Speaker Model Selection Using Bayesian Information Criterion for Speaker Indexing and Speaker Adaptation

Masafumi Nishida[†] *and Tatsuya Kawahara*[† ‡]

[†] PRESTO, Japan Science and Technology Corporation (JST)
[‡] School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
{nishida, kawahara}@ar.media.kyoto-u.ac.jp

Abstract

This paper addresses unsupervised speaker indexing for discussion audio archives. We propose a flexible framework that selects an optimal speaker model (GMM or VQ) based on the Bayesian Information Criterion (BIC) according to input utterances. The framework makes it possible to use a discrete model when the data is sparse, and to seamlessly switch to a continuous model after a large cluster is obtained. The speaker indexing is also applied and evaluated at automatic speech recognition of discussions by adapting a speaker-independent acoustic model to each participant. It is demonstrated that indexing with our method is sufficiently accurate for the speaker adaptation.

1. Introduction

We are studying unsupervised speaker indexing aiming at audio archiving of discussions and meetings. Speaker index is essential for retrieving the utterances of a specific speaker and also for improving automatic speech recognition performance based on speaker adaptation of the acoustic model.

Recently, speaker indexing has been studied mainly for voice mails [1] and Switchboard conversations [2]. In these tasks, the duration of a utterance is 10 seconds or longer. Thus, speaker models are obtained by adapting the universal background model, and speaker identification is performed based on likelihood ratio between the adapted model and the background model. In discussions and meetings, there are many short utterances and the variation of utterance duration is large. Therefore, it is not feasible to use the adaptation scheme and apply a uniform model.

To the problem, we have proposed a flexible framework that selects an optimal speaker model (GMM or VQ) based on the Bayesian Information Criterion (BIC) [3]. Conventionally, GMM and VQ-based methods are used in speaker recognition. It is well known that the recognition performance of GMM is higher than that of VQ when there is much training data [4], but GMM cannot be trained with a small size of data. In our framework, an optimal speaker model (GMM or VQ) is selected based on the BIC which reflects the amount of speech data, and the speaker models are directly estimated without using an adaptation technique.

In this paper, we present detailed experimental results by comparing with GMM or VQ alone. It is also compared with a method that controls a number of Gaussian distributions of GMM according to the training data [5, 6].

We also address automatic speech recognition based on speaker adaptation using the speaker indexing result. A simple method is to adapt a speaker-independent model by MLLR (Maximum Likelihood Linear Regression) using utterances of each indexed speaker. In this paper, we introduce a method that combines it with speaker adaptation based on speakers selection [7, 8]. The method selects a subset of speakers used for speaker-independent model who are acoustically close to the test speaker and is used to generate an adapted model.

The methods are evaluated using actual discussion data.

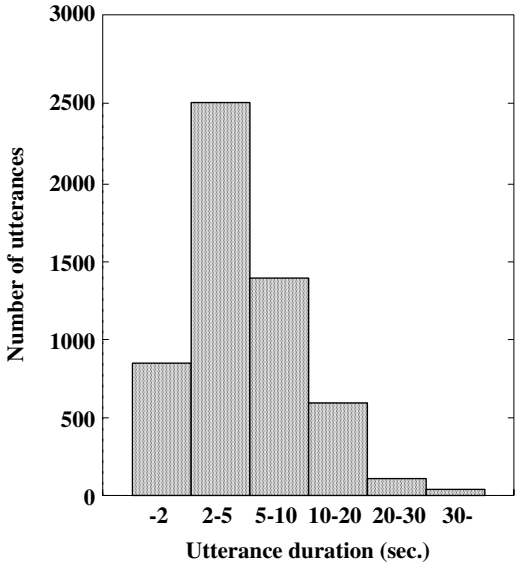

Figure 1: Distribution of utterance lengths

2. Database and Task

We use a one-hour forum for TV program that is broadcast on Sundays as the material for speaker indexing. In the program, politicians and journalists discuss the political and economic issues of Japan under the control of a moderator. For the test set, we picked 10 programs that were aired from June 2001 to January 2002.

The speech data is divided into segments based on energy and zero-crossing parameters, and the segments are regarded as utterances. For each discussion, there are 5 to 8 speakers with an average of 550 utterances. Fig. 1 shows the distribution of the duration of utterances. In Fig. 1, "$5-10$" shows the number of utterances of 5 to 10 seconds.

