論文 / 著書情報
Article / Book Information

Title	MDL-Based Cluster Number Decision Methods for Speaker Clustering and MLLR Adaptation
Authors	Zhipeng Zhang, Sadaoki Furui
Citation	ISCA Workshop on Adaptation Methods for Speech Recognition (ITR2001), Vol. , No. , pp. 41-44
Pub. date	2001, 9

MDL-Based Cluster Number Decision Methods for Speaker Clustering and MLLR Adaptation

Zhipeng Zhang and Sadaoki Furui

Tokyo Institute of Technology, Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
{zzp, furui}@furui.cs.titech.ac.jp

ABSTRACT

Speaker clustering is one of the major methods for speaker adaptation. MLLR (Maximum Likelihood Linear Regression) adaptation using transformation matrices corresponding to phone classes/clusters is another useful method especially when the length of utterances for adaptation is limited. In these methods, how to decide the most appropriate number of clusters is an important research issue. This paper proposes to use the MDL (Minimum Description Length) criterion to decide the optimum number of speaker clusters as well as phone clusters according to the size of utterances for clustering and adaptation. Experimental results of speaker clustering and MLLR-based speaker adaptation show that the MDL criterion derives the optimum number of clusters according to the size of utterances, which achieves the highest recognition performance.

1. INTRODUCTION

In large-vocabulary continuous-speech recognition, it is usually unrealistic to collect a sufficiently large set of training data for each speaker. Therefore, speaker adaptation of phone HMM is indispensable to achieve high recognition accuracy. We previously proposed an incremental speaker adaptation method in which an initial model for speaker adaptation was selected from a set of models made by speaker clustering [1]. In the adaptation phase, a cluster-dependent model that achieved the largest likelihood was selected and then MLLR (Maximum Likelihood Linear Regression) [2] was performed based on phone classes. This method was very effective to improve the recognition accuracy especially when the length of the utterance for speaker adaptation was short.

In these experiments, the optimum number of clusters for both speaker clustering and MLLR adaptation was determined experimentally. The system performance was largely dependent on the number of clusters for both speaker clustering and MLLR adaptation. The performance decreased if the number of clusters was too small and there was also a steep degradation if the number of clusters was too large. The optimum number of clusters depended on the size of data used for clustering and adaptation, respectively. Hence, how to decide the optimum number of clusters for both speaker clustering and MLLR adaptation is an important issue.

To decide the optimum number of clusters, we propose to use the MDL (Minimum Description Length) [3] criterion that has been proven to be effective in selecting the optimum model among various statistical models. In the case of speaker clustering, the description length (DL) according to each number of clusters is calculated and the number of clusters giving the shortest DL value is chosen. Once the initial model is chosen from a set of speaker-cluster-dependent HMMs, MLLR adaptation is applied to improve the performance. The MDL criterion is also used to decide the optimum number of MLLR transformation matrices according to the amount of adaptation data. These methods are tested with large-vocabulary continuous-speech recognition experiments using a spontaneous speech corpus.

Section 2 describes the methods of optimum number decision for speaker clustering and MLLR adaptation. Section 3 represents the results on spontaneous speech recognition experiments conducted to evaluate the performance of the MDL criterion to decide the optimum number of clusters for speaker clustering and MLLR adaptation. The paper concludes with discussion and future research issues.

2. MDL CRITERION FOR SPEAKER CLUSTERING AND MLLR ADAPTATION

2.1 Speaker clustering

Since directly clustering speech data uttered by many speakers needs huge amount of computation, we cluster speaker-dependent models rather than directly clustering speech data. Speaker-dependent phone HMM (SD-HMM) is first built based on the Baum-Welch algorithm for each speaker using a large set of speech data uttered by the speaker. Since all the SD-HMMs for different speakers have the same number of states, the distance between them can be defined with the following equation:

$$d(i,j) = -\frac{1}{S \times M} \sum_{s=1}^{S} \sum_{m=1}^{M} \{\log[b_{js}(\mu_{ism})] + \log[b_{is}(\mu_{jsm})]\} \qquad (1)$$

where b_{is} represents the output probability distribution function of the m-th mixture in the S-th state, and S and M represent the number of states and the number of mixtures in each state, respectively. In the distance calculation, only the mean values of distributions are taken into account. The differences of covariance, transition probabilities and weights are neglected.

