UNSUPERVISED SPEAKER INDEXING USING SPEAKER MODEL SELECTION BASED ON BAYESIAN INFORMATION CRITERION

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}@kuis.kyoto-u.ac.jp

ABSTRACT

This paper addresses unsupervised speaker indexing for discussion audio archives. In discussions, the speaker changes frequently, thus the duration of utterances is very short and its variation is large, which causes significant problems in applying conventional methods such as model adaptation and Variance-BIC (Bayesian Information Criterion) methods. We propose a flexible framework that selects an optimal speaker model (GMM or VQ) based on the BIC according to the duration of utterances. When the speech segment is short, the simple and robust VQ-based method is expected to be chosen, while GMM will be reliably trained for long segments. For a discussion archive having a total duration of 10 hours, it is demonstrated that the proposed method achieves higher indexing performance than that of conventional methods.

1. INTRODUCTION

Speaker indexing is useful for retrieving the utterances of a specific speaker and also for improving automatic speech recognition performance based on speaker adaptation of the acoustic model. In audio archives of discussions which we deal with in this paper, these two functions are very significant.

Speaker indexing is rather easy if we can train speaker models in advance. However, speakers are not always the same, and it is not practical to assume that speech samples for each speaker is available beforehand for many tasks including discussions. Therefore, we address unsupervised speaker indexing without prior speaker models. Moreover, we do not assume that the number of speakers is given.

Recently, speaker indexing has been studied mainly for voice mails [1], broadcast news, and Switchboard conversations [2]. In voice mail tasks, the duration of a message is 10 seconds or more. In the Switchboard Corpus, utterances in the telephone conversations have durations of 31 seconds on average and a minimum of 14 seconds [2]. In these tasks, speaker models are obtained by adapting the universal background model, and speaker clustering is performed based on the likelihood ratio between the adapted model and the background model [3]. In the discussion data we deal with, utterances have a duration of 6 seconds on average and the ratio of those utterances less than 10 seconds is about 85%. Therefore, it is not feasible to use adaptation techniques such as MLLR.

As an alternative approach for automatic speaker indexing and detection of speaker changes, a method based on BIC (Bayesian Information Criterion) [4] has been proposed. The method assumes a single Gaussian distribution for each segment and determines the number of clusters based on variances between segments. It is effective for broadcast news, where speech segments have long duration and the speaker change is not so frequent. In this paper, we call the method "Variance-BIC" because the likelihood is substituted by a variance. In discussion data, the variation in duration is much larger (minimum is 1 second and the maximum is 61 seconds), thus the comparison based on variances for such unbalanced data might not work. Moreover, the method assumes a single Gaussian distribution for each segment, and the speaker information may not be fully represented.

We propose a novel framework of model selection for speaker indexing. Conventionally, GMM [5] 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 a lot of training data [6], however, GMM cannot be estimated with a small size of data. In our framework, an optimal speaker model (GMM or VQ) is selected based on BIC which reflects the amount of speech data, and the speaker models are directly estimated without using an adaptation technique.

The methods are compared and evaluated using actual discussion data.

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 problems of Japan under the control of a moderator. For the test set, we picked 10 programs that were aired from June to December 2001.

Speaker indexing is performed based on utterance segmentation and speaker clustering. The speech data is divided into segments based on energy and zero-crossing parameters, and the segments are regarded as utterances. Table 1 shows the number of speakers and utterances in the discussions. Fig. 1 shows the distribution of the duration of utterances. In Fig. 1, "$X - Y$" shows the number of utterances of X to Y seconds.

The average duration is 6 seconds, the minimum is 1 second, and the maximum is 61 seconds. The utterances having durations less than 10 seconds occupy about 85% of the data. There are much greater number of short utterances and the variation in the duration is also larger. This causes a significant problem in applying a uniform model.