The average duration is 6 seconds, the minimum is 1 second,

and the maximum is 71 seconds. The utterances with durations less than 10 seconds occupy about 87% of the data. There are quite a number of short utterances and the variation of the duration is large. Therefore, an optimal model of suitable complexity should be selected depending on the data size.

3. Gaussian Mixture Size Selection

One way to control the complexity of the model is to control the number of Gaussian distributions based on the BIC according to the training data. We call the method "GMSS (Gaussian Mixture Size Selection)". The BIC of the GMM for a speaker s is formulated with the following function,

$$BIC_M^{(s)} = \log P(X|\lambda_M^{(s)}) - \frac{1}{2}M(2d+1)\log N \quad (1)$$

where $\log P(X|\lambda_M^{(s)})$ is a log likelihood of the training data X by the GMM when the number of mixtures is M, d is the dimension of the acoustic feature, N is the number of frames of the training data.

The mixture size of the GMM is determined by evaluating the following difference.

$$\begin{aligned}\Delta BIC^{(s)} &= BIC_M^{(s)} - BIC_{2M}^{(s)} \\ &= \log P(X|\lambda_M^{(s)}) - \log P(X|\lambda_{2M}^{(s)}) \\ &\quad + \frac{1}{2}M(2d+1)\log N \end{aligned} \quad (2)$$

The mixture size is incremented by double if $\Delta BIC^{(s)}$ is positive. Otherwise, it is determined to M.

If the training data is sparse, the mixture size is expected to be small or only one. But in that case, speaker information may not be fully represented.

4. Speaker Model Selection

Another way to cope with the sparse training data is to adopt a discrete model. Actually, a simple VQ-based method, which uses the VQ distortion as a distance measure, performs better than GMM [4] when a little data is available. Thus, we propose a flexible framework in which an optimal speaker model (VQ or GMM) is automatically selected based on the BIC according to the training data. We call the method "SMS (Speaker Model Selection)".

One problem in implementing this framework is that the model structure and distance measure are different for GMM and VQ. To solve the problem, we introduce a model called "CVGMM (Common Variance GMM)" that is an extension of VQ. CVGMM is modeled by assigning the same weights and covariances of the Gaussians to all mixture components. It realizes a normalization of the distance measure of VQ, so that it can be compared to the likelihood of GMM.

We first estimate a mixture of GMM as a speaker model. Then, we replace the covariance to generate the CVGMM and compute the BIC value for GMM and CVGMM. Specifically, the BIC for GMM and CVGMM of a speaker s is given by the followings, respectively.

$$BIC_{GMM}^{(s)} = \log P(X|\lambda_{GMM}^{(s)}) - \frac{1}{2}M(2d+1)\log N \quad (3)$$

$$\begin{aligned}BIC_{CVGMM}^{(s)} &= \log P(X|\lambda_{CVGMM}^{(s)}) \\ &\quad - \frac{1}{2}(M+1)d\log N \end{aligned} \quad (4)$$

Here, the mixture weights of CVGMM are assigned as $w_{CVGMM} = 1/M$ uniformly. The covariance of CVGMM is given by the average of the covariances of GMMs trained for all clusters.

$$\Sigma_{CVGMM} = \frac{1}{M \cdot S}\sum_{i=1}^{S}\sum_{j=1}^{M}\Sigma_{GMM_j}^{(i)} \quad (5)$$

Here, S is the number of clusters.

If the training data is sparse (i.e. duration is short), CVGMM is expected to be selected because GMM and CVGMM give comparable likelihoods and the model complexity of CVGMM is smaller. The method can dynamically change the model structure according to the data size. Thus, the appropriate speaker model can be constructed for any lengths of utterances.

