Transcription Cost Reduction for Constructing Acoustic Models Using Acoustic Likelihood Selection Criteria

Tomoyuki Kato, Tomiki Toda, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara, Japan 630-0192
{tomoyu-k, tomoki, sawatari, shikano}@is.naist.jp

Abstract

This paper describes a novel method for reducing the transcription effort in the construction of task-adapted acoustic models for a practical automatic speech recognition (ASR) system. We have to prepare actual data samples collected in the practical system and transcribe them for training the task-adapted acoustic models. However, transcribing utterances is a time-consuming and laborious process. In the proposed method, we firstly adapt initial models to acoustic environment of the system using a small number of collected data samples with transcriptions. And then, we automatically select informative training data samples to be transcribed from a large-sized speech corpus based on acoustic likelihoods of the models. We perform several experimental evaluations in the framework of 'Takemaru-kun', a practical speech-oriented guidance system. Experimental results show that 1) utterance sets with low likelihoods cause better task-adapted models compared with those with high likelihoods although the set with the lowest likelihoods causes the performance degradation because of including outliers, and 2) MLLR adaptation is effective for training the task-adapted models when the amount of the transcribed data is small and EM training outperforms MLLR if we transcribe more than around 10,000 utterances.

1. Introduction

Dramatic improvements of automatic speech recognition (ASR) techniques have enable ASR systems to be used practically. There are many ASR applications in various fields. In order to achieve the sufficient performance of ASR systems, we need to use proper acoustic and language models for specific conditions in which the systems are used (Gao et al., 2005; Lefevre et al., 2005). Since the conditions such as acoustic environments, speakers, and utterance contents are quite different between systems, it is impractical to prepare the models that cover every condition. Therefore, we need to construct task-adapted models for individual systems.

In order to train the task-adapted models, we have to prepare actual data samples collected in a practical system and transcribe them. However, transcribing utterances is a time-consuming and laborious process. This burden has become a critical issue for the practical use of ASR systems. The study for reducing the transcription effort has been investigated by many researchers and developers. It has been reported that the amount of transcribed data used for the training of language models can be reduced by about 30 % while keeping the recognition performance when using all of data samples (Hakkani-Tur et al., 2002).

Not only language models but also acoustic models should be adapted for a specific task in the ASR system. It is useful to investigate the reduction effect of transcribed data for the training of acoustic models as well. There is a little research for doing it (Kamm and Meyer, 2004). This paper describes a novel method for reducing the transcription effort in the construction of the task-adapted acoustic models. In the proposed method, we firstly adapt initial models to acoustic environment of the system using a small number of collected samples with transcriptions. And then, we automatically select informative training data samples to be transcribed from a large-sized speech corpus based on acoustic likelihoods of the models. We perform experimental evaluations of the proposed method using actual speech data collected in a practical speech-oriented guidance system in which the large vocabulary continuous speech recognition (LVCSR) is employed. We also investigate an appropriate training method of the task-adapted models for various amounts of the transcribed speech data.

The paper is organized as follows. In **Section 2.**, the proposed method is described. In **Section 3.**, we perform experimental evaluations for demonstrating the effectiveness of the acoustic likelihood selection criterion and investigating an appropriate model training method according to the amount of transcribed data. Finally, we summarize this paper in **Section 4.**.

2. Efficient Model Construction with Acoustic-Likelihood-Based Data Selection

Figure 1 shows the proposed method for selecting informative training data samples to be transcribed. Specific processes are shown as follows:

I. We collect a small number of speech samples and transcribe them. In order to adapt existing acoustic models, i.e., hidden Markov models (HMMs) for the environment of the system, we perform Maximum Likelihood Linear Regression (MLLR) adaptation (Gales and Woodland, 1996). The adapted models are named 'intermediate models.'

II. Using the intermediate models, we perform ASR for all utterances in the speech corpus, and then we calculate acoustic likelihoods of the models for individual utterances. We select an informative utterance set based on the likelihoods and transcribe the selected utterances.

