論文 / 著書情報
Article / Book Information

Title	Speaker Adaptation Based on Two-Step Active Learning
Authors	Koichi Shinoda, Hiroko Murakami, Sadaoki Furui
Citation	Proc. INTERSPEECH 2009, Vol. , No. , pp. 576-579,
Pub. date	2009, 9
Copyright	(c) 2009 International Speech Communication Association, ISCA
DOI	http://dx.doi.org/

INTERSPEECH 2009 BRIGHTON

Speaker Adaptation Based on Two-Step Active Learning

Koichi Shinoda, Hiroko Murakami, and Sadaoki Furui

Department of Computer Science, Tokyo Institute of Technology, 152-8552, Japan
{shinoda,furui}@cs.titech.ac.jp, murakami@ks.cs.titech.ac.jp

Abstract

We propose a two-step active learning method for supervised speaker adaptation. In the first step, the initial adaptation data is collected to obtain a phone error distribution. 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. We evaluated the method using a Japanese speech database and maximum likelihood linear regression (MLLR) as the speaker adaptation algorithm. We confirmed that our method had a significant improvement over a method using randomly chosen sentences for adaptation.
Index Terms: speech recognition ,speaker adaptation, active learning

1. Introduction

Speaker adaptation techniques (e.g., [1, 2]) to improve speech recognition performance by using a small number of utterances from users are often used in many applications. They fall into two categories: supervised adaptation and unsupervised adaptation. In supervised adaptation, users are asked to speak sentences prepared beforehand, and therefore, the transcriptions for the utterances are known. On the other hand, in unsupervised adaptation, the transcriptions are not given. In both categories, it is desirable to achieve a larger improvement with a smaller number of utterances.

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. Given this motivation, we focuse on sentence selection based on active learning for speaker adaptation.

Active learning has been extensively studied in speech recognition [3, 4, 5]. In most of these studies, it has been used to select sentences in training data for acoustic modeling in order to decrease the annotation effort. Since it is assumed that the transcriptions are not given beforehand, the focus of these studies has been to find an effective uncertainty measure for each utterance; those utterances whose transcriptions seem to be highly uncertain are preferred as training data. The same approach can be applied to unsupervised adaptation, where the transcriptions are not available.

Supervised adaptation can take a different approach, since the transcriptions are available. In this case, the focus is on how to design an adaptation sentence set for each speaker, and there have been a few studies on it [6, 7, 8]. Shen *et. al.* [6] selected a phonetically balanced phone sentence set for a given task. Huo *et. al.* [8] selected a vocabulary consisting of words that were expected to be highly confusable in a given task. These approaches, however, do not consider the acoustic characteristics of the speaker's voice. They should be classified as task adaptation methods, in the sense that they do not improve the recognition accuracies in comparison with the conventional adaptation frameworks when the initial speaker-independent acoustic model has already been tuned to the given recognition task.

In this paper, we propose a two-step active learning method for supervised speaker adaptation. In the first step, our method collects a small amount of utterances from a user to obtain his/her tendency in speech recognition errors. In the second step, it selects those sentences rich in phonetic units in the errors from a sentence pool, and it collects their utterances as additional data for adaptation. Since our method directly aims at decreasing recognition errors, it is expected to be highly discriminative. Our method has two critical issues: One is how to relate the recognition errors to the selection criterion of adaptation sentences. The other is how to set the size of the initial adaptation data in the first step; it should be as small as possible in order to decrease the user's effort, but it must be sufficient enough to estimate the tendency of errors precisely. We describe our approach to resolving these issues in this paper.

This paper is organized as follows. Section 2 explains our method, and Section 3 briefly explains the MLLR speaker adaptation method. Section 4 reports our evaluation experiments using a Japanese speech database, and Section 5 concludes the paper.

2. Sentence selection based on active learning

2.1. Overview

Figure 1 is the flowchart of our method. First, our method measures the recognition accuracy for each phone class of a target speaker from a small amount of his/her utterances and obtains the distribution of errors over all phone classes. Second, from an adaptation sentence set prepared beforehand, it selects those sentences whose distributions of phone-class occurrences are close to the phone-error distribution and asks the speaker to speak them. Finally, it carries out the adaptation process using the utterances collected in the first and second steps.

