論文 / 著書情報
Article / Book Information

Title	Active Learning Using Phone-Error Distribution for Speech Modeling
Authors	Hiroko MURAKAMI, Koichi SHINODA, Sadaoki FURUI
出典 / Citation	IEICE TRANS. INF. & SYST, Vol. E95-D, No. 10, pp. 2486-2494
発行日 / Pub. date	2012, 10
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2012 Institute of Electronics, Information and Communication Engineers.

IEICE TRANS. INF. & SYST., VOL.E95–D, NO.10 OCTOBER 2012

PAPER

Active Learning Using Phone-Error Distribution for Speech Modeling

Hiroko MURAKAMI[†*], *Nonmember*, **Koichi SHINODA**[†a)], *Senior Member*, *and* **Sadaoki FURUI**[†], *Fellow*

SUMMARY We propose an active learning framework for speech recognition that reduces the amount of data required for acoustic modeling. This framework consists of two steps. We first obtain a phone-error distribution using an acoustic model estimated from transcribed speech data. Then, from a text corpus we select a sentence whose phone-occurrence distribution is close to the phone-error distribution and collect its speech data. We repeat this process to increase the amount of transcribed speech data. We applied this framework to speaker adaptation and acoustic model training. Our evaluation results showed that it significantly reduced the amount of transcribed data while maintaining the same level of accuracy.

key words: *active learning, speaker adaptation, acoustic modeling, phone error distribution, Kullback-Leibler divergence*

1. Introduction

Statistical methods such as hidden Markov models (HMMs) have been successfully applied to speech recognition. A large amount of transcribed speech data is usually provided for model estimation to achieve sufficiently high recognition accuracy. However, it is costly to collect such a large amount of data. Many studies have been done with the objective of reducing the amount of transcribed data while maintaining the same level of accuracy. There are two major approaches, unsupervised learning and active learning. Unsupervised learning effectively uses speech data without transcription, whereas active learning selects speech data to be transcribed.

Active learning has been extensively studied for acoustic modeling in speech recognition [1]–[6]. In most of these studies, it has been used to select utterances from untranscribed speech data. Their focus has been on finding an effective uncertainty measure for each utterance; those utterances whose transcriptions seem to be highly uncertain are preferred as training data. Several methods have used active learning in a different way, where they first select a sentence set from a text corpus and collect its read-speech data [7]–[10]. We take this latter approach in this paper.

The key problem with this approach is finding good criteria for sentence selection. Iso et al. [7] proposed designing a phonetically balanced sentence set, which employs a maximum entropy criterion for selecting sentences. While this approach is useful to avoid the data sparseness problem, it does not directly increase the recognition performance. Huo et al. [10] selected vocabulary consisting of words that are expected to be highly confusable in a given task. This method is indeed effective, but may not be significantly effective in general large vocabulary continuous speech recognition (LVCSR). We therefore need sentence selection criteria that directly relate to error reduction and that can be used for general large vocabulary speech recognition.

In this study, we try to improve the overall recognition accuracy by improving the acoustic models of phones having relatively high recognition errors. Assuming that the more speech data for training, the better the model becomes, we collect a training sentence set having phone-occurrence distribution which is similar to the error distribution among phones.

We propose a novel active learning framework for acoustic modeling in speech recognition. It consists of two steps. We first obtain a phone-error distribution using an acoustic model estimated from transcribed speech data. Then, from a text corpus prepared beforehand, we select a sentence whose phone-occurrence distribution is close to the phone-error distribution, and collect its speech data. We use Kullback-Leibler divergence (KLD) [11] as the distance measure between the two distributions. We repeat this process to increase the amount of transcribed speech data. We apply this framework to two tasks, speaker adaptation [12] and acoustic model training for LVCSR [13].

Speaker adaptation techniques (e.g., [14], [15]) that involve using a small number of utterances from users to improve speech recognition performance are often used in many applications. These techniques fall into two categories: supervised adaptation and unsupervised adaptation. In supervised adaptation, users are asked to speak sentences prepared beforehand. Our focus is hence on how to design an adaptation sentence set for each speaker in supervised adaptation. Each speaker has different acoustic characteristics; for example, the phones with low recognition accuracies vary from user to user. Collecting utterances rich in those phones is expected to be an effective way to improve adaptation performance or to reduce the amount of adaptation data while maintaining the same level of recognition accuracy. We evaluated this method in Japanese phone recognition.