III. Using those utterances, we train task-adapted models with EM (Forward-Backward) algorithm or MLLR.

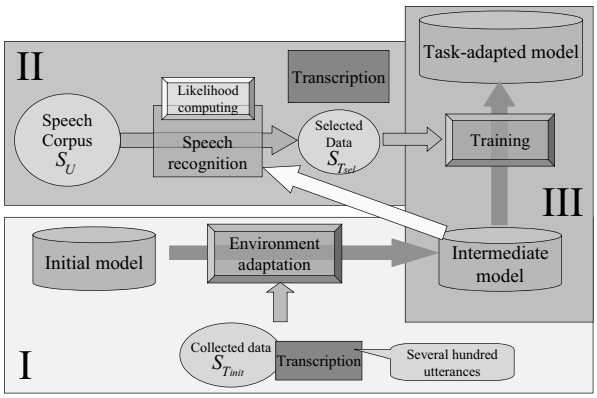

Figure 1: Algorithm for efficient construction of task-adapted models.

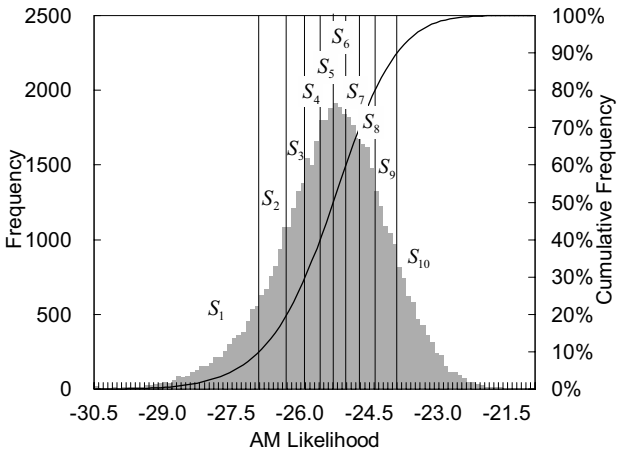

Figure 2: Frequency distribution of acoustic likelihoods for utterances in a speech corpus collected for 13 months.

It is necessary to investigate the following points:

- *How to select the informative samples:*
  It is expected that data samples causing low likelihoods of the intermediate model are informative for training the task-adapted models because they are not represented very well by the current models. One of selection methods is just to select samples with the low likelihoods. We need to experimentally demonstrate whether this method works well or not. We also need to compare the likelihood with a confidence measure that is used as a selection criterion in the similar researches (Hakkani-Tur et al., 2002; Kamm and Meyer, 2004).

- *How to train the task-adapted models:*
  It is reasonable to choose an appropriate training method of the task-adapted models according to the amount of transcribed data. An adaptation-based method such as MLLR works very well when using a small amount of training data (Gales and Woodland, 1996). On the other hand, the EM algorithm powerfully works when using a large amount of training data. We actually compare these two from a view of the amount of transcribed data.

3. Experimental Evaluations

We performed experimental evaluations of the proposed method in the framework of 'Takemaru-kun', a practical speech-oriented guidance system (Nisimura et al., 2000). Takemaru-kun system has been developed to study a spoken dialogue interface through long-term operation in a public place. We have collected natural human-machine interaction data for more than two and a half years.

3.1. Experimental Conditions

We used 80,666 utterances that had been collected for 19 months (Apr. 2003 through Oct. 2004) as a speech corpus. These utterances didn't include noisy data that had already been removed manually. As a test set, we used 1,000 utterances which were not included in the speech corpus. These test utterances were selected from a speech database collected for 5 months (Nov. 2002 through Mar. 2003). They didn't also include noisy data. Voices of various age-groups such as child, junior, adult, and senior were included in both the speech corpus and the test set.

We used phonetic tied-mixture (PTM) models (Lee et al., 2000) trained with JNAS (Japanese Newspaper Article Sentences) database (Itou et al., 1999) as initial acoustic models. The number of states was set to 2000. The number of mixtures was set to 64. As for language models, we used 40,000 words 3-gramm models adapted for Takemaru-kun system. We employed Julius ver. 3.5 decoder (Lee et al., 2001).