Based on a distance matrix containing the distances between each SD-HMMs, a clustering procedure originally proposed for the "SPLIT" speech recognition system [4] is carried out to cluster speakers. After clustering, phone HMM for each cluster (Cluster-HMM) is made using all the utterances belonging to the cluster. Each Cluster-HMM is trained using the MAP (Maximum A Posteriori probability) [5] method.

2.2 MDL Criterion

MDL criterion has been proven to be effective in selecting the optimum model among various statistical models. It selects the model with the minimum DL for given amount of data. When a set of models $\{1,\ldots,i,\ldots,I\}$ is given, DL $l^{(i)}$ for data $X=\{x_1,\ldots,x_N\}$ and a model i is defined by:

$$l^{(i)} = -\log P_{\hat{\theta}^{(i)}}(X) + \frac{\alpha_i}{2}\log N + \log I \qquad (2)$$

where α_i is the number of free parameters of the model i, and $\hat{\theta}^{(i)}$ represents the estimated value of the parameter $\theta^{(i)}$ of the model i. The first term on the right-hand side represents the negative of the log likelihood. The second term is related to the complexity of model i and the number of data samples N. The third term is assumed to be constant. As the model becomes more complex, the value of the first term decreases and the second term increases. When comparing various models, the model with the shortest DL is considered to have the most appropriate size and complexity.

2.3 MDL-based number of cluster decision for speaker clustering

We apply the MDL criterion to decide the optimum number of speaker clusters. The DL can be written as:

$$l^{(i)} = -\log P_{\hat{\theta}^{(i)}}(X) + \frac{p\times C\times S\times n}{2}\log N + \log I \qquad (3)$$

where C is the number of clusters, S is the number of states, n is the number of vector dimensions, and p is a penalty parameter. The DL for each number of clusters is calculated and the number of clusters giving the minimum DL value is chosen.

2.4 MDL-based number of cluster decision for MLLR

We apply the MDL criterion to the MLLR-based speaker adaptation using phone classes/clusters. The Gaussian mean parameters of each HMM are updated according to the following equation:

$$\hat{\mu} = A\mu + b \qquad (4)$$

where A is an $n\times n$ matrix and b is an n-dimensional vector. The number of free parameters for MLLR transformation can be considered to be $n\times(n+1)$ and, therefore, the DL can be defined by:

$$l^{(i)} = -\log P_{\hat{\theta}^{(i)}}(X) + \frac{p\times C\times n\times(n+1)}{2}\log N + \log I \qquad (5)$$

where C is the number of clusters, n is the number of vector dimensions, and p is a penalty parameter. The DL for each number of clusters is calculated and the number of cluster giving the minimum DL value is chosen.

3. EXPERIMENTS ON A JAPANESE SPONTANOUS SPEECH RECOGNITION SYSTEM

3.1 Language models

A national project, entitled “Spontaneous Speech: Corpus and Processing Technology” was started in 1999 organized by the Science and Technology Agency of Japan under the supervision of Furui [6]. This project will be conducted over a 5-year period in pursuit of the following three major goals: (1) building a large-scale spontaneous speech corpus mainly consisting of presentation utterances, (2) acoustic and linguistic modeling for spontaneous speech understanding and summarization, and (3) constructing a prototype for automatic speech summarization.

Speech recognition experiments for evaluating the MDL-based method proposed in this paper were conducted using a part of the corpus built in the national project. The training set consisted of 610 presentations: 274 academic conference presentations and 336 simulated presentations. The simulated presentations talking about a wide variety of topics including the subjects’ experiences in their daily lives were specially recorded for the project.