The development of LVCSR systems requires a large amount of speech data with transcription for acoustic model training. More than 100 hours of data are needed to achieve sufficient recognition accuracy, but collecting such

Manuscript received January 12, 2012.
Manuscript revised May 15, 2012.
[†]The authors are with Tokyo Institute of Technology, Tokyo, 152–8552 Japan.
[*]Presently, with NTT Corporation, Tokyo, 100–8116 Japan.
a) E-mail: shinoda@cs.titech.ac.jp
DOI: 10.1587/transinf.E95.D.2486

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 a high cost. Each language has different acoustic characteristics, and hence, the phones with low recognition accuracy vary from language to language. We applied our framework to collect utterances rich in those phones and proved its effectiveness in Japanese LVCSR. Additionally, in order to apply our method to the first category of active learning, in which utterances are selected from untranscribed speech data, we examine a semi-supervised 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 explains our active learning framework and Sect. 3 explain our sentence selection algorithm. Sections 4 and 5 explain our speaker adaptation method and our acoustic modeling method, respectively. Section 6 reports on our evaluation experiments using Japanese speech databases, and Sect. 7 concludes the paper.

2. Two-Step Active Learning

Our proposed active learning framework can be used both for *selecting* sentences from a text corpus and *generating* sentences from scratch. For simplicity, we explain the framework for sentence selection. A flowchart for this is shown in Fig. 1.

First, we prepare an acoustic model M and a small amount of data, Data E, and we recognize Data E using M to estimate the distribution of error occurrences. Let Data A be another transcribed data set, which will be augmented by active learning. In acoustic model training, Data A consists of transcribed speech data used to construct the acoustic model M. In speaker adaptation, Data A is empty, and the speaker-independent model is used as M.

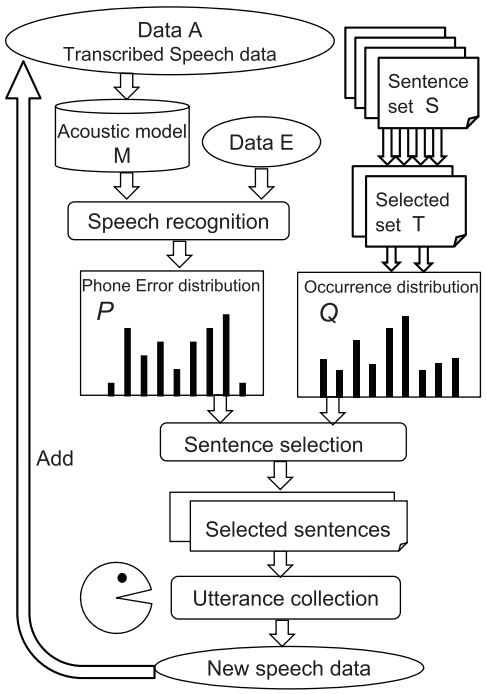

Fig. 1 Flow of two-step active learning.

Let U be a set of phones. The phone-error distribution $P(u)$ over phones $u \in U$ is defined as:

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

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.

Next, from a large text corpus (sentence set), S, prepared beforehand, we select those sentences whose distribution of phone occurrences is close to the phone-error distribution P. Let $c_X(u)$ be the number of occurrences of phone u in a set X and C_X be the total number of occurrences of all phones in X. Then, the phone-occurrence distribution Q_T of a set T of selected sentence is defined as:

$$Q_T(u) = \frac{c_T(u)}{C_T}. \tag{2}$$

Kullback-Leibler divergence (KLD) [11] 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 add them to Data A.

We can iterate this two-step process by adding the selected data to Data A and updating the acoustic model M using Data A. The resulting phone-error distribution P will become more precise because the data amount for estimating M increases. It will also represent more precisely the distribution of recognition errors of the updated model M. Accordingly, this iteration may enhance the effectiveness of our active learning framework.

We should address several issues when we apply this framework to acoustic modeling. Two major issues are:

1. How large should the amount of transcribed data, Data E, be at the beginning? If it is too small, the phone-error distribution P would be unreliable. On the other hand, we would like to keep it as small as possible to save on the transcription cost.
2. How often should we update model M? There is a trade-off between the performance and the computational cost. We can update it every time we select one sentence to obtain the highest performance. However, this frequent updating is very costly, since we should train model M and recognize Data E each time.

Apparently, we have no theoretically correct answers for these issues. In this study, we determine those two parameters, the data size E and the update frequency, empirically (i.e., in heuristic ways). We leave their optimization to our future work.

3. Sentence Selection Algorithm

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

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

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

When the number of phones is large, for example, when we use triphones in LVCSR, a relatively high computational cost is required to calculate KLD in Eq. (3) for every sentence s in a large corpus S. To reduce the cost, we approximate the difference Δ_s between the present KLD $D(P\|Q_{T\cup\{s\}})$ with a new sentence s and the KLD $D(P\|Q_T)$ in the previous step by using Taylor expansion.