In a preliminary experiment, we investigated how large amount of adaptation data is necessary for constructing the intermediate models (the process I in **Fig. 1**). An experimental result showed that 350 utterances that had been collected for only one day is enough for constructing the proper intermediate models with MLLR. Therefore, we used the intermediate models adapted with those utterances in the following evaluations.

3.2. Evaluation of Selection Method

We calculated an acoustic likelihood of the intermediate models divided by the number of frames for each utterance in a part of the speech corpus consisting of 52,161 utterances collected for 13 months (Apr. 2003 through Apr. 2004). A frequency distribution of the likelihoods is shown in **Fig. 2**. We prepared several utterance sets by dividing all utterances into several partitions ($S_1 \sim S_{10}$ in the figure) based on the likelihoods while keeping the number of utterances in each set equal. And then, we trained the task-adapted models with the individual sets. MLLR was employed for the model training.

3.2.1. Effectiveness of likelihood selection

Figure 3 shows word accuracies when using individual task-adapted models. As references, it also shows a result of using all utterances and that of using utterances selected at random. We can see a tendency that the utter-

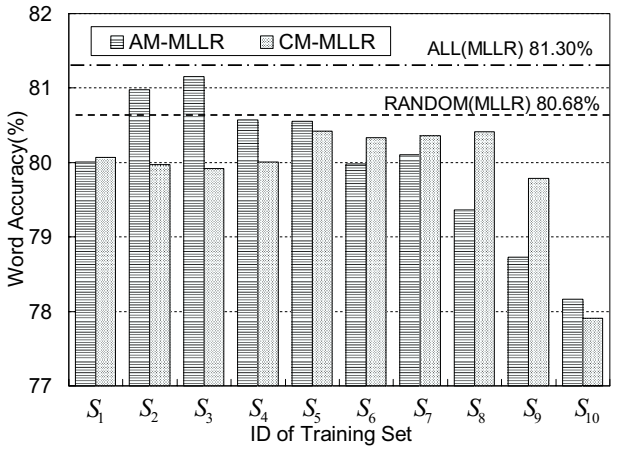

Figure 3: Word accuracies when using two kinds of task-adapted models. One, 'AM-MLLR' was trained using an utterance set selected with acoustic likelihood. The other, 'CM-MLLR' was trained using an utterance set selected with confidence measure. We selected only 10 % samples from a speech corpus collected for 13 months.

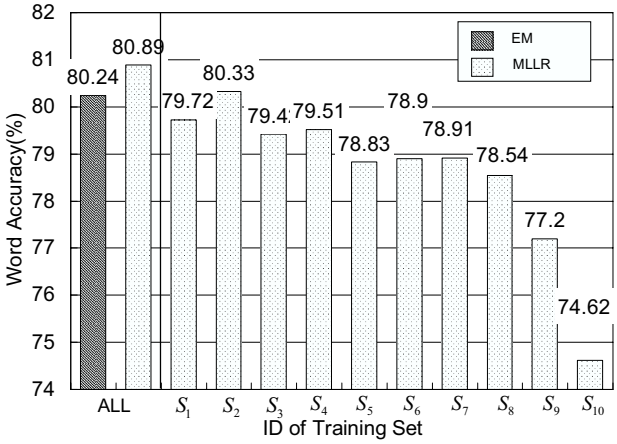

Figure 4: Word accuracy when using task-adapted models trained with each utterance set selected with acoustic likelihoods. We used a speech corpus collected for 3 months.

