Speaker Model Selection Based on the Bayesian Information Criterion Applied to Unsupervised Speaker Indexing

Masafumi Nishida and Tatsuya Kawahara, *Member, IEEE*

***Abstract*—In conventional speaker recognition tasks, the amount of training data is almost the same for each speaker, and the speaker model structure is uniform and specified manually according to the nature of the task and the available size of the training data. In real-world speech data such as telephone conversations and meetings, however, serious problems arise in applying a uniform model because variations in the utterance durations of speakers are large, with numerous short utterances. We therefore propose a flexible framework in which an optimal speaker model (GMM or VQ) is automatically selected based on the Bayesian Information Criterion (BIC) according to the amount of training data available. 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 amount of data is obtained. The proposed framework was implemented in unsupervised speaker indexing of a discussion audio. For a real discussion archive with a total duration of 10 hours, we demonstrate that the proposed method has higher indexing performance than that of conventional methods. The speaker index is also used to adapt a speaker-independent acoustic model to each participant for automatic transcription of the discussion. We demonstrate that speaker indexing with our method is sufficiently accurate for adaptation of the acoustic model.**

***Index Terms*—Automatic speech recognition, Bayesian information criterion, discussions, speaker adaptation, speaker model selection, speaker recognition, unsupervised speaker indexing.**

I. Introduction

TEXT-INDEPENDENT speaker recognition has been the subject of intensive research for the past several decades. Vector quantization (VQ) [1]–[4] and Gaussian mixture model (GMM) [5]–[11] have been used as the speaker model. When the speaker recognition tasks assume read-speech corpora, the amount of training data is almost the same for each speaker. Thus, the structure of the speaker model is uniform and is specified manually according to the nature of the task and available amount of the training data. When a sufficient amount of training data is available, GMM usually achieves high recognition accuracy. However, it is not effective if the training data size is insufficient to enable the covariance matrices of mixture densities to be estimated reliably. It has been reported that recognition with the simple VQ-based method is even higher than that with GMM when limited training data is available [12]. Recently, the main target of speaker recognition research has shifted to spontaneous speech such as telephone conversations [13], [14] and meetings [15]. In these real-world data, 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.

In this paper, we propose a novel approach in which an optimal speaker model (GMM or VQ) is selected based on the Bayesian information criterion (BIC) [16], which reflects the amount of speech data. The framework makes it possible to use a discrete model when the training data is sparse, and to seamlessly switch to a continuous model after sufficient data is obtained. Thus, the method we propose enables the model structure to be changed dynamically according to data size. We call this method statistical speaker model selection (SSMS).

The proposed framework is applied to speaker indexing of discussion audio archives. Speaker indexing is essential in retrieving the utterances of a specific speaker and also in improving automatic speech recognition based on speaker adaptation of the acoustic model.

Speaker indexing generally consists of a sequence of speaker identification processes. Studies have been reported on methods under both supervised and unsupervised training conditions. Under the supervised training conditions, it is assumed that training data is available for the target speakers, and GMMs are trained to represent the speech of the target speakers [17]–[19]. Furthermore, methods using Kullback information based on a codebook [20] and an ergodic HMM [21] have been proposed by assuming that only the number of speakers is known beforehand. Speaker indexing is rather easy if we can train speaker models in advance and test data consists of only trained speakers and is assumed to fit one of the trained models. In the audio archives of discussions that we deal with here, the speakers are not always the same, and it is not practical to assume that speech samples of individual speakers will be available beforehand for various tasks including discussions. We therefore focus on unsupervised speaker indexing without prior speaker models. Moreover, we do not assume that the number of speakers is given.

Recently, unsupervised speaker indexing is vigorously studied using the database of not only broadcast news but also switchboard conversations and meetings which are dealt

Manuscript received August 28, 2003; revised May 6, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ramesh A. Gopinath.

M. Nishida is with Graduate School of Science and Technology, Chiba University, Chiba 263-8522, Japan (e-mail: nishida@faculty.chiba-u.jp).

T. Kawahara is with the Academic Center for Computing and Media Studies, Kyoto University, Kyoto 606-8501, Japan (e-mail: kawahara@i.kyoto-u.ac.jp).