Table 1. Test set of discussion speech

	A	B	C	D	E
#Speaker	5	5	5	8	6
#Utterance	449	711	578	569	672
	F	G	H	I	J
#Speaker	8	5	8	5	5
#Utterance	367	281	340	554	557

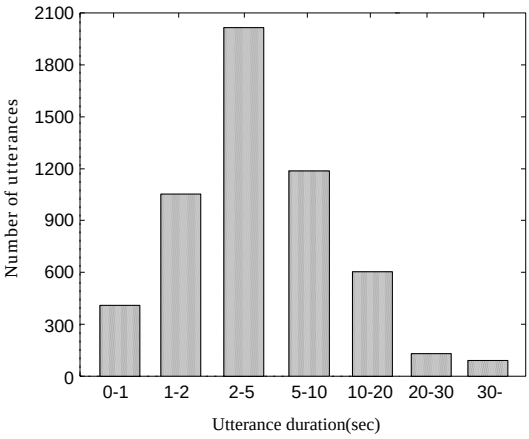

Fig. 1. Distribution of utterance lengths

3. SPEAKER INDEXING USING VARIANCE-BIC

BIC is a likelihood criterion penalized by the model complexity or the number of parameters in the model. Specifically let $X = \{x_j \in \Re^d : j = 1, \cdots, N\}$ be the data set, and $\lambda = \{\lambda_i : i = 1, \cdots, K\}$ be the candidates of parametric models and β_i is the number of parameters in the model λ_i. The BIC is defined as:

$$BIC_i = \log P(X|\lambda_i) - \alpha\frac{1}{2}\beta_i \log N \tag{1}$$

where α is a penalty weight.

The conventional method of speaker indexing based on Variance-BIC [4] also consists of speaker segmentation and clustering processes. The decision of a speaker turn, or to decide if two consecutive segments are uttered by different speakers, is based on the BIC for variances of two clusters and formulated with the following function,

$$\Delta BIC^i{}_{variance} = -\frac{n_1+n_2}{2}\log|\Sigma_0| + \frac{n_1}{2}\log|\Sigma_1| + \frac{n_2}{2}\log|\Sigma_2| + \alpha\frac{1}{2}(d+\frac{1}{2}d(d+1))\log(n_1+n_2) \tag{2}$$

where Σ_0 is a covariance of the merged segment, Σ_1 is a covariance of the first segment, and Σ_2 is a covariance of the second segment. And n_i represents the data size (number of frames) of respective segments. If ΔBIC is positive, the two clusters are merged.

In discussion data, the variance of utterance duration is large, hence, reliable estimation and fair comparison of variances is difficult especially for very short speech segments. The method has another problem that the penalty weight is task-dependent and has to be tuned for every new task [7].

4. SPEAKER INDEXING USING SPEAKER MODEL SELECTION

4.1. Speaker Model Selection

We explore a novel approach that directly generates speaker models depending on the data size. GMM is an appropriate statistical model, but needs a lot of training data for reliable parameter estimation. When little data is available on the contrary, a simple VQ-based method, which uses the VQ distortion as a distance measure, performs better than GMM [6]. In conventional speaker recognition tasks, the amount of training data is almost same for each speaker, and the speaker model is specified manually according to the nature of the task or available size of the training data. In this paper, 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.

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. CVGMM becomes the VQ model by replacing the covariance matrix with the identity matrix.

We first estimate the mixtures of GMM for each speech segment. Then, we replace the covariance to generate CVGMM and compute the BIC value for GMM and CVGMM. Specifically, the BIC for GMM of a cluster s is given by

$$BIC^{(s)}_{GMM} = \log P(X|\lambda^{(s)}_{GMM}) - \alpha M d \log N \tag{3}$$

where $\log P(X|\lambda^{(s)}_{GMM})$ is a log likelihood of the training data X by the GMM, M is the number of mixtures, N is the number of frames of the training data, and the penalty weight α is set to 1. The BIC for CVGMM is given by below.

$$BIC^{(s)}_{CVGMM} = \log P(X|\lambda^{(s)}_{CVGMM}) - \alpha\frac{1}{2}d(M+1)\log N \tag{4}$$

Here, the mixture weights of CVGMM are given by $\bar{w}^{(s)} = \frac{1}{M}$ uniformly. Estimation of the covariance of CVGMM is very difficult for a cluster having a small amount of training data. So, we replace it with the average of the covariances of GMMs trained for all clusters by Eq. (5).

$$\bar{\Sigma}^{(s)}_{CVGMM} = \frac{1}{M \cdot S}\sum_{i=1}^{S}\sum_{j=1}^{M}\Sigma^{(i)}_{GMM_j} \tag{5}$$

Here, S is the number of clusters.

If the training data is sparse, 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 and the discriminant measure according to the data size. Thus, it can perform speaker indexing for any lengths of utterances.