3.2 Acoustic models

The 25 elements feature vector extracted from the speech samples consisted of 12 MFCC parameters, their delta features (derivatives) and normalized logarithmic power. The MFCC parameters were normalized by the cepstral mean subtraction (CMS) method.

Speech data consisting of 111,177 utterances (approximately 59 hours) uttered by 338 male speakers were used to train the speaker-independent (SI) HMM. The SI-HMM was gender-dependent shared-state triphone HMM and was designed using the tree-based clustering. The number of states in the model was 3000 and there were 16 Gaussian mixtures in each state.

3.3 Evaluation data

Speech data consisting of 150 presentation sentences uttered by 3 male speakers were extracted as a test set. The total duration

of the utterances from each speaker was respectively 6,7 and 6 minutes.

3.4 Experimental results of optimum number of cluster decision in speaker clustering

First, a cluster-selection-based speaker adaptation experiment was performed. The total number of utterances for constructing the cluster-dependent HMM was 21,704 (approximately 15 hours). It was about 1/4 of the data from 338 speakers that were used to train the SI-HMM. A test utterance was decoded using the SI-HMM and a phone label sequence was produced. Phone HMMs for each cluster-dependent HMM (Cluster-HMM) were concatenated according to the label sequence and used for re-calculating the likelihood. The best Cluster-HMM yielding the maximum likelihood was selected and used to re-recognize the input utterance. Figure 1 shows the DL value as a function of the number of clusters. Figure 2 shows the mean word recognition accuracy for the test set according to the number of clusters. The highest accuracy is obtained when the number of clusters is 4, which corresponds to the condition derived by the MDL criterion. The penalty p in the equation for calculating DL was set at 1.1 in the series of experiment in this section.

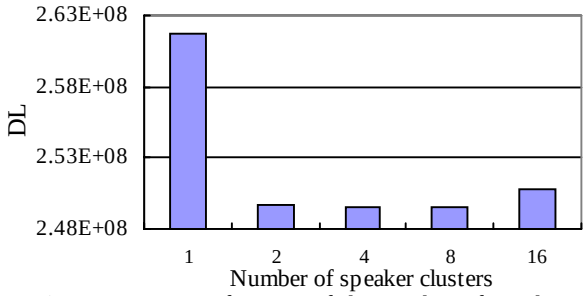

Figure 1: DL as a function of the number of speaker clusters (Number of utterances for clustering: 21,704).

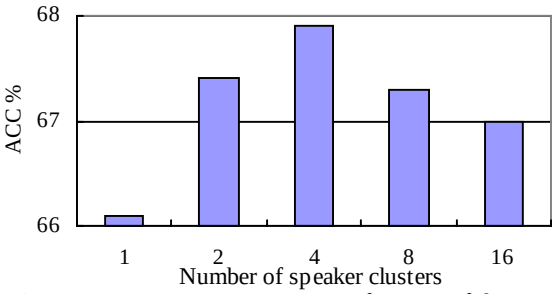

Figure 2: Recognition accuracy as a function of the number of speaker clusters (Same condition as Figure 1).

Then we increased the amount of data for constructing the cluster-dependent model to 50,303 utterances (approximately 35 hours). It was about 3/5 of the data from 338 speakers that were used to train the SI-HMM. Figure 3 shows the DL value for each number of clusters. Figure 4 shows the word accuracy for the test set according to the number of clusters. The highest accuracy is obtained when the number of clusters is 16, which corresponds well to the optimum condition derived by the MDL.