Digital Object Identifier 10.1109/TSA.2005.848890

1063-6676/$20.00 © 2005 IEEE

with at the NIST evaluation tasks. Approaches to unsupervised speaker indexing are mainly classified into metric-based and model-based methods. The metric-based methods try to detect speaker changes and to perform speaker clustering by differences between segments based on distance measures such as the Generalized Likelihood Ratio (GLR) [22], [23], Kullback-Leibler distance (KL) [24] and BIC [25], [26]. The BIC based method assumes a single Gaussian distribution for each segment and determines speaker changes based on variances between segments. It is effective in broadcast news, where speech segments are long and changes in speaker are not so frequent. We call this method "Variance-BIC" in this paper because likelihood is represented by a variance. In conversations and meetings, however, speakers change frequently, and there are many short utterances and large variations in speech length. Thus, the comparison based on such fluctuating data might not work. Moreover, speaker information may not be fully represented with a single Gaussian distribution for each segment.

Model-based approaches use an ergodic HMM network consisting of nodes representing speakers [27], or train a GMM for each segment [28], [29]. Recently, one of the most widely-used methods is to set up speaker models by adapting the universal background model (UBM) [30] based on maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP). In this kind of methods, speakers are identified based on the likelihood ratio between the adapted and background models [31], [32]. Several studies introduce hierarchical clustering as a post-processing based on the cross likelihood ratio (CLR) [33], [34]. Another method adopts HMM, as in LIA system, where each state represents a speaker and the transitions correspond to speaker changes. It is trained by adapting with data segmented with Viterbi decoding [35]. This kind of adaptation approach may work well for relatively longer utterances. However, it encounters difficulty in applying to conversations and meetings where there are numerous turns and short utterances. In the discussion data we deal with here, utterances have a duration of 6 seconds on average and the ratio of the utterances that are less than 10 seconds is about 87%. It is not feasible to reliably perform adaptation for these short utterances.

To cope with real-world audio archives, where the duration of utterances ranges from very long to very short, we introduce a statistical speaker model selection (SSMS) scheme through which optimal models of suitable complexity are selected depending on utterance length. The method we propose is evaluated at speaker indexing of a real discussion archive with a total duration of 10 h. The results obtained from speaker indexing are then used to adapt the acoustic model used to automatically transcribe the archive. The indexing accuracy is also evaluated in terms of speech recognition.

The paper is organized as follows. Section II introduces the proposed SSMS scheme which selects an optimal speaker model (GMM or VQ) based on the BIC. Section III describes the speaker indexing based on the proposed SSMS. The specification of the discussion data and the experimental results of speaker indexing are presented in Section IV. Section V describes automatic speech recognition based on the speaker indexing, and Section VI concludes the paper.

II. Statistical Speaker Model Selection

A. Bayesian Information Criterion

The problem of selecting an appropriate model is formulated as choosing from a set of candidates that describe a given data set. Generally, the likelihood of training data is improved by increasing the number of parameters in the model. However, when there are too many parameters, the model encounters problems of overfitting and lacks robustness.

The BIC (Bayesian information criterion) has been introduced to model selection based on the Bayesian estimation of model parameters. Specifically, let $X = \{x_j \in \Re^d : j = 1, \ldots, N\}$ be the data set, and $\lambda = \{\lambda_i : i = 1, \ldots, K\}$ be the candidates for parametric models and β_i be the number of parameters in model λ_i. The BIC is then defined as

$$\mathrm{BIC}_i = \log P(X \mid \lambda_i) - \alpha \frac{1}{2} \beta_i \log N \tag{1}$$

where $\log P(X \mid \lambda_i)$ is the log likelihood of training data X for model λ_i and α is the weight of the second term.

If the likelihood given by two models is comparable, the simpler model would give a better BIC value, and be selected. Thus, the BIC considers the balance between likelihood and model complexity and avoids overfitting.

B. Speaker Model Selection Based on the BIC

We explore a novel approach that selects speaker models depending on the data size. One way of coping with sparse training data is to adopt a discrete model. When little data is available, a simple VQ-based method, which uses VQ distortion as a distance measure, performs better than GMM [12]. GMM is an appropriate statistical model, but needs a lot of training data for reliable parameter estimation. If we can assume that the amount of training data is almost the same for each speaker, the speaker model is uniform and specified manually according to the nature of the task or the available size of the training data. As the assumption does not hold for real-world data such as discussion archives, we propose a flexible framework in which an optimal speaker model (GMM or VQ) is automatically selected based on the BIC according to the amount of training data. As previously stated, we call this framework "SSMS (statistical speaker model selection)".

