SPEAKER INDEXING AND ADAPTATION USING SPEAKER CLUSTERING BASED ON STATISTICAL MODEL SELECTION

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

[†] Graduate School of Science and Technology, Chiba University
Inage-ku, Chiba 263–8522, Japan
[‡] School of Informatics, Kyoto University
PRESTO, Japan Science and Technology Corporation (JST)
Sakyo-ku, Kyoto 606-8501, Japan
nishida@faculty.chiba-u.jp, kawahara@i.kyoto-u.ac.jp

ABSTRACT

This paper addresses unsupervised speaker indexing and automatic speech recognition of discussions. In speaker indexing, there are two cases, where the number of speakers is unknown and known beforehand. When the specified number is unknown, it is difficult to apply to various data because it needs to determine several parameters like threshold. In addition, serious problems arise in applying a uniform model because variations in the utterance durations of speakers are large. We thus propose a method which can robustly perform speaker indexing for the two cases using a flexible framework in which an optimal speaker model (GMM or VQ) is selected based on the BIC. Moreover, we propose a combination method of speaker adaptation based on speaker selection and the indexing method. For real discussion archives, we demonstrated that indexing performance is higher than that of conventional methods for the two cases and speech recognition performance was improved by the combination method.

1. INTRODUCTION

Recently, speaker indexing has been studied mainly for voice mails [1] and Switchboard conversations [2]. In these tasks, the duration of an utterance is 10 seconds or longer. Thus, speaker models are obtained by adapting the universal background model, and speaker clustering is performed based on the likelihood ratio between the background model and the adapted model. In discussions and meetings, the utterance length of speakers is not fixed and there are a large number of short utterances as well as very long ones, which causes serious problems in applying a uniform model. Therefore, it is not feasible to use an adaptation scheme.

As an alternative approach to speaker clustering and detection of speaker changes, a method based on Bayesian Information Criterion (BIC) has been proposed [3]. The method assumes a single Gaussian distribution for each segment and performs a speaker clustering based on the variance ratio between segments. We call this method "Variance-BIC" because the likelihood is replaced by a variance. However, it is difficult to apply to various data, since it is necessary to determine a penalty weight in order to control the balance between variance and model complexity.

To the problem, we have proposed a flexible framework in which an optimal speaker model (GMM or VQ) is automatically selected based on the BIC [4]. Since the proposed framework has been applied to speaker indexing of discussions in the case where the number of speakers is unknown beforehand, it was necessary to determine a threshold of indexing. It may be possible to know the number of participants beforehand in discussions. Thus, we propose a method which can robustly perform speaker indexing in the both cases where the number of speakers is unknown and known. We carry out speaker indexing experiments for the two cases and demonstrate that the proposed method is robust and less sensitive to the threshold value for indexing.

We also address automatic speech recognition (ASR) based on speaker adaptation using the indexing result. For adaptation using the indexing result, there is a simple method that adapts a speaker-independent (SI) acoustic model by MLLR using the utterances of each indexed speaker. In the SI model, however, the variation of speakers is large and all speakers are not necessarily matched to the test speaker. Therefore, adaptation methods based on speaker selection or clustering have been studied [5, 6]. These methods perform speaker clustering for training speakers of the SI model and select a subset of speakers matched to the test speaker. However, the optimal speakers cannot necessarily be chosen in the case where there is an acoustical difference between the test data and the SI model. Thus, we propose a combination method of speaker adaptation based on speaker selection and the proposed indexing method in order to select the optimal speakers when the test data is acoustically different from the SI model.

The methods are compared and evaluated using actual discussion data.

2. STATISTICAL SPEAKER MODEL SELECTION

We explore a flexible framework in which an optimal speaker model (GMM or VQ) is automatically selected based on the BIC. We call this "Speaker Model Selection (SMS)".