First, Δ_s can be rewritten as follows:

$$\begin{aligned}
\Delta_s &= D(P\|Q_{T\cup\{s\}}) - D(P\|Q_T), \\
&= \sum_{u\in U} P(u) \log \frac{Q_T(u)}{Q_{T\cup\{s\}}(u)}, \\
&= \sum_{u\in U} P(u) \log \left(\frac{c_T(u)}{C_T} \cdot \frac{C_T + C_{\{s\}}}{c_T(u) + c_{\{s\}}(u)} \right), \\
&= \sum_{u\in U} P(u) \left(\log \left(1 + \frac{C_{\{s\}}}{C_T} \right) - \log \left(1 + \frac{c_{\{s\}}(u)}{c_T(u)} \right) \right). \quad (4)
\end{aligned}$$

Note that Δ_s should be negative to decrease KLD between P and Q.

Then, since it can be safely assumed that $c_{\{s\}}(u) \ll c_T(u)$ for all u, $C_{\{s\}} \ll C_T$, and $c_{\{s\}}(u)/c_T(u)$ and $C_{\{s\}}/C_T$ are in the same order,

$$\begin{aligned}
\Delta_s &\sim \sum_{u\in U} P(u) \left(\frac{C_{\{s\}}}{C_T} - \frac{c_{\{s\}}(u)}{c_T(u)} \right), \\
&= \frac{C_{\{s\}}}{C_T} \left(1 - \sum_{u\in U} P(u) \frac{Q_{\{s\}}(u)}{Q_T(u)} \right), \quad (5)
\end{aligned}$$

where $Q_{\{s\}}$ is the phone-occurrence distribution in sentence s. When we use a large set of recognition units such as triphones, most of them do not appear in a single sentence s. Since we can skip the addition in Eq. (5) for such units, the computational cost required for calculating Eq. (5) is much smaller than that for Eq. (3).

We calculate Δ_s for all sentences in S and choose the sentence which gives the smallest Δ_s. We repeat this sentence selection process until when Δ_s for all the remaining sentences in S becomes $\Delta_s \geq 0$.

In the sentence selection, we ignore phones 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.

4. Speaker Adaptation

4.1 MLLR

While we can apply our framework to any adaptation techniques, we apply it to one of the major techniques, maximum likelihood linear regression (MLLR) [15]. This method restricts the mapping from the initial model to the target speaker's model to be an affine transformation in the feature space, and it estimates the mapping parameters from the user's utterances. It updates the mean vector $\mu = (\mu_1, \cdots, \mu_n)^t$ in each Gaussian component in the output probabilities of the HMMs as follows

$$\hat{\mu} = A\mu + b, \quad (6)$$

where n is the dimension of the input feature vectors, A is an $n \times n$ matrix, and b is an n-dimensional vector. A and b are obtained by maximum likelihood estimation. A speaker-independent (SI) model is often used as the initial acoustic model for adaptation.

4.2 Active Learning in Speaker Adaptation

We apply our active learning for supervised adaptation, where each user speaks predetermined sentences to register their voice in the speech recognition system.

First we collect a certain amount of data, Data E, to measure the recognition accuracy. We do not collect Data A, since we can use the initial SI model for recognizing Data E. Then we move to the sentence selection process. Let us assume we would like to select N additional sentences in total in this process. There are two possible approaches: batch adaptation and sequential adaptation.

In batch adaptation, we select all the N sentences at the same time by using the phone-error distribution estimated by using the SI model to recognize Data E. Then, we collect speech data corresponding to the selected N sentences and carry out speaker adaptation using both Data E and this speech data where the SI model is used for the initial model for adaptation. In sequential adaptation, we update the acoustic model by speaker adaptation every time we collect one utterance. That is, we repeat N times the two-step process in Sect. 2. We employ batch adaptation in this study because it is much simpler and requires a lower computational cost.

5. Acoustic Model Training

5.1 Active Learning for Acoustic Model Training

While implementation of the proposed active learning to acoustic model training is rather straightforward, a few issues should still be discussed.

The number of utterances we collect is usually very large, more than ten thousand. Therefore, it is not affordable from the view point of computational cost to update the model each time we collect one sentence, as explained in Sect. 2. We apply the following strategy to avoid this problem. When we collect speech data of N sentences, we first collect Data A and Data E with K sentences. Then we divide the rest of $N - K$ sentences to be further collected into several blocks, and apply the two-step active learning process for each block. The block size (the number of sentences in each block) should be determined heuristically.

In LVCSR, we usually use triphones as recognition units. Since the number of triphones is fairly large, usually more than 1,000, its error distribution may not be reliable when it is estimated from a small number of error samples. We use the following strategy to avoid this problem. First, we estimate monophone error distribution. When its KLD from the monophone occurrence distribution converges, that is, when we cannot find any sentences that reduce the KLD in the provided text corpus, we stop the sentence selection process using monophones. We repeat the same process for diphone distributions and then move to triphones when the KLD again converges.