One problem in implementing this framework for 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)", which is an extension of the VQ-based model [36]. EVQ is modeled by assigning the same mixture weight and covariance to all Gaussian mixture components. It normalizes the distance measure of VQ, so that it can be compared to the likelihood of GMM. EVQ becomes a VQ model by replacing the covariance matrix with an identity matrix.

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

$$\mathrm{BIC}^{(s)}_{\mathrm{GMM}} = \log P\left(X \,\middle|\, \lambda^{(s)}_{\mathrm{GMM}}\right) - \frac{1}{2} M(2d+1) \log N \tag{2}$$

where $\log P(X \mid \lambda^{(s)}_{\mathrm{GMM}})$ is the log likelihood of training data X 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. Penalty weight α is set to 1 in this case. Here, d-dimensional means and variances plus mixture weights are counted. [1]

We then generate the EVQ. The mixture weights of EVQ are uniformly assigned as $w_{\mathrm{EVQ}} = 1/M$. We also replace its covariance with the average covariances of GMMs trained for all speakers as follows:

$$\Sigma_{\mathrm{EVQ}} = \frac{1}{M \cdot S} \sum_{i=1}^{S} \sum_{j=1}^{M} \Sigma^{(i)}_{\mathrm{GMM}_j} \tag{3}$$

where S is the number of speakers, and $\Sigma^{(i)}_{\mathrm{GMM}_j}$ is the covariance of jth mixture component of speaker i. We can use a mean vector of GMM obtained through an EM algorithm using a fixed variance for a mean vector of EVQ. However, we generate the EVQ-based model by using a mean vector obtained through an LBG (Linde, Buzo, and Gray) algorithm [37] and estimated variance with (3) because the proposed framework is to choose a discrete model (VQ) or a continuous model (GMM) according to the amount of training data available.

The BIC for the EVQ is given by

$$\mathrm{BIC}^{(s)}_{\mathrm{EVQ}} = \log P\left(X \;\middle|\; \lambda^{(s)}_{\mathrm{EVQ}}\right) - \frac{1}{2}(M+1)d \log N \tag{4}$$

where d-dimensional means for M Gaussians are counted as well as one common variance.

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 larger. Thus, an appropriate speaker model can be constructed for any utterance length. Moreover, the framework makes it possible to use the discrete model (VQ) when the training data is sparse, and to seamlessly switch to a continuous model (GMM) after the training data is fully obtained.

C. Gaussian Mixture Size Selection

Another way to control the complexity of the model is to change the number of Gaussian distributions based on the BIC according to the amount of training data. This has been used to control the complexity of the acoustic model (HMM) in automatic speech recognition [38], [39]. We call this method "GMSS (Gaussian mixture size selection)." The BIC of the GMM for speaker s is given by the same (2), and we use script M which stands for the number of components of GMM.[2]

$$\mathrm{BIC}^{(s)}_{M} = \log P\left(X \;\middle|\; \lambda^{(s)}_{M}\right) - \frac{1}{2}M(2d+1)\log N. \tag{5}$$

[1]The strictly accurate definition of the BIC of the GMM is $\mathrm{BIC}^{(s)}_{\mathrm{GMM}} = \log P(X \mid \lambda^{(s)}_{\mathrm{GMM}}) - (1/2)(2dM + M - 1)\log N$ because the number of free parameters of mixture weights is smaller by one. However, we used (2) in this paper because the difference is negligible.

[2]The value of parameter α in this BIC was set to 1 after preliminary experiments.

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

$$\begin{aligned}\Delta\mathrm{BIC}^{(s)} &= \mathrm{BIC}^{(s)}_{M} - \mathrm{BIC}^{(s)}_{2M} \\ &= \log P\left(X \;\middle|\; \lambda^{(s)}_{M}\right) - \log P\left(X \;\middle|\; \lambda^{(s)}_{2M}\right) \\ &\quad + \frac{1}{2}M(2d+1)\log N. \end{aligned} \tag{6}$$

The size of the mixture is doubled if $\Delta\mathrm{BIC}^{(s)}$ is negative. Otherwise, it is determined as M.