4.2. Speaker Indexing

Speaker indexing is performed by selecting GMM or CVGMM based on the BIC. We call the method "SMS (Speaker Model Selection)".

The procedure is described as follows:

1. Training: For each cluster, GMM and CVGMM are trained. In the initial training, each utterance makes one cluster.
2. Model selection: An optimal model is selected for each cluster between GMM and CVGMM based on the BIC.
3. Distance computation: The distance between clusters is computed based on the Cross Likelihood Ratio [8]. The Cross Likelihood Ratio d_{ij} is given by

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

   where X_i is the utterances and λ_i is the selected model (GMM or CVGMM) in the cluster i.
4. 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 the same, they are merged.

   Step 1, 2, 3 and 4 are repeated until no more clusters can be merged.
5. 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 1, 2, 3 and 5 are repeated until distances for all cluster pairs are larger than the threshold θ.

The first merging procedure (step 4) is introduced for initial clusters of short segments for which stable likelihood is not obtained. Then speaker clustering based on the likelihood is performed (step 5).

5. EXPERIMENTAL EVALUATION

5.1. Experimental Condition and Evaluation Measure

All ten discussion data described in Section 2 are used in the experiments. We compared our method with the conventional method based on the Variance-BIC method and the GMM-based method. GMM is same as the proposed method, but we assume it is selected for all clusters.

The speech data is sampled at 16 kHz and the acoustic features consist of 26 components of 12MFCCs, energy and their deltas. The threshold θ of the proposed method is optimized in a preliminary experiment. The penalty weight α in the Variance-BIC is set to 3.0 using the discussion data other than the test set.

In this study, we use the BBN metric to evaluate the indexing performance. The BBN metric [9] is given by

$$I_{BBN} = \sum_{i=1}^{N_c} n_i p_i - Q N_c, \tag{7}$$

where n_i is the number of utterances in a candidate cluster i, N_c is the number of candidate clusters, and p_i is purity of a cluster i. Purity is defined as $p_i = \sum_{j=1}^{N_s} (\frac{n_{ij}}{n_i})^2$, where N_s is the number of speakers and n_{ij} is the number of utterances of speaker j in cluster i. A variable Q is a system design parameter that controls the degree to which fewer and larger clusters are favored at the expense of decreased purity. We set the parameter $Q = 0.5$.

Table 2. Speaker indexing result

	Index	Spk num		
		RE	PR	F-value
Variance-BIC	0.81	1.00	0.83	0.91
GMM				
(4 mix)	0.86	0.95	0.66	0.78
(8 mix)	0.94	1.00	0.75	0.86
(16 mix)	0.93	0.98	0.91	0.94
(32 mix)	0.90	0.98	0.89	0.93
SMS				
(4 mix)	0.86	0.95	0.66	0.78
(8 mix)	0.93	1.00	0.74	0.85
(16 mix)	0.95	0.98	0.88	0.93
(32 mix)	0.97	1.00	0.86	0.92

We perform evaluation using the ratio of the BBN metric by the automatic indexing methods and that by the correct indexing as well as the accuracy of the number of speakers. The accuracy of the number of speakers is measured by the recall rate (RE) and the precision rate (PR) and F-value. These are defined as follows, respectively:

$$RE = \frac{Number\ of\ correctly\ indexed\ speakers}{Actual\ number\ of\ speakers} \tag{8}$$

$$PR = \frac{Number\ of\ correctly\ indexed\ speakers}{Number\ of\ indexed\ speakers} \tag{9}$$

$$F - value = \frac{2 \cdot RE \cdot PR}{RE + PR} \tag{10}$$

5.2. Experimental Results

The average indexing performance obtained by the methods is shown in Table 2. In Table 2, "Index" denotes the ratio of the BBN metric and "Spk num" denotes the accuracy of the number of speakers.

The proposed SMS method achieved accuracy of 97% in indexing and 92% (F-value) in estimation of the number of speakers when the number of mixtures was 32. It outperforms the method based on the Variance-BIC and the conventional GMM.

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 and the accuracy of the number of speakers is lower.

In the SMS method, it is possible to train the models with 16 or larger mixtures because of the introduction of Common Variance GMM as an extension of VQ. The method realizes flexible modeling of data and accurately performs the cluster merging process with cross verification. Actually, CVGMM is more likely to be selected as the number of mixtures gets larger.