5.2 Selection from Untranscribed Speech Data

We have so far explained our active learning method for selecting a sentence set to collect read speech data. As explained in Sect. 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. On the contrary, our framework can be used only for collecting read speech data. Here we try to modify our framework to be applicable also for collecting spontaneous speech, in order to broaden the field of its application. For this purpose, we employ a semi-supervised learning method.

Basically, we apply a 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 available training data. 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 Sect. 2 and the previous subsection.

6. Experiment

We evaluated our active learning framework on speaker adaptation and acoustic model training tasks. In both cases, we simulated this framework by using fully transcribed speech databases. This is because it was impractical to collect speech data each time we updated the acoustic models. 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, instead of recording read speech for s. While we cannot measure the cost for collecting speech, we can evaluate the recognition performance of our framework as accurately as that in real situations.

6.1 Speaker Adaptation

6.1.1 Experimental Conditions

We evaluated our speaker adaptation method based on active learning in concatenated phone recognition using monophone HMMs. We used a database of Japanese newspaper article sentences (JNAS) [16] spoken by adults and senior citizens. To create this database, each speaker speaks about 100 sentences from newspapers and 50-100 phonetically balanced sentences. We used 522 speakers (261 speakers for each gender) for training, and 44 speakers (22 speakers for each gender) for testing.

The frame period for speech analysis was 10 ms, and the frame width was 25 ms. The input feature vector was 25-dimensional, consisting of 12-order mel-frequency cepstral coefficients (MFCCs), 12-order delta MFCCs, and a delta power. In the phone recognition experiment using monophone HMMs, we built a three-state speaker-independent HMM for each of 43 phone classes. There were 16 mixture components in each state.

For each test speaker, we used 100 sentences from newspaper articles, i.e., 60 sentences for adaptation and 40 sentences for testing. From the 60 sentences for adaptation, we randomly chose 5 sentences for Data E in the first adaptation step and used the remaining 55 sentences as the sentence pool in the second sentence selection step.

We used the batch adaptation framework. For the second step of this framework, we added a predetermined number of sentences to the 5 sentences selected in the first step and used these sentences in the supervised adaptation using MLLR. In all the experiments, the number of clusters for MLLR was 32, which gave the best performance when all the adaptation sentences were used.

We evaluated our method by using three different sentence sets in the first step (Sets 1, 2, and 3) in order to exclude unexpected biases in the evaluation. Figure 2 illustrates the data set design.

In the sentence selection, we ignored phone classes that rarely appeared, since their effect on the overall recognition accuracy was very small. We used the 27 phones indicated in Table 1.

In the evaluation, we employed concatenated phone recognition using a grammar representing the Japanese syllable structure. We used phone accuracies as the evaluation measures.

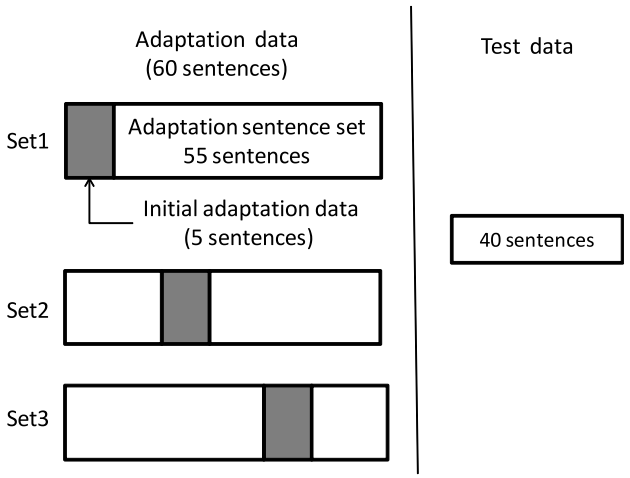

Fig. 2 Data set design for each speaker. The number of utterances from each speaker was 100.

Table 1 The 27 phone classes used in our speaker adaptation evaluation. Here, /Q/ is *sokuon*, /N/ is *hatsuon*, and /u:/ and /o:/ are long vowels.