5. Speaker Indexing Method

Speaker indexing is performed by training and incrementally merging speaker models. The procedure is described as follows:

1. Training: For each cluster, the speaker model is trained. In the initial training, each utterance makes one cluster.
2. Distance computation: The distance between clusters is computed based on the Cross Likelihood Ratio [9]. The Cross Likelihood Ratio d_{ij} for cluster i and cluster j is given by

$$d_{ij} = \log\frac{P(X_i|\lambda_i)}{P(X_i|\lambda_j)} + \log\frac{P(X_j|\lambda_j)}{P(X_j|\lambda_i)} \quad (6)$$

$$\log P(X_i|\lambda_j) = \frac{1}{n_i}\sum_{k=1}^{n_i}\log P(x_{ik}|\lambda_j)$$

   where X_i is all utterances of a cluster i, x_{ik} is the k-th utterance, n_i is the number of utterances and λ_i is the selected model.
3. Cluster merging with cross identification: For each cluster, the closest cluster whose distance is minimum is found and if the closest one of two clusters are each other, they are merged.

   Step 1, 2 and 3 are repeated until no more clusters can be merged.
4. Cluster merging with cross verification: The minimum distance among clusters is computed and if it is smaller than a threshold θ, these two clusters are merged.

   Step 2 and 4 are repeated until distances for all cluster pairs are larger than the threshold θ.

After sufficient training data is obtained for each cluster by the first merging procedure (Step 3), the training procedure (Step 1) is not performed for efficiency. Then, speaker clustering based on the matching likelihood is performed (Step 4).

6. Speaker Indexing Experiments

6.1. Experimental Condition and Evaluation Measure

All ten discussion data described in Section 2 are used in the experiments. The speech data is sampled at 16 kHz and the acoustic features consist of 26 components of 12 MFCCs, energy and their deltas. We compared our method (SMS) with the conventional method: the VQ-based method and the GMM-based method including GMSS. VQ and GMM are same as those

Table 1: Speaker indexing result

	Speaker indexing accuracy	Speaker number accuracy
VQ		
(4 cb)	63.2	66.7
(8 cb)	87.7	84.2
(16 cb)	92.2	93.0
(32 cb)	91.2	87.7
GMM		
(4 mix)	72.5	80.7
(8 mix)	93.6	87.7
(16 mix)	95.7	91.2
(32 mix)	93.3	91.2
GMSS	91.0	89.5
SMS		
(4 mix)	72.5	78.9
(8 mix)	93.6	78.9
(16 mix)	96.3	89.5
(32 mix)	97.2	93.0

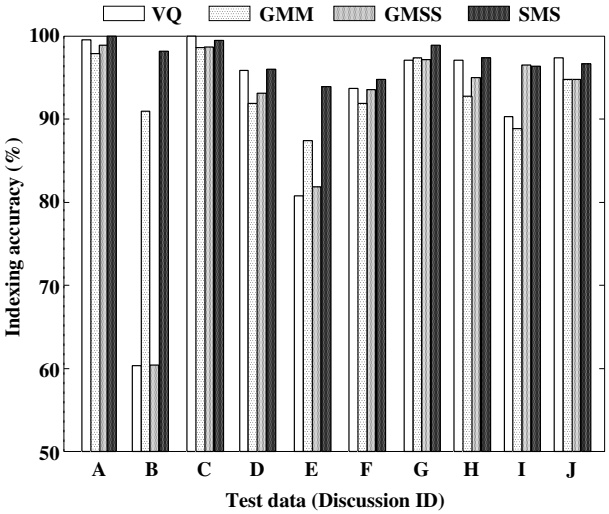

Figure 2: Indexing accuracy for each discussion

used in the proposed method, but we assume it is selected for all clusters.

We made evaluation using a speaker indexing accuracy and an accuracy of the number of speakers. The speaker indexing accuracy is defined as the ratio of the BBN metric [10] by the automatic indexing method and that by the correct indexing. It becomes 0 at the worst and 1 at the best case. The accuracy of the number of speakers is defined as,

$$SA = \{1 - \frac{\sum_{i=1}^{D} |S_i - C_i|}{\sum_{i=1}^{D} S_i}\} \times 100 \quad (7)$$

where S_i is the actual number of speakers and C_i is the number of clusters in the i-th discussion, D is the number of total discussions.

6.2. Experimental Results

The average indexing performance obtained by the described methods is shown in Table 1. The threshold θ of the speaker clustering procedure (Step 4) is determined so that the accuracy of the number of speakers gets the maximum in each case. It is less sensitive in the proposed method because likelihoods for speaker models are stable by appropriately choosing variances of Gaussian distributions according to the training data.