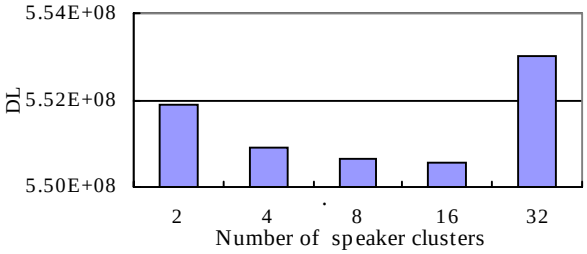

Figure 3: DL as a function of the number of speaker clusters (Number of utterances for clustering: 50,303)

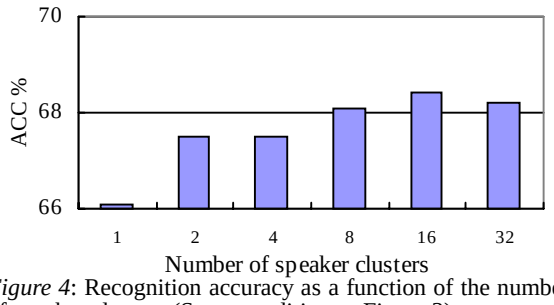

Figure 4: Recognition accuracy as a function of the number of speaker clusters (Same condition as Figure 3).

3.5 Experimental results of optimum number of cluster decision for MLLR adaptation

The next experiment was performed to evaluate the effectiveness of the MDL-based phone class number decision for improving the MLLR-based adaptation. Supervised MLLR adaptation was performed using 50 utterances for each speaker. Figure 5 shows the DL for each number of classes, and Figure 6 shows the mean word accuracy for the test set. The opti-

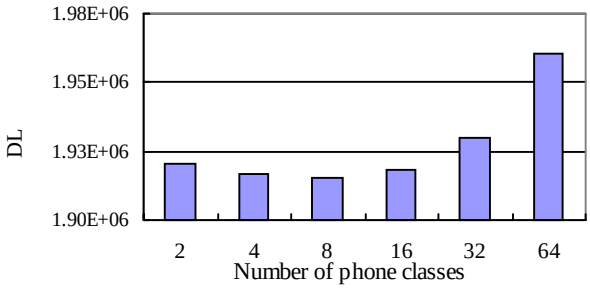

Figure 5: DL as a function of the number of phone classes (Number of utterances for speaker adaptation: 50, nitial model: speaker-independent model).

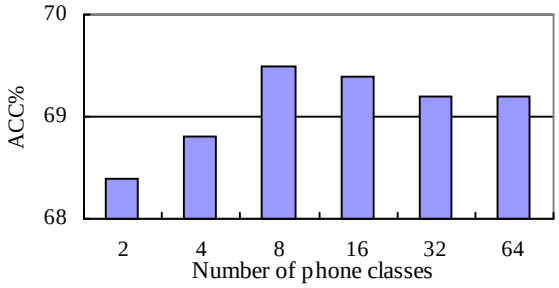

Figure 6: Recognition accuracy as a function of the number of phone classes (Same condition as Figure 5).

mum number of classes is 8 and it corresponds to the condition derived by the MDL criterion. The penalty *p* was set at 1 for the experiments in this section.

As the next step for improving the performance, we used the Cluster-HMMs instead of the SI-HMM as an initial model for MLLR adaptation. Supervised MLLR adaptation was performed using the same adaptation data as the previous experiment for each speaker. Figure 7 shows the DL value of each phone class number and Figure 8 shows the mean word accuracy for the test set. Comparing with the result given in Figure 6, it is observed that cluster-specific initial HMM is better than the SI-HMM for the MLLR adaptation. The former method reduces word error rates by 11.5% relative to the baseline using the SI-HMM for recognition. The results also show that the MDL method derives number of clusters that produced only slightly less than optimum recognition.

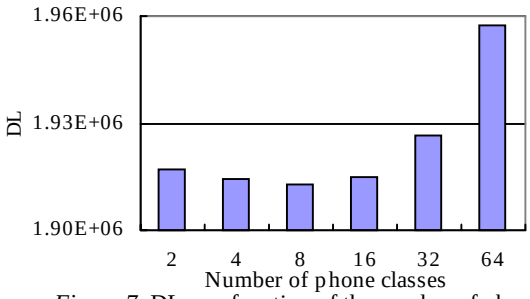

Figure 7: DL as a function of the number of phone classes (Number of utterances for peaker adaptation: 50, initial model: speaker-cluster model).