/a/,/i/,/u/,/e/,/o/,/u:/,/o:/,/N/,/w/ /y/,/j/,/ky/,/t/,/k/,/ts/,/ch/,/b/,/d/ /g/,/z/,/m/,/n/,/s/,/sh/,/h/,/r/,/Q/

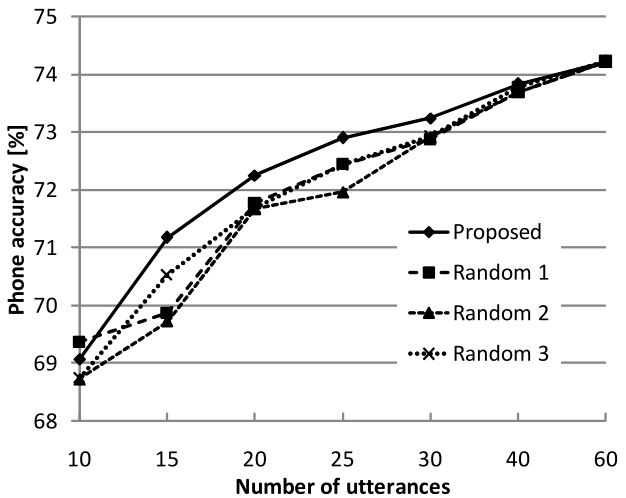

Fig. 3 Comparison of proposed method with random selection. Random 1, Random 2, and Random 3 are results obtained using three different random selections of adaptation sentences.

6.1.2 Results

First, we evaluated the proposed method on different numbers of the sentences selected in the second step. We chose Set 1 as the initial adaptation set. We compared our method with a *random selection* method, where the adaptation sentences used in the second step were randomly selected from the 55 sentences. We tested the random selection method three times with different seeds. The results averaged over all the phones are shown in Fig. 3 and Table 2. The phone accuracy obtained by the speaker-independent model averaged over all the test speakers was 64.0%.

Table 2 Phone accuracies of the proposed method and the three random selections (%). This table conveys the same information as Fig. 3.

No. of utterances	Proposed	Random 1	Random 2	Random 3
10	69.1	69.4	68.7	68.7
15	71.2	69.9	69.7	70.5
20	72.3	71.8	71.7	71.7
25	72.9	72.4	72.0	72.4
30	73.2	72.9	72.9	72.9
40	73.8	73.7	73.7	73.8
60	74.2	74.2	74.2	74.2

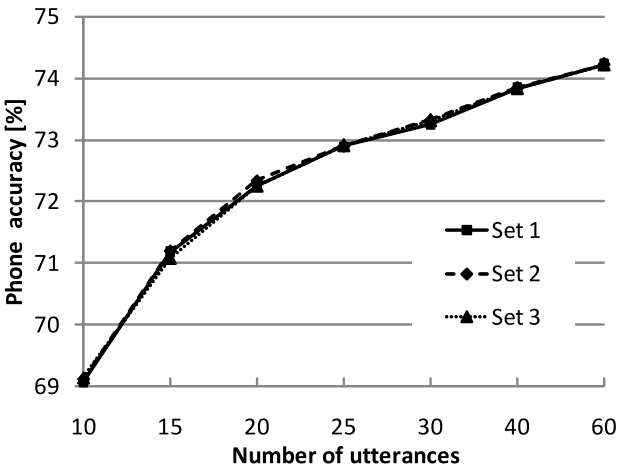

Fig. 4 Results of proposed method using different initial adaptation data.

The proposed method performed better than random selection in almost all cases; one random method performed better than the proposed method in the adaptation using 10 sentences. The improvement from the random selection level was the largest when 15 sentences were used. The accuracy was 1.1 absolute points higher than the average of the three random selection results. We also found that the difference between the accuracy of our method and each of the other two methods was statistically significant at 1% level when the numbers of utterances were 15, 20, 25, 30. These results indicate the effectiveness of the proposed method. Since there were 55 sentences in the sentence pool, the accuracies obtained by the proposed method and by the random selection converged to the same values as the number of sentences increased.

If the initial adaptation set is different, the additional sentences to be selected in the second step may also be different. To confirm the robustness of our method to changing the initial adaptation set, we changed the initial adaptation sentence set (by selecting Set 1, Set 2, or Set 3) and compared the corresponding results. Each of these sets contained 5 sentences. The results, shown in Fig. 4, proved to be almost the same as those for the initial adaptation set. It is therefore safe to say that the sentence selection for the initial adaptation set does not affect the performance of our method.

Figure 5 shows the results of the proposed method for

each speaker when there were 15 additional sentences. The accuracies for most speakers increased.

The phone error distributions were largely different from speaker to speaker. Figure 6 shows their examples. We also confirmed that, accordingly, the selected sentences were also largely different from speaker to speaker. Also, we did not find any cases in which the selected sentences by the proposed method coincidentally became similar to those by the random method.