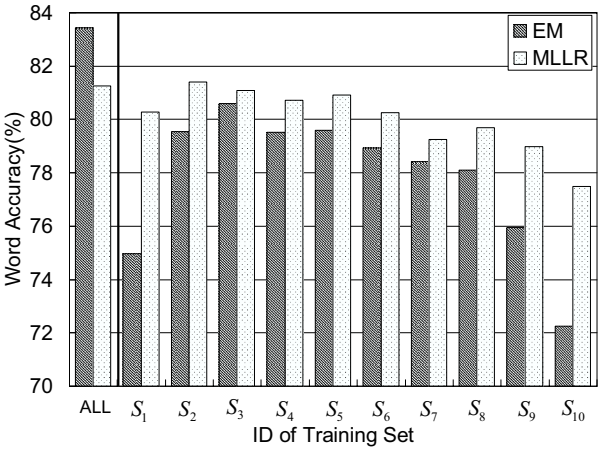

Figure 5: Word accuracy when using task-adapted models trained with each utterance set selected with acoustic likelihoods. We used a speech corpus collected for 19 months.

ance sets with low likelihoods causes better task-adapted models compared with those with high likelihoods as mentioned in **Section 2.**. However, the set with the lowest likelihoods causes the performance degradation because it might include outliers. Consequently, the utterance set with the second or the third lowest likelihoods often causes the best task-adapted models. Note that using those sets causes better results than using the set selected at random.

We compared the acoustic likelihood criterion with a confidence measure (Lee et al., 2004). Results of using the confidence measure are also shown in **Fig. 3**. It is shown that the acoustic likelihood works better than the confidence measure. It might be possible that the confidence measure is not always effective for the data selection in the practical LVCSR system. One of advantages of the likelihood criterion is less affected by the performance of language models than the confidence measure.

3.2.2. Validation of likelihood selection

In order to validate whether the same tendency as mentioned above is observed or not when using other speech corpora, we performed the same evaluations using two different sizes of the speech corpora. One consisted of 10,014 utterances collected for 3 months (Apr. 2003 through June 2003) and the other consisted of all utterances collected for 19 months.

Experimental results are shown in **Fig. 4** and **Fig. 5**. We can again see the similar results to the previous one. Namely, the set with the lowest likelihoods causes the performance degradation although the sets with low likelihoods are more informative than those with high likelihoods.

3.3. Evaluation of Training Method

Results of employing EM training are also shown in **Fig. 5**. We can see the same tendency as described in the MLLR case. It is shown that EM training with the selected utterance sets is inferior to MLLR because the amount of training data is insufficient for estimating all pdfs. If we have the sufficient amount of training data, EM training outperforms MLLR as shown in the figure ('EM' bar at 'ALL').

In order to investigate an appropriate training method according to the amount of training data, we varied the size of the selected utterance set from 10 % to 50 % of the speech corpus collected for 13 months.

Figure 6 shows word accuracy as a function of the ratio of the number of training utterances to all utterances in the corpus. In the proposed likelihood selection, we selected an utterance set causing the best word accuracy for each size of the training data as shown in **Table 1**. As for the random selection, we randomly selected training utterances ten times and calculated an average of word accuracies. A range between the minimum and maximum word accuracies is also shown as a gray area when employing EM training. It is

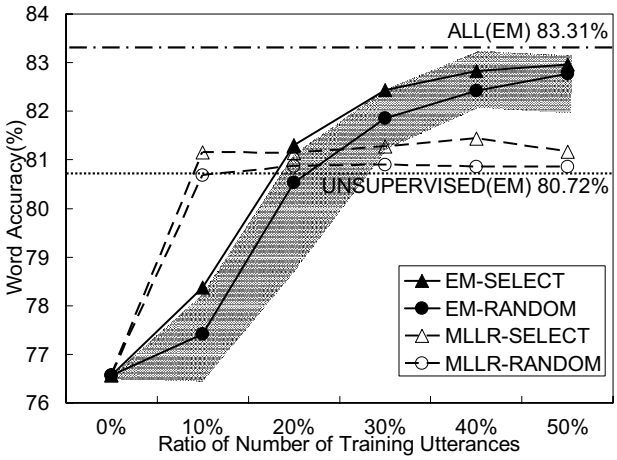

Figure 6: Word accuracy as a function of ratio of the number of training utterances to all utterances in a speech corpus collected for 13 months. 'ALL (EM)' and 'UNSUROERVISED (EM)' show results of using all utterances in the corpus for supervised EM training and unsupervised EM training, respectively. The unsupervised EM uses transcriptions automatically generated with the intermediate models.