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

III. Speaker Indexing Algorithm

A. Speaker Indexing Based on the Variance-BIC

The conventional method of speaker indexing based on the Variance-BIC is formulated as follows [25]. It assumes a single Gaussian distribution for each segment and performs speaker clustering based on the variance ratio. In this paper, one or more utterance units are called a segment. Initially, each utterance makes a segment.

To decide if any pair of segments is uttered by the same speaker, the method computes the difference in the BIC value $\mathrm{BIC}^{0}_{\mathrm{var}}$ for a hypothesis that the segments are uttered by the same speaker, and the value $\mathrm{BIC}^{12}_{\mathrm{var}}$ for a hypothesis that the segments are uttered by different speakers, which are given by (7) and (8), respectively.[3]

$$\begin{aligned}\mathrm{BIC}^{0}_{\mathrm{var}} &= -\frac{N_1+N_2}{2}\log|\Sigma_0| \\ &\quad - \frac{\alpha}{2}\left(d + \frac{d(d+1)}{2}\right)\log(N_1+N_2) \end{aligned} \tag{7}$$

$$\begin{aligned}\mathrm{BIC}^{12}_{\mathrm{var}} &= -\frac{N_1}{2}\log|\Sigma_1| - \frac{N_2}{2}\log|\Sigma_2| \\ &\quad - \alpha\left(d + \frac{d(d+1)}{2}\right)\log(N_1+N_2) \end{aligned} \tag{8}$$

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

The difference in the BIC values is rewritten as

$$\begin{aligned}&\Delta\mathrm{BIC}_{\mathrm{var}} \\ &\quad = \mathrm{BIC}^{0}_{\mathrm{var}} - \mathrm{BIC}^{12}_{\mathrm{var}} \\ &\quad = -\frac{N_1+N_2}{2}\log|\Sigma_0| + \frac{N_1}{2}\log|\Sigma_1| + \frac{N_2}{2}\log|\Sigma_2| \\ &\qquad + \alpha\frac{1}{2}\left(d + \frac{d(d+1)}{2}\right)\log(N_1+N_2). \end{aligned} \tag{9}$$

If $\Delta\mathrm{BIC}_{\mathrm{var}}$ is positive, the two segments are merged. Speaker clustering is performed by repeating the process. When the $\Delta\mathrm{BIC}_{\mathrm{var}}$ values between all segment pairs become negative,

[3]The original equation of the $\mathrm{BIC}^{0}_{\mathrm{var}}$ has a term of $(-(N_1+N_2)/(2) \times (d/2) \times \log(2\pi))$. We omitted the term because it is common (constant) for all models. That also holds true for (8).

the clustering process is finished and the set of speakers is defined together with their speech segments. Thus, the BIC is used as a termination criterion in a hierarchical speaker clustering procedure. As previously stated, we call this method "Variance-BIC" because the likelihood is represented by a variance.

As variations in the duration of utterances in the discussion data are large, reliable estimation and fair comparison of variances are difficult especially for very short speech segments. Another problem is that penalty weight α is task-dependent and must be tuned for every new task [40]. Moreover, the method assumes a single Gaussian distribution for each segment, and speaker information may not be fully represented.

B. Speaker Indexing Based on Speaker Model Selection

Next, a speaker indexing procedure is presented based on the SSMS scheme we propose. Speaker indexing is performed through an iterative process of training speaker models for segments and then merging them. In this step, optimal speaker model is chosen between GMM and EVQ based on the BIC. As the distance measure between models, we adopt the cross likelihood ratio (CLR) which is based on likelihoods for corresponding utterances, thus reliably applicable to both GMM and EVQ. At the early stage of clustering, there are so many clusters that are different in sizes and selected model types. So, we perform identification scheme which merges clusters assigned with a same label. At the later stage, we expect that most of the clusters are made of sufficient data size and modeled with GMM. So, we conduct verification scheme which merges clusters within a certain distance of the CLR Fig. 1 shows a flow diagram of this procedure.

The detailed procedure is described as follows.