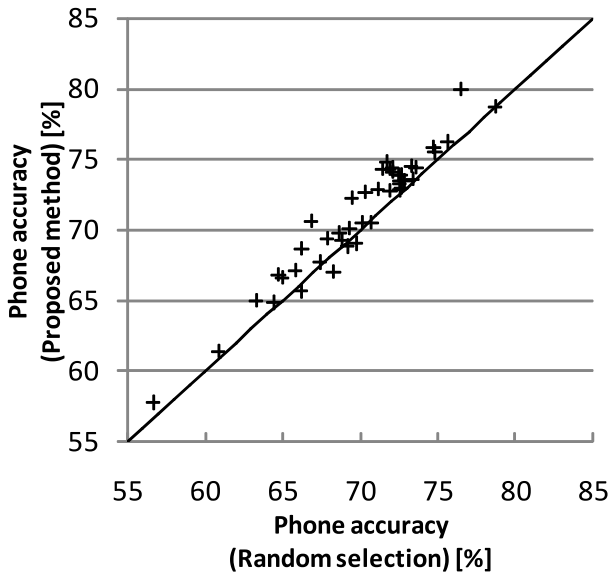

Fig. 5 Comparison of the proposed method with the average of the three random selection methods. The "+" symbols indicate the result for each speaker.

6.2 Acoustic Model Training

6.2.1 Experimental Conditions

We evaluated our acoustic model training method based on active learning using lecture-speech data obtained from male speakers in the Corpus of Spontaneous Japanese [17]. We used 198,807 utterances (152 h) from 666 speakers as training data, and 2,328 utterances (1.95 h) from 10 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 E. The other data from the training data (185,779 utterances, 142 h) were used as a text corpus 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 Sect. 3 to 10,000. 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)

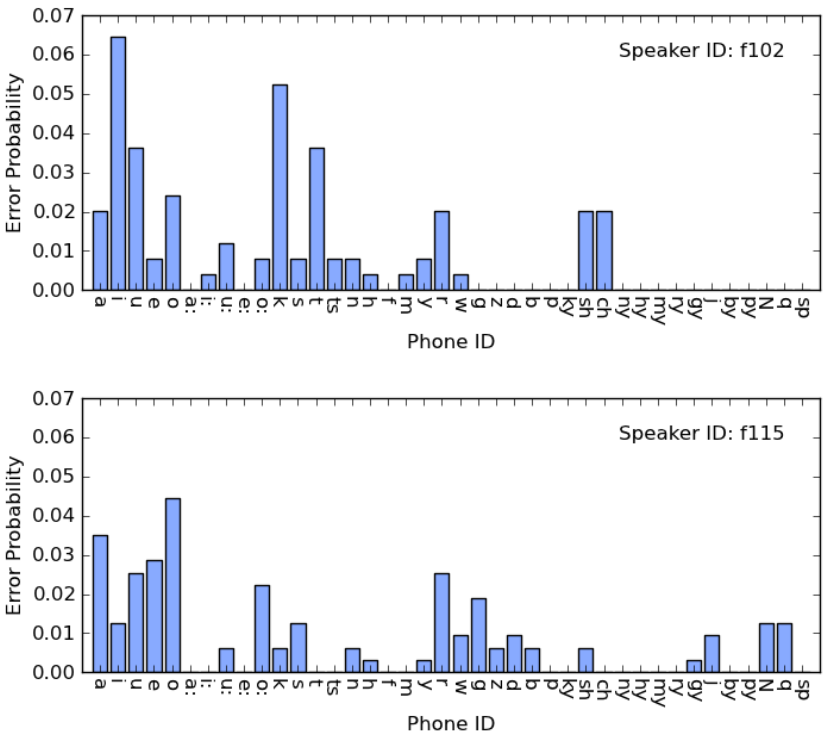

Fig. 6 Two examples of the phone error distribution.

with three states in phone recognition to estimate phone-error distribution P. There were 64 mixture components in each state. We used concatenated phone recognition with the same grammar as in Sect. 6.1

To evaluate recognition accuracy, we used triphone HMMs with 3,000 states, each of which had a Gaussian-mixture probability density function. There were 16 mixture components 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.

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 [7], which selects a sentence set such that the entropy of its phone distribution becomes maximum.

6.2.2 Comparison with Other Methods

Figure 7 plots the recognition results. We compared the proposed, random selection, and phonetically balanced sentence selection methods. We tested the random selection method three times with different seeds. Their averages are shown in Fig. 7. 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, whereas the other methods required 152 h. At the end of each phase, monophone, diphone, and triphone, the improvements of our method from the other two methods were statistically significant at the 1% level. The accuracy of the phonetically balanced method was almost the same as that with the random selection method. The phonetically balanced method is effective when there is an insufficient number of phones with low occurrence in the training data. However, in our situation, the amount of training data was large, and such phones occurred frequently enough in the training data. Because of this, the phonetically balanced method was not effective.