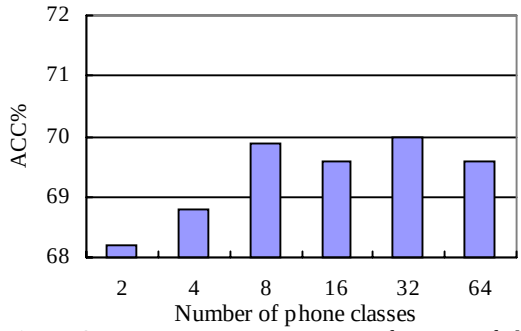

Figure 8: Recognition accuracy as a function of the number of phone classes (Same condition as Figure 7).

We then reduced the amount of adaptation data for each speaker to 10 utterances. Figure 9 shows the DL value of each class number and Figure 10 shows the mean word accuracy for each class number condition. It is shown that the optimum number of classes is 4 and the MDL method derives number of clusters that produced only slightly less than optimum recognition.

4. CONCLUSION

An MDL-based optimum number of cluster decision method has been investigated in the framework of speaker adaptation using initial models made by speaker clustering for MLLR adaptation with phone classes. Experimental results show that

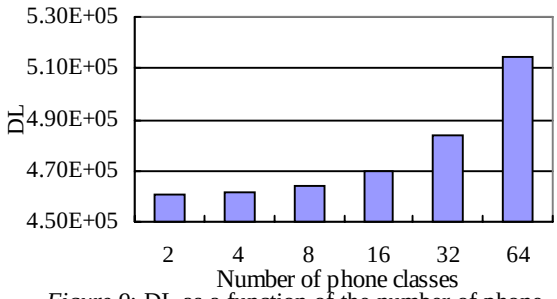

Figure 9: DL as a function of the number of phone classes (Number of utterances for speaker adaptation: 10, initial model: speaker-cluster model).

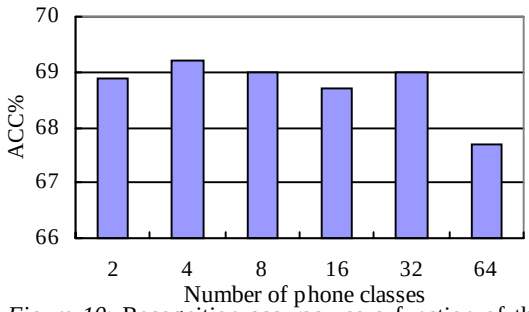

Figure 10: Recognition accuracy as a function of the number of phone classes (Same condition as Figure 9).

the proposed method is effective to decide the number of speaker clusters and phone classes according to the size of utterances for clustering and adaptation, respectively. The method of combining the speaker clustering and the MLLR adaptation reduced WER by 11.5% relative to the baseline using speaker-independent HMM.

Future research includes constructing more precise cluster-specific models and more general adaptation methods that can contend with not only speaker-to-speaker variability but also noise and channel variations to make recognition systems more robust against a wider range of mismatch conditions.

ACKNOWLEDGMENTS

The authors would like to thank to the members of “Spontaneous Speech” national project for constructing the corpus used in this investigation.

REFERENCES

[1] Z.P. Zhang et al., Proc of ICSLP, pp.694-697, 2000

[2] C.J. Leggetter et al., Computer Speech and Language, Vol.9, No.2, pp. 171-185, 1995.

[3] J. Rissanen, IEEE Transactions on IT, Vol.30, No.4, pp.629-636, 1984

[4] N. Sugamura et al., Proc of Speech Tech. Committee Meeting ASJ, pp.64-65, 1982

[5] J.L. Gauvain et al., IEEE Transactions on SAP, Vol.2 No.2, pp.291-298, 1994

[6] S. Furui et al., Proc of ICSLP, pp. 518-521, 2000