1) *Training:* For each cluster, GMM and EVQ are trained using all utterances of the cluster. In the initial step, each utterance forms one cluster.
2) *Model selection:* An optimal model is selected for each cluster between GMM and EVQ based on the BIC.
3) *Distance calculation:* The distance between clusters is computed based on the cross likelihood ratio (CLR) [41]. The CLR d_{ij} for clusters i and j is given by

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

$$\log P(X_i \mid \lambda_j) = \sum_{k=1}^{n_i} \log P(x_{ik} \mid \lambda_j) \tag{10}$$

where X_i is all utterances of cluster i, x_{ik} is its kth utterance, n_i is the number of utterances, λ_i is the selected model (GMM or EVQ) for cluster i, and $\log P(X_i \mid \lambda_j)$ is the average log likelihood of utterances of cluster i given by model λ_j.

4) Merging clusters with identification: For each cluster, the closest cluster with the minimum distance is identified. Then, for a pair of clusters, if they share the same closest cluster (which is different from themselves), they are merged. Namely, merge clusters i and j if $\arg\min_k d_{ik} = \arg\min_k d_{jk} (i \neq j \neq k)$.

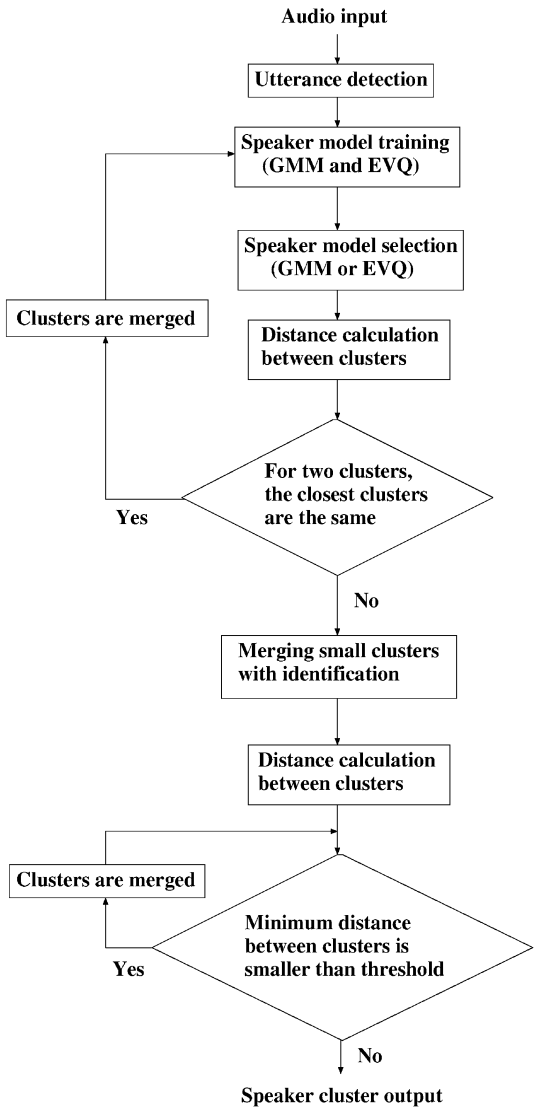

Fig. 1. Flow diagram of speaker indexing based on statistical speaker model selection.

Steps 1, 2, 3, and 4 are repeated until no more clusters can be merged.

5) Merging small clusters with identification: If there are small clusters remaining whose selected model is EVQ even after step 4, they are merged with large clusters whose speaker model is GMM and gives maximum likelihood to them. (The merged cluster keeps original component speaker models for distance calculation of the next step.)
6) Merging clusters with verification: The minimum distance between clusters is computed and if it is smaller than threshold θ, these clusters are merged. Namely, merge clusters i and j if $d_{ij} < \theta$. This process is done once for all pairs of clusters after step 5 and not repeated.

After sufficient training data is obtained for each cluster through the first merging procedure (step 4), the training procedure (step 1) is not performed for efficiency. At this phase, we assume that all clusters can be modeled with GMM, and small clusters that consist of short utterances are false alarms or irregular samples. They are merged with the other identification

TABLE I
TEST SET OF DISCUSSION AUDIO

	A	B	C	D	E
#Speaker	5	5	5	8	6
#Utterance	534	665	609	541	612
	F	G	H	I	J
#Speaker	8	5	5	5	5
#Utterance	474	371	613	559	524

procedure (step 5).[4] Speaker clustering is then performed based on the matching likelihood (step 6). For accurate clustering in this process, we keep all original model parameters merged in step 5. And in step 6, the distance between clusters is defined by measuring distances with all models kept in the cluster and taking the minimum.