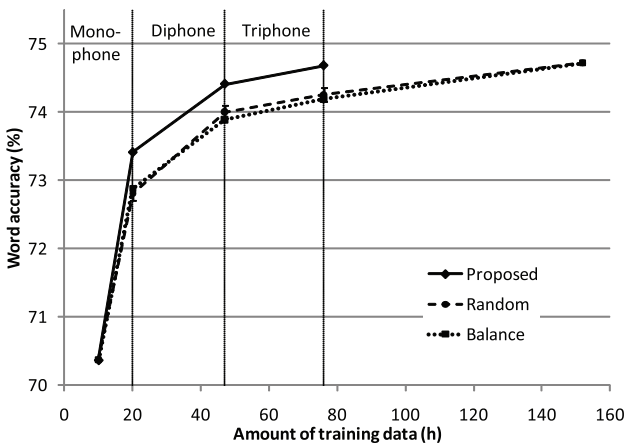

Fig. 7 Comparison of proposed method with random selection (Random) and phonetically balanced (Balance) methods.

6.2.3 KLD Values

Figure 8 plots 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 selected sentences increases. The final KLD value for each recognition unit class increases as the number of recognition units increases. Accordingly, it becomes more difficult to achieve Q closer to P.

6.2.4 Approximation

Table 3 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 computation time for sentence selection by 55% for diphones and 44% for triphones. For comparison, we report the time required for the other computation processes: training an acoustic model and recognizing Data E. For diphones, 6.5 h was required for training, and 1.0 h for recognition. Therefore, our method reduced the total computational costs by 16%. For triphones, 12.5 h was required for training, and 1.0 h for recognition. Thus, our method reduced the total computational cost by 9%. It should be noted that a large cost is also required to collect speech data.

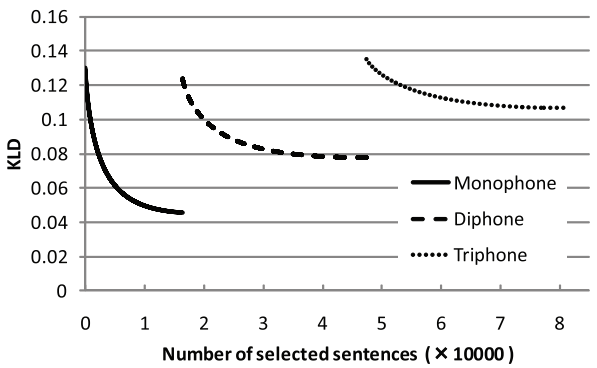

Fig. 8 Change in KLD values between the phone-error distribution and the accumulated phone-occurrence distribution in accordance with the number of selected sentences.

Table 3 Comparison of proposed method using approximation. Org indicates results without approximation, and App indicates results obtained from Eq. (5). The table shows recognition accuracy and time required for sentence selection. An Intel (R) Xeon (R) CPU (E5540, 2.53 GHz) was used for calculation. The 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

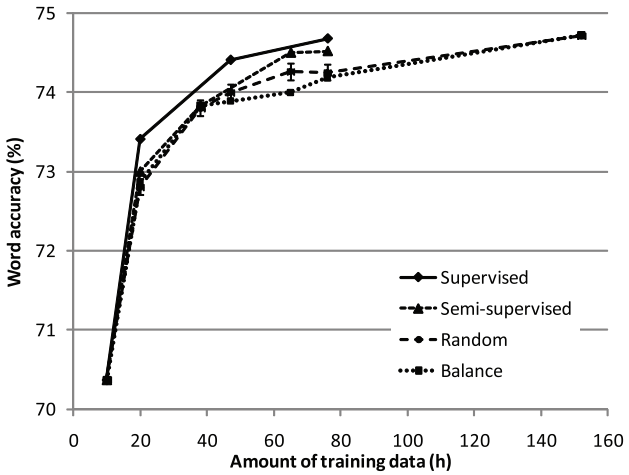

Fig. 9 Comparison of recognition results from semi-supervised training framework, supervised training framework, and two other methods.

6.2.5 Selection from Untranscribed Data

Figure 9 compares the recognition results of our semi-supervised learning method for selecting utterances from untranscribed data, explained in Sect. 5. We compared this method with the random selection and phonetically balanced methods. We also show our supervised learning method, where we assume correct transcriptions were given (Oracle). When there was 76 h of training data (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 appeared very often 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 the real situation.

7. Conclusion

We have proposed an active learning framework for constructing a speech 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 applied this framework to speaker adaptation and acoustic model training.

In speaker adaptation, our evaluation using phone recognition confirmed that it improved the phone accuracy by 1.1 absolute points from that of random selection. The database we used in this study was not large, and there were only 55 adaptation sentences in the adaptation sentence pool. Our method should be able to improve recognition performance even more if it is given more choices in the sentence selection process. We are planning to build an online evaluation scheme in which a large text-only database is prepared beforehand, and the sentences to be spoken by a subject are determined from the speech recognition results of his/her previous utterances.