2.2. Phone-error distribution

In the first step of the adaptation, we ask a speaker to speak a small number of sentences. Then, we apply continuous phone recognition using the initial speaker-independent (SI) model to these initial adaptation data and obtain the phone accuracy $a(u_i)$ for each phone u_i. Given the phone-error rate $r(u_i) = 1 - a(u_i)$ for each phone u_i, a phone-error distribution $P(\mathrm{u})$of the target speaker over all the phone classes is defined as follows

$$P(u_i) = \frac{r(u_i)}{\sum_{j=1}^{n} r(u_j)}, \quad i = 1, \cdots, n \tag{1}$$

where n is the number of phone classes in the target language. Since the number of utterances in the initial adaptation data is

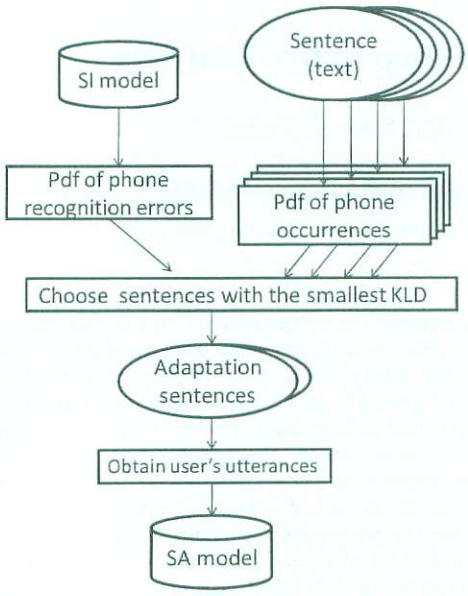

Figure 1: Flow of the proposed method

small, some phone classes do not appear in the initial adaptation data. Here, we simply set n to be the number of phone classes which appear in the initial adaptation data. It should be noted that we normalize the number of occurrences of each phone in Eq. 1 to decrease the influence of frequently appearing phones, such as vowels.

2.3. Phone distribution in a sentence set

An adaptation sentence set is prepared before the second adaptation step. It does not involve the sentences used in the first step. Let N be the number of sentences in the set and t_k be the k-th sentence in the set where $k = 1, 2, \ldots, N$. The phone distribution in t_k is defined as

$$Q_k(u_i) = \frac{s_k(u_i)}{\sum_{j=1}^{n} s_k(u_j)}, \quad i = 1, \ldots, n, \tag{2}$$

where s_k is the number of occurrences of each phone class u_i. In each sentence t_k, some phone classes u_i may not appear, and thus, their $Q_k(u_i)$'s become zero. We set a small non-zero value δ to $s_k(u_i)$ when u_i is not observed in sentence t_k to avoid calculation difficulties in the following process.

2.4. Kullback-Leibler divergence

In our active learning approach, we select from the adaptation sentence set those sentences whose phone distribution $Q_k(\mathrm{u})$ is close to the phone-error distribution $P(\mathrm{u})$ of the target speaker. For this purpose, we calculate Kullback-Leibler divergence (KLD) from $P(\mathrm{u})$ to $Q_k(\mathrm{u})$ of each sentence k, which is defined as follows

$$D_k(Q_k(u_i)||P(u_i)) = \sum_{i=1}^{n} Q_k(u_i) \log \frac{Q_k(u_i)}{P(u_i)} \tag{3}$$

Those sentences with small KLD values are selected as additional sentences for adaptation. In our implementation described in Section 4, we set a threshold M on the number of additional sentences and select M sentences with smaller KLDs than those of the other sentences.

2.5. Adaptation process

In the second step, the system asks the target speaker to speak the sentences selected in the previous process, and it uses the collected utterances for adaptation. There are two possible approaches for adaptation: batch adaptation and sequential adaptation. Here, we would like to select M additional sentences in total for the second step.