One problem in implementing this framework in selecting the speaker model is that the model structure and distance measure are different for GMM and VQ. To solve this, we introduce a model called "Extended VQ (EVQ)". EVQ is modeled by assigning the same weights and covariances of Gaussians to all mixture components, and EVQ becomes a VQ model by replacing the covariance matrix with the identity matrix.

We first estimate a Gaussian mixture of GMM as a speaker model. Only the diagonal components of covariances are used. Specifically, the BIC of the GMM for speaker s is given by

$$BIC^{(s)}_{GMM} = \log P(X|\lambda^{(s)}_{GMM}) - \frac{1}{2}M(2d+1)\log N \qquad (1)$$

where $\log P(X|\lambda^{(s)}_{GMM})$ is a log likelihood of training data X obtained by GMM, M is the number of mixture components, d is the dimension of the acoustic feature, and N is the number of frames of training data.

We then generate the EVQ. The mixture weights of EVQ are uniformly assigned as $w_{EVQ} = 1/M$. Estimating the covariance of EVQ is very difficult for a speaker with a small amount of training data. Therefore, we replace it with the average covariances of GMMs trained for all speakers as follows.

$$\Sigma_{EVQ} = \frac{1}{M \cdot S} \sum_{i=1}^{S} \sum_{j=1}^{M} \Sigma^{(i)}_{GMM_j} \tag{2}$$

Here, S is the number of speakers. The BIC for the EVQ is given by Eq. (3).

$$BIC^{(s)}_{EVQ} = \log P(X|\lambda^{(s)}_{EVQ}) - \frac{1}{2}(M+1)d \log N \tag{3}$$

When the training data size is small, the VQ model will be selected because its complexity is much smaller. After a large amount of training data is obtained, GMM is expected to be selected because its likelihood is large.

3. SPEAKER INDEXING ALGORITHM

3.1. Speaker Indexing based on Variance-BIC

The conventional method of speaker indexing based on Variance-BIC is formulated as follows [3]. In this paper, one or more utterance units are called a segment. Initially, each utterance makes a segment. To decide if two consecutive segments are uttered by the same speaker, the difference in the BIC values is computed as below.

$$\begin{aligned} &\Delta BIC_{var} \\ &= -\frac{N_1+N_2}{2} \log|\Sigma_0| + \frac{N_1}{2} \log|\Sigma_1| + \frac{N_2}{2} \log|\Sigma_2| \\ &+\alpha \frac{1}{2}(d + \frac{d(d+1)}{2}) \log(N_1+N_2) \end{aligned} \tag{4}$$

Here, Σ_0 is the covariance of the merged segment, and Σ_1 and Σ_2 are those for the first and second segments, respectively. Full covariances are used. N_i represents the number of frames of respective segments, d is the dimension of the acoustic feature, and α is the penalty weight.

If ΔBIC_{var} is positive, the two segments are merged. Speaker clustering is performed by repeating the process. When the ΔBIC_{var} values between all segment pairs become negative, the clustering process is finished. As variations in the duration of utterances in the discussion data are large, hence, reliable estimation and fair comparison of variances are difficult especially for very short speech segments.

3.2. Speaker Indexing based on Speaker Model Selection

We propose a speaker indexing procedure based on the SMS scheme for two cases, where the number of speakers is both unknown and known in advance.

The detailed procedure is described as follows.

1. Training and model selection: The GMM and EVQ are trained for each cluster. In the initial step, each utterance forms one cluster. An optimal model is selected between GMM and EVQ for each cluster based on the BIC.
2. Distance calculation: The distance between clusters is computed based on the Cross Likelihood Ratio (CLR) as follows,

   $$d_{ij} = \log \frac{P(X_i|\lambda_i)}{P(X_i|\lambda_j)} + \log \frac{P(X_j|\lambda_j)}{P(X_j|\lambda_i)} \tag{5}$$

   where X_i is all utterances of cluster i, λ_i is the selected model (GMM or EVQ) for cluster i, and $\log P(X_i|\lambda_j)$ is the average log likelihood of utterances of cluster i obtained by model λ_j.
3. Merging clusters with cross identification: For each cluster, the closest cluster with the minimum distance is found and if the closest one of two clusters are the same (other) cluster, they are merged. Namely, merge clusters i and j if $argmin_k d_{ik} = argmin_k d_{jk}$ $(i \neq j \neq k)$.