Performance of GMM and SMS is comparable when the number of mixtures is 4 and 8. GMM can be well trained and gets a

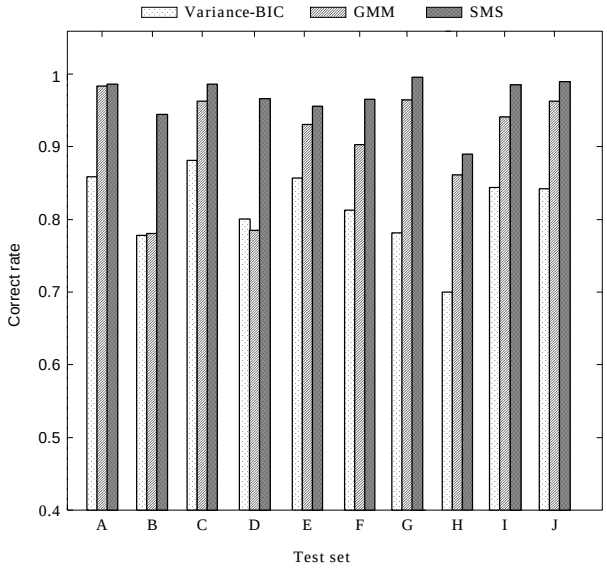

Fig. 2. Index performance (Ratio of BBN metric) for each discussion

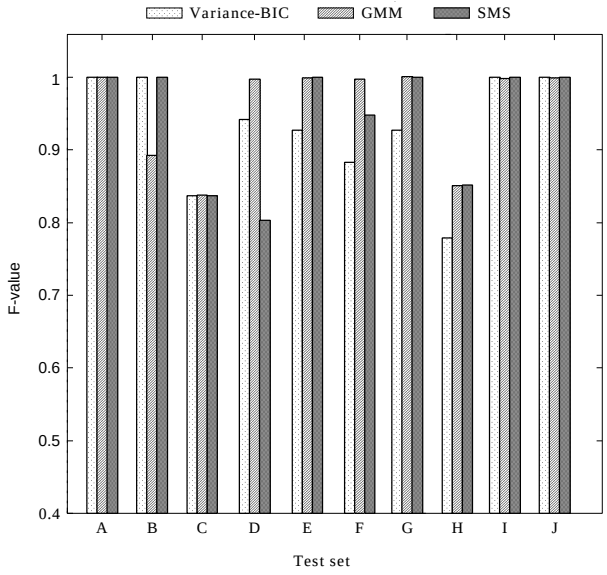

Fig. 3. Estimation result of number of speakers for each discussion

better BIC value than CVGMM in most cases when the number of mixtures is small. Although the Variance-BIC method realizes high accuracy in estimating the number of speakers, the indexing performance is low. Most of very short utterances are incorrectly clustered because the fixed penalty weight α is used in spite of a large variation in the duration of utterances.

The indexing performance for each discussion is shown in Fig. 2 and the accuracy of the number of speakers is shown in Fig. 3. In Figs. 2 and 3, "GMM" and "SMS" denote the result when the number of mixtures is 32. The SMS method achieves the best performance over most of the data.

6. CONCLUSION

We presented a method of unsupervised speaker indexing for discussions, in which the speaker changes frequently, thus the duration of utterances is short and the variation in the duration is large. The proposed method selects an optimal speaker model among VQ and GMM based on the BIC according to the duration of utterances. It was demonstrated that the method realizes high indexing performance. The method works without specifying the number of speakers and the models of each speaker in advance.

As a future work, we will perform automatic speech recognition based on unsupervised speaker adaptation using the speaker indexing results.

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] D.A. Reynolds, "Comparison of Background Normalization Methods for Text-Independent Speaker Verification Systems," Proc. EUROSPEECH, pp. 963-966, 1997.

[4] 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.

[5] D.A. Reynolds and R.C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE Trans.SAP, vol. 3, no. 1, pp. 72-83, 1995.

[6] 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.

[7] A. Tritschler and R. Gopinath, "Improved Speaker Segmentation and Segments Clustering Using the Bayesian Information Criterion," Proc. EUROSPEECH, vol. 2, pp. 679-682, 1999.

[8] 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.

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