The indexing performance for individual discussions is shown in Fig. 2. In Figs. 2, "VQ", "GMM" and "SMS" denote the result when the size of mixtures or codebooks is 32.

The proposed SMS method achieved an accuracy of 97.2% in indexing and 93.0% in estimation of the number of speakers when the number of mixtures was 32. It outperforms the VQ-based method and the GMM-based method. It achieved the best performance over almost all data. It is also verified that the discrete VQ is chosen when the utterances are short, and the stochastic GMM is chosen for large clusters.

For the GMM-based method, it gets harder to estimate large mixtures with the data because there are so many short utterances for which variances of some mixture components becomes too small, which cause false matching. So the clusters of same speakers are not correctly merged.

The GMSS method that adaptively controls the mixture size did not perform better. The method selects single Gaussian distribution that is poor representation than VQ, and most of very short utterances are incorrectly clustered.

With the VQ-based method, it is possible to train the stable model even for a little training data. However, it does not represent the speaker information after sufficient size of clusters is obtained. Actually, the GMM-based method achieves better performance when the number of mixtures and codebooks is same.

7. ASR based on Speaker Adaptation Using Speaker Indexing

Then, we perform automatic speech recognition (ASR) based on speaker adaptation using the indexing result by the proposed SMS method (32-mix.).

For adaptation using the indexing result, there is a simple method that adapts a speaker-independent acoustic model by MLLR (Maximum Likelihood Linear Regression) with utterances of each indexed speaker. In the speaker-independent model, however, a variation of speakers is large and all speakers are not necessarily matched to the test speaker. Therefore, adaptation methods based on speakers selection or clustering have been studied [7, 8]. In this approach, a subset of speakers who are acoustically close to the test speaker is selected and an adapted model for the selected speakers is used in automatic speech recognition.

In this paper, we investigate the combination of speaker indexing and speaker adaptation based on speakers selection in order to obtain the acoustic model that well represents speaker information of each participant of discussions.

The procedure is described as follows:

1. Training: Each speaker used for training the speaker-independent acoustic model is modeled by GMM of 64 mixtures.
2. Speakers selection: For each indexed speaker of the test data, the likelihoods are calculated by GMMs of training speakers and speakers with higher likelihoods are selected.
3. Adaptation 1: The speaker-independent model is adapted by MLLR with the training data of the selected speak-

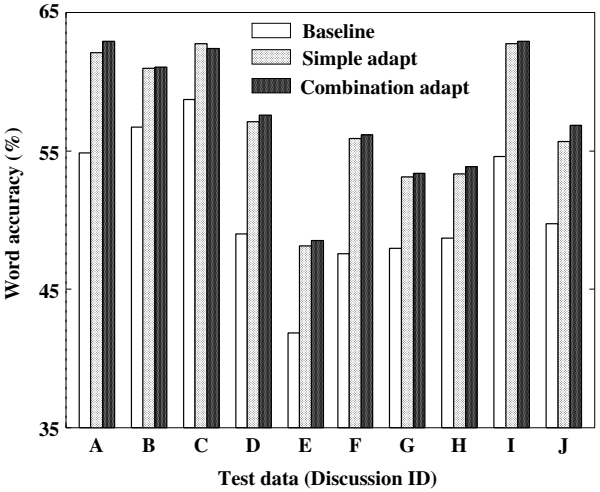

Figure 3: Automatic speech recognition result

ers.

4. Adaptation 2: Then, the model is adapted using the utterances of each indexed speaker by MLLR to generate the adapted model used in ASR.

The baseline acoustic model is a phonetic tied-mixture triphone HMM (3000 states and 16K Gaussians) trained with the Corpus of Spontaneous Japanese (CSJ) [11]. The language model is a weighted combination of lecture-based and minutes-based models [12]. We used our Julius 3.3 [13] decoder. There are 381 training speakers for the speaker-independent acoustic model.

The word accuracy for each discussion is shown in Fig. 3. Here, "Baseline" denotes the case using the baseline model without adaptation. "Simple adapt" denotes the unsupervised adaptation using the speaker indexing result and initial ASR result with the baseline model, and "Combination adapt" denotes the combinated adaptation method of speaker indexing and speakers selection.