   Steps 1, 2, and 3 are repeated until no more clusters can be merged.
4. Merging clusters with cross verification: The minimum distance between clusters is computed and if the distance is smaller than threshold θ, these two clusters are merged. Namely, merge clusters i and j if $d_{ij} < \theta$.

   Steps 2 and 4 are repeated until the distances for all cluster pairs are large than threshold θ. When the number of speakers is given, cluster merging (Step 4) continues until the number of obtained clusters reaches the specified number by disregarding threshold θ.

4. SPEAKER INDEXING EXPERIMENTS

4.1. Experimental Conditions and Evaluation Measure

We used a one-hour forum TV program as the material for speaker indexing experiments. During the program, politicians and journalists discuss Japanese political and economic issues under the control of a moderator. We selected ten programs that were aired from June 2001 to January 2002 for the test set.

The speech data was sampled at 16 kHz and the acoustic features consist of 26 components of 12 MFCCs, energy and their deltas. For each discussion, there were five to eight speakers with an average of 550 utterances. The total number of speakers was 57. The average duration was six seconds, the minimum was one second, and the maximum was 71 seconds. Utterances with durations of less than ten seconds represented about 87% of the data. There were quite a few short utterances and there were large variations in duration.

We compared our method (SMS) with conventional methods, i.e., the Variance-BIC, the VQ-based and the GMM-based methods. The VQ and GMM were the same as those used in the proposed method, but we assumed the model was uniformly selected for all clusters. We carried out the speaker indexing experiments for two cases, where the number of speakers was both unknown and known in advance.

We did evaluations using speaker indexing accuracy and accuracy on the number of speakers. Speaker indexing accuracy was defined as the ratio of the BBN metric [7] obtained by automatic indexing and that by correct indexing. Thus, this is given by,

$$SIA = \frac{\sum_{i=1}^{C} n_i p_i - QC}{n - QS} \times 100 \tag{6}$$

where n_i is the number of utterances in candidate cluster i, and C is the number of candidate clusters. $p_i = \sum_{j=1}^{S} (\frac{n_{ij}}{n_i})^2$ is the

purity of cluster i, and n_{ij} is the number of utterances by speaker j in cluster i. n is the total number of utterances, and S is the actual number of speakers. We set the system design parameter to $Q = 0.5$. Indexing performance increases with a larger value for the BBN metric. It became zero at worst and one at best. Accuracy on the number of speakers is defined as,

$$SNA = \{1 - \frac{\sum_{k=1}^{D} |S_k - C_k|}{\sum_{k=1}^{D} S_k}\} \times 100 \quad (7)$$

where S_k is the actual number of speakers, C_k is the number of obtained clusters in the k-th discussion, and D is the total number of discussions. This is used only when the number of speakers is unknown.

4.2. Experimental Results

We investigated the sensitivity of the speaker indexing accuracy to threshold θ of the speaker clustering procedure (Step 4). SIA has been plotted by changing threshold θ in Fig. 1. Also, Fig. 2 plots the SNA when threshold θ is changed for all cases. These graphs plot the results when the size of the mixtures or codebooks is 32. The penalty weight α in the Variance-BIC was set to 5.0 after the preliminary experiments.