IV. SPEAKER INDEXING EXPERIMENTS

A. Database and Task

As the material for speaker indexing experiments, we used a one-hour forum-type TV program that is broadcast on Sundays. 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 were divided into utterance units based on energy and zero-crossing parameters. The error rate of the utterance segmentation or the ratio of cases where two speakers are contained in a segmented utterance was 2.9%. The correct (reference) label for speaker indexing in this case was defined as a speaker with the longest utterance duration in the segment. Table I shows the numbers of speakers and utterances in the discussions. Each discussion comprised five to eight speakers with an average of 550 utterances. Fig. 2 shows the distribution of the duration of utterances. Here, "5–10" means the number of utterances with a duration from 5 to 10 s.

The average duration was 6 s, the minimum was 1 s, and the maximum was 71 s. Utterances with durations of less than 10 seconds represented about 87% of the data. There were numerous short utterances and also large variations in duration. This suggests a serious problem in applying a uniform model and the necessity for a framework by which an optimal model of suitable complexity can be selected depending on data size.

B. Experimental Conditions and Evaluation Measures

Audio material of ten discussions previously described was used in the unsupervised speaker indexing experiments. The speech data were sampled at 16 kHz and the acoustic features consist of 26 components of 12 MFCCs, energy and their deltas.

We compared our method (SSMS) with conventional methods, i.e., the Variance-BIC, the VQ-based and the GMM-based methods including GMSS. 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. In the VQ-based method, the log likelihoods in the cross likelihood ratio of (10) are replaced with the Euclidean distance between centroids, which assumes identity covariances. In the GMSS method, the GMM is trained for each cluster and its mixture size is determined by the BIC value of (6). Clustering is performed based on the cross likelihood ratio between the obtained GMMs.

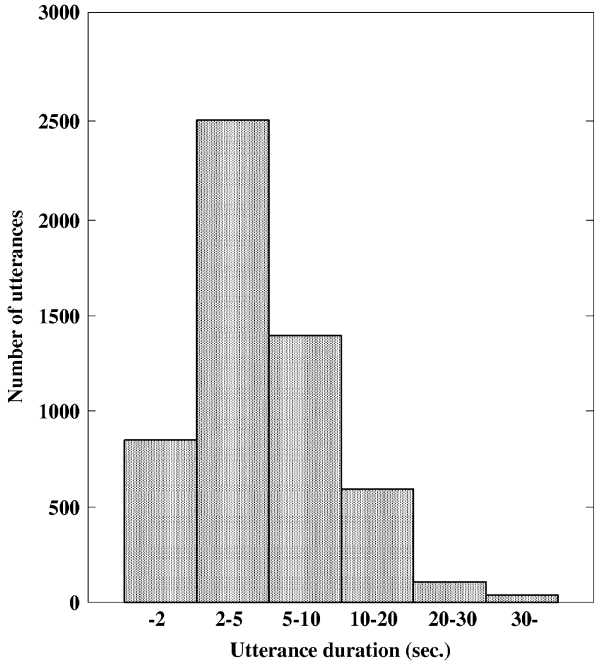

Fig. 2. Distribution of utterance lengths.

We carried out the speaker indexing experiments for two cases, where the number of speakers was both unknown and known in advance because it was easy to obtain the number beforehand in discussions and meetings. Where the number of speakers was given beforehand, cluster merging (step 6) continued until the number of obtained clusters reached the specified number by disregarding threshold θ.

To evaluate indexing performance, we use the BBN metric [29] which is given by

$$I_{\mathrm{BBN}} = \sum_{i=1}^{C} n_i p_i - QC$$
$$p_i = \sum_{j=1}^{S} \left(\frac{n_{ij}}{n_i} \right)^2 \tag{11}$$

where n_i is the number of utterances in candidate cluster i, and C is the number of candidate clusters. p_i is the purity of cluster i, S is the number of actual speakers and n_{ij} is the number of utterances by speaker j in cluster i. Indexing performance increases with a larger value for the BBN metric. 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 this parameter to $Q = 0.5$ based on the balance with the number of speakers and utterances.