Table 1: Utterance-set causing the best recognition performance when using the speech corpus collected for 13 months

	10%	20%	30%	40%	50%
EM	S_3	$S_2 \sim S_3$	$S_2 \sim S_4$	$S_3 \sim S_6$	$S_1 \sim S_5$
MLLR	S_3	$S_3 \sim S_4$	$S_2 \sim S_4$	$S_3 \sim S_6$	$S_1 \sim S_5$

observed that the performance of the proposed selection is almost equal to that of the best case in the random selection. When the size of the selected utterance set is 10 %, MLLR adaptation works better than EM training. The performance of EM reaches the same level of MLLR when using 20 % of the corpus (around 10,000 utterances). EM clearly outperforms MLLR when using more utterances.

4. Conclusion

We proposed an efficient training method of task-adapted acoustic models for reducing the transcription effort. The proposed method automatically selected informative training data samples to be transcribed from a large-sized speech corpus based on acoustic likelihoods. Experimental evaluations were conducted in the framework of a practical speech-oriented guidance system. As a result, it was shown that 1) utterance sets with low likelihoods cause better task-adapted models compared with those with high likelihoods although the set with the lowest likelihoods causes the performance degradation because of including outliers, and 2) MLLR adaptation effectively works when the amount of the transcribed data is small and EM training outperforms MLLR if we transcribe more than around 10,000 utterances.

In development of a practical ASR system, it is impractical to collect a sufficient amount of actual speech data for training the task-adapted models in the short term. Furthermore, because the cost for the system development is practically limited, the number of transcribed data would also be limited. The experimental results in this paper suggest one good approach for the construction of the practical system: 1) we collect a small number of actual speech samples, e.g., for one day, and transcribe them for adapting initial acoustic models to the environment of the system, 2) we automatically select a certain amount of speech data that can be transcribed from collected speech data based on acoustic likelihoods of the adapted models, and 3) we train task-adapted models employing MLLR or EM according to the amount of the selected data.

This paper doesn't address a problem of the language model training. It is necessary to investigate an efficient training method of the language models as well.

Acknowledgment:

This research was supported in part by MEXT e-Society leading project.

5. References

Y. Gao, L. Gu, and H.-K. Jeff Kuo. Portability challenges in developing interactive dialogue systems. *Proc. ICASSP*, pp. 1017–1020, 2005.

F. Lefevre, J.-L Gauvain, and L. Lamel. Genericity and portability for task-independent speech recognition. *Computer Speech and Language*, Vol. 19, pp. 345–363, 2005.

D. Hakkani-Tur, G. Riccardi, and A. Gorin. Active learning for automatic speech recognition. *Proc. ICASSP*, pp. 3904–3907, 2002.

T.M. Kamm and G.G.L. Meyer. Robustness aspects of active learning for acoustic modeling. *Proc. ICSLP*, pp. 1095–1098, 2004.

M.J.F. Gales and P.C. Woodland. Mean and variance adaptation within the MLLR framework. *Computer Speech and Language*, Vol. 10, No. 4, pp. 249–264, 1996.

R. Nisimura, Y. Nishihara, R. Tsurumi, A. Lee, H. Saruwatari, K. Shikano. Takemaru-kun: speech-oriented information system for real world research platform. *Proc. Proc. International Workshop on Language Understanding and Agents for Real World Interaction*, pp. 70–78, 2003.

A. Lee, T. Kawahara, K. Takeda, and K. Shikano. A new phonetic tied-mixture model for efficient decoding. *Proc. ICASSP*, pp. 1269–1272, 2000.

K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. shikano, and S. Itahashi. JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. *The Journal of the Acoustical Society of Japan(E)*, Vol. 20, pp. 199–206, 1999.

A. Lee, T. Kawahara, and K. Shikano. Julius - an open source real-time large vocabulary recognition engine. *Proc. Eurospeech2001*, pp. 1691–1694, 2001.

A. Lee, T. Kawahara, and K. Shikano. Real-time word confidence scoring using local posterior probabilities on tree trellis search. *Proc. ICASSP*, pp. 793–796, 2004.