In batch adaptation, we select M sentences at the same time by using the phone-error distribution estimated by using the SI model to recognize the initial adaptation data. Then, speaker adaptation is carried out using both the initial adaptation data and the additional adaptation data, where the initial model for adaptation is the SI model.

Sequential adaptation is carried out as follows until the number of additional sentences becomes M.

1. Set the initial adaptation data to the adaptation data set A.
2. Carry out speaker adaptation using A.
3. Use the adapted model to recognize A in order to estimate the phone-error distribution.
4. Select one sentence s by using the updated phone-error distribution.
5. Ask the target speaker to speak sentence s and add the resulting utterances to A. Go to 2.

In our evaluation described in Section 4, we employed batch adaptation, since its implementation was easier.

3. MLLR

We employ a speaker adaptation algorithm using maximum likelihood linear regression (MLLR) [2]. This algorithm 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. We update the mean vector $\mu = (\mu_1, \cdots, \mu_n)^t$ in each Gaussian component in the output probabilities of the HMMs as follows

$$\hat{\mu} = A\mu + b, \tag{4}$$

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.

4. Experiment

We evaluated our method using a Japanese read-speech database. We used concatenated phone recognition using monophone HMMs to confirm the effectiveness of our method.

4.1. Experimental conditions

The evaluation used a database of Japanese newspaper article sentences (JNAS) [9] spoken by adults and seniors. In 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.

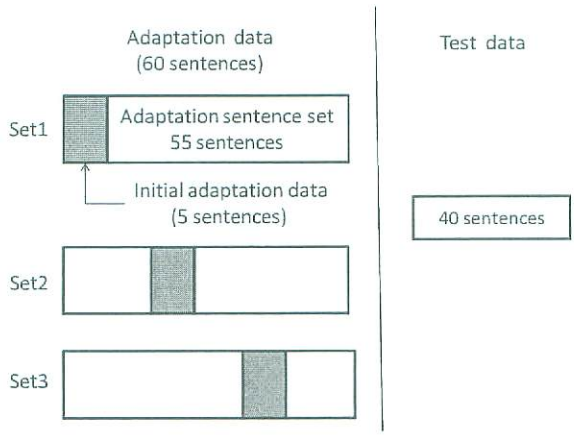

Figure 2: Data set design for each speaker. The number of utterances from each speaker was 100.

/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/

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

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

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 five sentences in the first adaptation step and used the remaining 55 sentences as the sentence pool in the second adaptation step.

We used the batch adaptation framework. For the second step of this framework, we added a predetermined number of sentences to the five sentences selected in the first step and used these sentences in the supervised adaptation using MLLR. The number of clusters for MLLR was 32.

We evaluated our method by using three different sentence sets in the first step (Set 1, Set 2, and Set 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 influence on the overall recognition accuracy was very small. We used the 27 phone shown in Table 1. We set the non-negative small value δ described in Subsection 2.3 to 1.0×10^{-25} according to the result of our preliminary experiment.

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

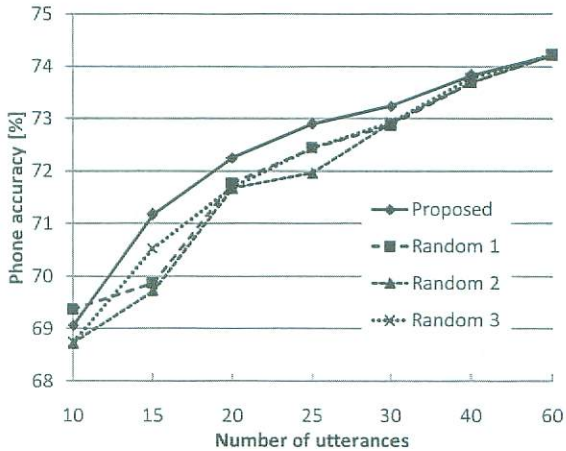

Figure 3: Comparison of proposed method with random selection. Random 1, Random 2, and Random 3 are results obtained by three different random selections of adaptation sentences.