The SMS method achieved SIA of 97.0% when SNA was maximum. It outperformed the Variance-BIC, the VQ-based, and the GMM-based methods. It achieved the best performance over almost all the data. For Variance-BIC method, the accuracies are plotted by changing the penalty weight α. Although the scale of α is different from that of θ, we see the accuracy is sensitive to this value, and the peaks of SIA and SNA are obtained at totally different values of α. SIA in the VQ-based method is less sensitive to variations in threshold θ. However, SNA changes with slight variations in threshold θ compared with the other methods. SMS maintains consistent SIA and SNA against variations in threshold θ. It is less sensitive because it can appropriately choose and reliably estimate speaker models according to the amount of training data.

The average indexing performance when the number of speakers is given in advance is shown in Table 1. SMS achieved SIA of 97.0% with 32-mixture and again outperformed the other methods. Indexing accuracy here was the same as where the number of speakers was unknown. This shows that we can obtain sufficiently high indexing by choosing optimal threshold θ and also that specifying the number of speakers has the same effect as using optimal threshold θ. This does not necessarily hold for other methods, e.g. VQ-based method with the codebook size of 32, because the best SIA was obtained where more than the actual numbers of speaker clusters were used.

5. SPEAKER ADAPTATION BASED ON SPEAKER SELECTION USING INDEXING RESULT

We explore a combination method of speaker indexing and speaker adaptation based on speaker selection. The indexing result by the proposed SMS method (32-mix.) is used for speaker adaptation.

We use the speaker-independent (SI) acoustic model trained with the Corpus of Spontaneous Japanese (CSJ) [8], which consists of lecture speech, because there is not enough discussion data to train the SI model. Therefore, it is difficult to select the optimal speakers by the conventional adaptation based on speaker selection since the test data is acoustically different from the SI model.

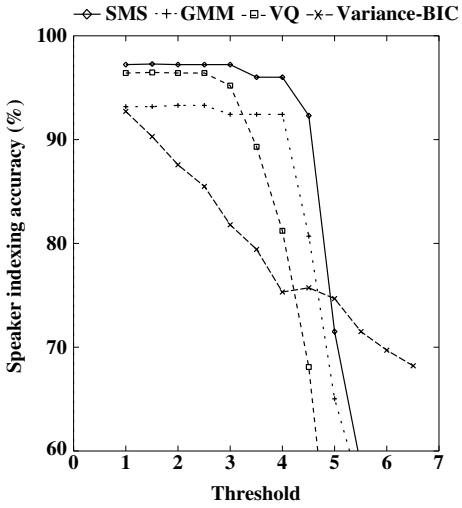

Fig. 1. Speaker indexing accuracy when varying the threshold θ

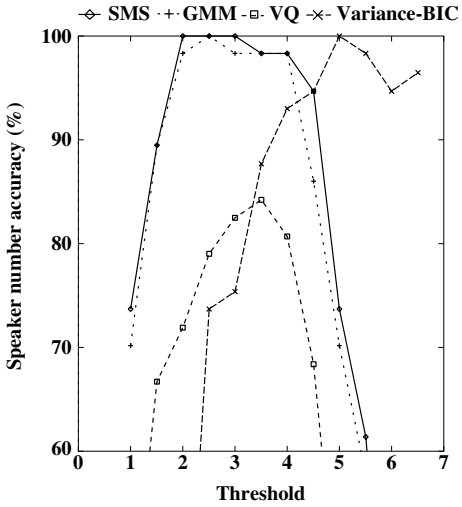

Fig. 2. Speaker number accuracy when varying the threshold θ

After speaker clustering is performed by the proposed indexing method for each discussion audio, we chooses a subset of speakers who are acoustically close to the test speaker from all discussion audios using the obtained speaker models. We adapts the SI model using the utterances of these speakers. This method does not need to newly train speaker models for the adaptation based on speaker selection and can treat the speaker indexing and adaptation process systematically.

The procedure is described as follows.