In acoustic model training, we evaluated our method with simulation experiments using CSJ. Texts for 76 h of 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 applied to a semi-supervised training framework using untranscribed speech data, where a hypothesis transcription obtained by a speech decoder was used. 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.

As we described in the end of Sect. 2, the two control parameters in our active learning, the size of Data E and the update frequency, were determined empirically in this study. We need more study for their optimization.

While we used maximum-likelihood estimation for model parameter estimation in our evaluation, the combination of our framework and discriminative learning is also expected to yield higher recognition accuracies. We would like to implement them with 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 situation in which we generate texts whose corresponding speech data are expected to be effective in reducing errors. We plan to construct an on-line training system for this purpose. We also plan to extend our method to recognition units with longer contexts such as words.

Acknowledgements

This work was partly supported by Grant-in-Aid for Scientific Research (B) 20300063.

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. Interspeech 2005, 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] Y. Hamanaka, K. Shinoda, T. Tsutaoka, S. Furui, T. Emori, and T. Koshinaka, "Committee-based active learning for speech recognition," IEICE Trans. Inf. & Syst., vol.E94-D, no.10, pp.2015–2023, Oct. 2011.
[7] 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.
[8] 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," Comput. Speech Lang., vol.13, pp.79–97, 1999.
[9] X. Cui and A. Alwan, "Efficient adaptation text design based on the Kullback-Leibler measure," Proc. ICASSP2002, pp.I-613–616, 2002.
[10] Q. Huo and W. Li, "An active approach to speaker and task adaptation based on automatic analysis of vocabulary confusability," Proc. Interspeech 2007, pp.1569–1572, 2007.
[11] S. Kullback, R.A. Leibler, and J.B. MacQueen, "On information and sufficiency," Annals of Mathematical Statistics, vol.22, no.1, pp.79–86, 1951.
[12] H. Murakami, K. Shinoda, and S. Furui, "Speaker adaptation based on two-step active learning," Proc. INTERSPEECH2009, pp.576–579, 2009.
[13] H. Murakami, K. Shinoda, and S. Furui, "Designing text corpus using phone-error distribution for acoustic modeling," Proc. IEEE 2011 Automatic Speech Recognition and Understanding Workshop, pp.191–195, 2011.
[14] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. Speech Audio Process., vol.2, no.2, pp.291–298, 1994.
[15] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous-density hidden Markov models," Comput. Speech Lang., vol.9, pp.171–185, 1995.
[16] K. Itou, M. Yamamoto, K. Takeda, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," J. Acoust. Soc. Jpn. (E), vol.20, no.3, pp.199–206, 1999.
[17] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," Proc. LREC, vol.2, pp.947–952, 2000.

Hiroko Murakami received her B.S. in 2009 and her M.S. in 2011, both in computer science from the Tokyo Institute of Technology, Japan. She is now with Nippon Telegraph and Telephone Corporation.

Koichi Shinoda received his B.S. in 1987 and his M.S. in 1989, both in physics, from the University of Tokyo. He received his D.Eng. in computer science from the Tokyo Institute of Technology in 2001. In 1989, he joined NEC Corporation, Japan, and was involved in research on automatic speech recognition. From 1997 to 1998, he was a visiting scholar with Bell Labs, Lucent Technologies, in Murray Hill, NJ. From June 2001 to September 2001, he was a Principal Researcher with Multimedia Research Laboratories, NEC Corporation. From October 2001 to March 2002, he was an Associate Professor with the University of Tokyo. He is currently an Associate Professor at the Tokyo Institute of Technology. His research interests include speech recognition, statistical pattern recognition, and human interfaces. Dr. Shinoda received the Awaya Prize from the Acoustic Society of Japan in 1997 and the Excellent Paper Award from the Institute of Electronics, Information, and Communication Engineers IEICE in 1998. He is an Associate Editor of Computer Speech and Language. He is a member of IEEE, ACM, ASJ, IPSJ, and JSAI.

Sadaoki Furui received his B.S., M.S., and Ph.D. in mathematical engineering and instrumentation physics from Tokyo University, Tokyo, Japan in 1968, 1970, and 1978. He is engaged in a wide range of research on speech analysis, speech recognition, speaker recognition, speech synthesis, and multimodal human-computer interactions, and has authored or coauthored over 800 published articles. He has received Paper Awards and Achievement Awards from the IEEE, the IEICE, the ASJ, the ISCA, the Minister of Science and Technology, and the Minister of Education. He has also been a recipient of the prestigious Purple Ribbon Medal from the Japanese Emperor.