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

Table 2: The phone accuracies of the proposed method and the three random selections[%]. This table conveys the same information as Fig. 3.

4.2. Results

4.2.1. Phone recognition

First, we evaluated the proposed method on different numbers of the selected sentences 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 speak-independent model averaged over all the test speakers was 64.0%.

The proposed method performed better than random selection in most cases; one random method was better than the proposed method in the adaptation using ten sentences. The improvement from the random selection level was the largest when 15 sentences were used. The accuracy was 1.1 points absolute higher than the average of the three random selection results. This result demonstrated the effectiveness of the proposed method. Since the number of sentences in the sentence pool was 55, the accuracies obtained by the proposed method and by the random selection converged to the same values as

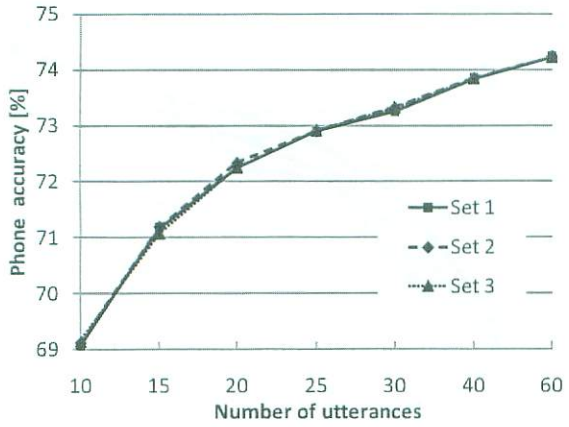

Figure 4: Results of the proposed method using different initial adaptation data

the number of sentences increased.

If the initial adaptation set is different, the additional sentences to be selected in the second step may 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 five 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 the number of the additional sentences was 15. The accuracies for most speakers increased.

5. Conclusion

We proposed an active learning method for supervised adaptation that selects adaptation sentences that are expected to be effective in improving recognition performance and uses their utterances for adaptation. Our evaluation using phone recognition confirmed that it improved the phone accuracy by 1.1 points absolute from that of random selection.

We plan to apply our method for large-vocabulary continuous speech recognition using triphone HMMs to confirm the effectiveness of our method in larger tasks. While we evaluated our method in the batch adaptation mode, sequential adaptation is likely to be more efficient. We would like to continue our research in this direction.

The database we used in this study was not large, and the number of adaptation sentences in the adaptation sentence pool was only 55. 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.

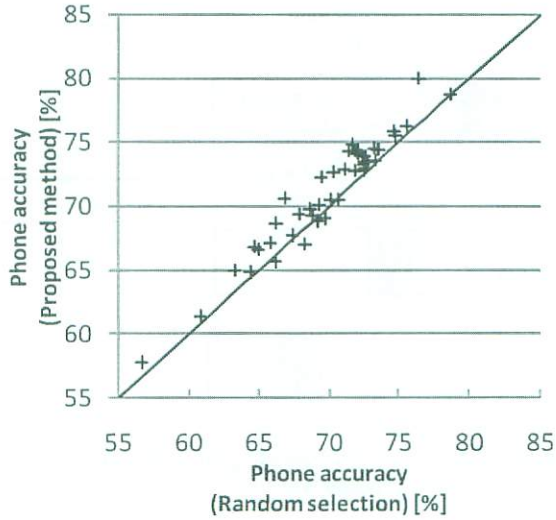

Figure 5: Comparison of the proposed method with the average of the three random selection methods. The symbol "+" indicates the result for each speaker.

6. Acknowledgement

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

7. References

[1] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. Speech and Audio Processing, vol. 2, pp. 291-298, 1994.

[2] C.J. Leggetter *et al.*, "Maximum likelihood linear regression for speaker adaptation of continuous-density hidden Markov models," Computer Speech and Language, vol. 9, pp. 171-185, 1995.

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

[4] H.-K. Kuo and V. Goel, "Active learning with minimum expected error for spoken language processing," Proc. Interspeech2005, pp. 437-440, 2005.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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 - 06, 1999.

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