With the baseline model, the accuracy was 51.0% on average. The simple adaptation method improved it to 57.2%. This demonstrates that the unsupervised speaker adaptation based on the speaker indexing is very effective. For reference, we also performed supervised adaptation using correct speaker labels (transcription is given by ASR) and the accuracy was 57.2%, which was comparable to the totally unsupervised case. The result demonstrates that the speaker indexing performance by the SMS is sufficient for adaptation of the acoustic model.

In the combinated method, the accuracy was 57.6% on average. It was improved for 35 of 57 speakers of the all discussions compared with the simple adaptation method. Some speakers such as chairmans appear at several discussions, but it is observed that different speakers are selected from the speaker-independent model at different discussions. This demonstrated that speakers selected by the likelihood of GMM are not necessarily optimal. Further investigations are needed on this issue.

8. Conclusions

We have presented a method that selects an optimal speaker model among VQ and GMM based on the BIC according to the input utterances. In speaker indexing of discussions, the proposed method achieved higher indexing performance than the conventional method that controls the Gaussian mixture size using the BIC. Moreover, speech recognition performance was improved by speaker adaptation combining the speaker indexing and speakers selection. It was also shown that the indexing accuracy by the proposed method is sufficiently high for adaptation of the acoustic model.

9. References

[1] D. Charlet, “Speaker Indexing for Retrieval of Voicemail Messages,” Proc. ICASSP, Vol. 1, pp. 121-124, 2002.

[2] S. Meignier, J. F. Bonastre, and I. M. Chagnolleau, “Speaker Utterances Tying Among Speaker Segmented Audio Documents Using Hierarchical Classification: Towards Speaker Indexing of Audio Databases,” Proc. ICSLP, pp. 577-580, 2002.

[3] M. Nishida and T. Kawahara, “Unsupervised Speaker Indexing Using Speaker Model Selection based on Bayesian Information Criterion,” Proc. ICASSP, Vol. 1, pp. 172-175, 2003.

[4] T. Matsui and S. Furui, “Comparison of Text Independent Speaker Recognition Methods Using VQ Distortion and Discrete/Continuous HMMs,” Proc. ICASSP, Vol. 2, pp. 157-160, 1992.

[5] K. Shinoda and T. Watanabe, “Acoustic Modeling based on the MDL Criterion for Speech Recognition,” Proc. EUROSPEECH, Vol. 1, pp. 99-102, 1997.

[6] S. S. Chen and R. A. Gopinath, “Model Selection in Acoustic Modeling,” Proc. EUROSPEECH, Vol. 3, pp. 1087-1090, 1999.

[7] Y. Gao, M. Padmanabhan, and M. Pichey, “Speaker Adaptation based on Pre-clustering Training Speakers,” Proc. EUROSPEECH, Vol. 4, pp. 2091-2094, 1997.

[8] S. Yoshizawa, A. Baba, K. Matsunami, Y. Mera, M. Yamada, and K. Shikano, “Unsupervised Speaker Adaptation based on Sufficient HMM Statistics of Selected Speakers,” Proc. ICASSP, Vol. 1, pp. 337-340, 2001.

[9] D. A. Reynolds, E. Singer, B. A. Carlson, G. C. O’Leary, J. J. McLaughlin, and M. A. Zissman, “Blind Clustering of Speech Utterances based on Speaker and Language Characteristics,” Proc. ICSLP, pp. 3193-3196, 1998.

[10] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, “Clustering Speakers by Their Voices,” Proc. ICASSP, pp. 757-760, 1998.

[11] H. Nanjo and T. Kawahara", “Speaking-rate Dependent Decoding and Adaptation for Spontaneous Lecture Speech Recognition,” Proc. ICASSP, pp. 725-728, 2002.

[12] Y. Akita, M. Nishida, and T. Kawahara, “Automatic Transcription of Discussions Using Unsupervised Speaker Indexing,” Proc. SSPR, pp. 79-82, 2003.

[13] A. Lee, T. Kawahara, and K. Shikano, “Julius — an Open Source Real-Time Large Vocabulary Recognition Engine,” Proc. EUROSPEECH, pp. 1691-1694, 2001.