1. Speaker selection: For each indexed speaker of the test data, the CLRs were calculated by GMMs of indexed speakers and speakers with lower CLRs were selected. The CLR d_{ij} for indexed speakers i and j is given by

$$d_{ij} = \log \frac{P(\mu_i|\lambda_i)}{P(\mu_i|\lambda_j)} + \log \frac{P(\mu_j|\lambda_j)}{P(\mu_j|\lambda_i)} \quad (8)$$

where μ_i is a set of mean vectors of GMM for speaker i, λ_i is the selected model (GMM) for speaker i, and

Table 1. Indexing result when number of speakers is known

	Speaker Indexing Accuracy (%)
Variance-BIC	74.7
VQ (4 cb)	61.8
(8 cb)	82.2
(16 cb)	91.9
(32 cb)	94.4
GMM (4 mix)	66.8
(8 mix)	89.6
(16 mix)	91.3
(32 mix)	93.3
SMS (4 mix)	66.8
(8 mix)	89.4
(16 mix)	91.6
(32 mix)	97.0

Table 2. Automatic speech recognition result

	Word accuracy (%)
Baseline	51.0
Simple adaptation	57.2
Speaker selection adaptation (30 spks)	57.7
Combination adaptation (5 spks)	58.1
Supervised adaptation	59.4

$\log P(\mu_i|\lambda_j)$ is the average log likelihood of utterances of speaker i obtained by model λ_j. We compute the CLRs by using the mean vectors of GMM instead of the feature vectors of utterances because the processing cost is large when CLRs are calculated using the feature vectors of utterances.

2. Adaptation 1: The SI model is adapted by MLLR with the utterances of the selected indexed speakers.

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

The baseline acoustic model is a phonetic tied-mixture triphone HMM (3000 states and 16K Gaussians in total) trained with the CSJ. We use 43 phones, and all of them are modeled with the left-to-right HMM of three states. The training data consisted of spontaneous oral presentations by 381 speakers that amounted to 60 hours. The language model is a back-off word trigram, which is a weighted combination of a model trained with the CSJ and one constructed from the minutes taken at the National Diet of Japan [9]. There are 36,053 vocabulary items. We used our Julius 3.3 decoder [10] for recognition with these models.

The average word accuracy obtained by the described methods is shown in Table 2. Here, "Baseline" denotes the case using the baseline model without adaptation. "Simple adaptation" denotes unsupervised adaptation using the indexing result and initial ASR result with the baseline model. "Speaker selection adaptation" denotes unsupervised adaptation based on conventional speaker selection. "Combination adaptation" denotes unsupervised adaptation based on the combination method of speaker indexing and speaker selection. "Supervised adaptation" denotes supervised adaptation using correct speaker labels and phoneme transcriptions.

With the baseline model, the accuracy was 51.0% on average. The simple adaptation method improved accuracy to 57.2%. This demonstrates that unsupervised speaker adaptation based on speaker indexing is very effective. The accuracy achieved by supervised adaptation was 59.4%. In speaker selection adaptation, the best accuracy was 57.7% when 30 training speakers of the SI model were selected for each indexed speaker. In the combination method, the best accuracy was 58.1% when 5 speakers are selected from all discussion data for each indexed speaker. The proposed method chose fewer speakers than the conventional adaptation based on speaker selection and exhibited higher recognition performance than the conventional method. Therefore, we demonstrated that the optimal speakers matched to each indexed speaker were selected.

6. CONCLUSION

We presented a method which can robustly perform speaker indexing for two cases, where the number of speakers is unknown and known beforehand. For actual discussion archives, we demonstrated that the proposed method achieves higher indexing performance than conventional methods such as Variance-BIC, VQ-based and GMM-based methods for the two cases. We also found that the proposed method is less sensitive to the threshold value for clustering.

Moreover, we applied a combination method of speaker adaptation based on speaker selection and proposed indexing method. This method obtained the acoustic model more matched to each indexed speaker by choosing a subset of speakers who are acoustically close to the test speaker from all discussion audios using speaker models obtained by speaker indexing. Speech recognition performance was improved by the combination method.

7. 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] S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[4] 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.

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

[6] 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.

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

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

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

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