selection method, and the phonetically-balanced method. We also showed our supervised learning method where we assume correct transcriptions were given (oracle). When the amount of training data was 76 h (half of all data), the accuracies were 74.5% for our semi-supervised training framework, 74.7% for our supervised training framework, 74.3% for the random selection method, and 74.2% for the phonetically-balanced method. While the accuracy of our method was slightly higher than those of the other two methods, it was lower than that in our supervised training framework. This is because we used erroneous recognition results as the transcription for the training utterances and used them in selection. Some phones with low recognition accuracies may not have often appeared in the hypothesis transcription. It should be noted that the language model we used was trained using the transcribed text provided, which was not available in real situation.

VI. Conclusion and future work

We have proposed an active learning method for constructing a training data set for acoustic modeling. It generates a text corpus for read speech data, whose occurrence distribution of recognition units is expected to be close to their error distribution. We used KLD as a distance measure between distributions. We evaluated our method with simulation experiments using CSJ. Texts for 76-hour training data were selected with our method, which achieved recognition accuracy of 74.7%, while the conventional training methods required 152 h to achieve the same accuracy. We also proved that our method can be applicable to a semi-supervised training framework using untranscribed speech data, where hypothesis transcription obtained by speech decoder was used.

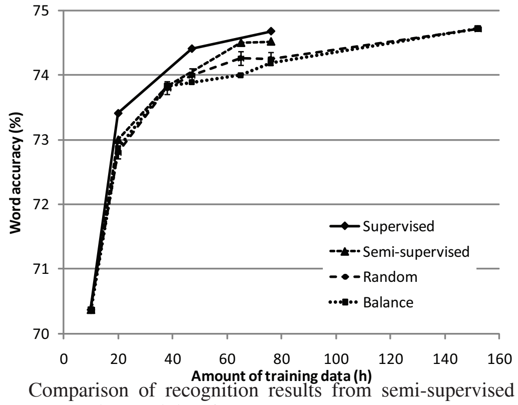

Fig. 4. Comparison of recognition results from semi-supervised training framework, supervised training framework and two methods.

In the future, we first have to conduct further investigations to achieve significant effectiveness in our semi-supervised training framework. We believe it should be combined with conventional active learning methods for untranscribed speech data. While we used maximum-likelihood estimation for acoustic model training in our evaluation, the combination of our method and discriminative training methods is also expected to yield higher recognition accuracies. We would like to implement them to our framework.

In our evaluation discussed above, we used a selection method with which we selected sentences from a text corpus prepared beforehand. In the future, we will apply our method in a more realistic and general situation in which we generate texts whose corresponding speech data is expected to be effective in reducing errors. We plan to construct an online training system for this purpose. We also plan to extend our method to recognition units with longer contexts such as words.

References

[1] T. M. Kamm and G. G. L. Meyer, “Robustness aspects of active learning for acoustic modeling,” Proc. ICSLP2004, pp. 1095-1098, 2004.

[2] H.-K. Kuo and V. Goel, “Active learning with minimum expected error for spoken language processing,” Proc. INTERSPEECH2005, pp. 437-440, 2005.

[3] D. Hakkani-Tur, G. Riccardi, and G. Tur, “An active approach to spoken language processing,” ACM Trans. Speech and Language Processing, vol. 3, no. 3, pp. 1-31, 2006.

[4] B. Varadarajan, D. Yu, L. Deng, and A. Acero, “Maximizing global entropy reduction for active learning in speech recognition,” Proc. ICASSP2009, pp. 4721-4724, 2009.

[5] Y. Hamanaka, K. Shinoda, S. Furui, T. Emori and T. Koshinaka ”Speech modeling based on committee-based active learning” Proc. ICASSP2010, SP-L8.1, 2010.

[6] K. Iso, T. Watanabe and H. Kuwabara, “Design of a Japanese sentence list for a speech database,” Proc. Acoust. Soc. Japan (March), vol. 1, pp. 89-90, 1988.

[7] J.-L. Shen, H.-M. Wang, R.-Y. Lyu, and L.-S. Lee, “Automatic selection of phonetically distributed sentence sets for speaker adaptation with application to large vocabulary Mandarin speech recognition,” Computer Speech and Language, vol. 13, pp. 79-97, 1999.

[8] X. Cui and A. Alwan, “Efficient adaptation text design based on the Kullback-Leibler measure,” Proc. ICASSP2002, pp. I-613-616, 2002.

[9] Q. Huo and W. Li, “An active approach to speaker and task adaptation based on automatic analysis of vocabulary confusability,” Proc. INTERSPEECH2007, pp. 1569-1572, 2007.

[10] H. Murakami, K. Shinoda, and S. Furui ”Speaker adaptation based on two-step active learning” INTERSPEECH2009, pp. 576-579, 2009.

[11] S. Kullback, and R. A. Leibler, J. B. MacQueen, “On information and sufficiency,” Annals of Mathematical Statistics, vol. 22, no. 1, pp. 79-86, 1951.

[12] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, “Spontaneous speech corpus of Japanese,” Proc. LREC, vol. 2, pp. 947-952, 2000.

[13] Hidden Markov Model Toolkit (HTK), http://htk.eng.cam.ac.uk/