We did evaluations using speaker indexing accuracy and accuracy on the number of speakers. Speaker indexing accuracy

[4]This assumption makes it impossible to cope with data where some speakers articulate one or two really short utterances. In an earlier implementation, we did not include step 5. We observed that introducing this step had little effect on the accuracy of speaker indexing, but improved the estimation of the number of speakers by eliminating false irregular clusters.

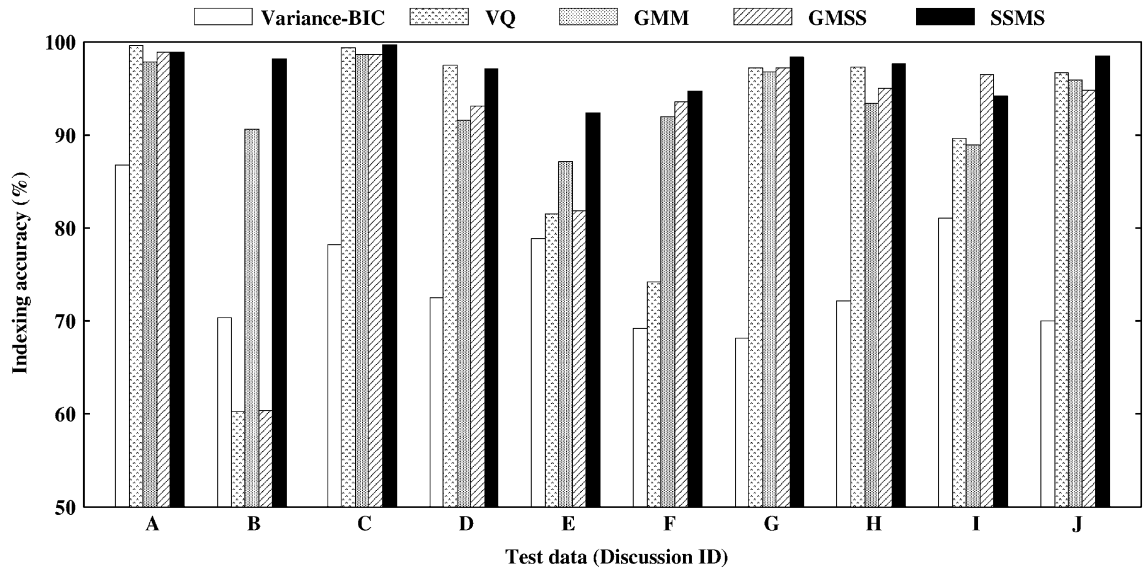

Fig. 3. Speaker indexing accuracy (SIA) for each discussion when the number of speakers is unknown.

is defined as the ratio of the BBN metric obtained by automatic indexing and that by correct indexing. Thus, this is given by

$$\mathrm{SIA} = \frac{I_{\mathrm{BBN}}}{n - QS} \times 100 = \frac{\sum_{i=1}^{C} n_i p_i - QC}{n - QS} \times 100 \tag{12}$$

where n is the total number of utterances and S is the actual number of speakers. It became 0 at worst and 100 at best. Accuracy on the number of speakers is defined as

$$\mathrm{SNA} = \left\{ 1 - \frac{\sum_{k=1}^{D} |S_k - C_k|}{\sum_{k=1}^{D} S_k} \right\} \times 100 \tag{13}$$

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

TABLE II
SPEAKER INDEXING ACCURACY WHEN THE NUMBER OF SPEAKERS IS UNKNOWN

	Speaker Indexing Accuracy (%)	Speaker Number Accuracy (%)
Variance-BIC	74.7	100.0
VQ		
(4 cb)	57.8	80.7
(8 cb)	80.7	86.0
(16 cb)	92.4	89.5
(32 cb)	89.3	84.2
GMM		
(4 mix)	72.3	75.4
(8 mix)	93.8	86.0
(16 mix)	95.8	96.5
(32 mix)	93.3	100.0
GMSS	91.0	89.5
SSMS		
(4 mix)	72.3	73.7
(8 mix)	93.5	84.2
(16 mix)	96.1	93.0
(32 mix)	97.0	100.0

cb: codebook size, mix: mixture size

C. Experimental Results

The average indexing performance when the number of speakers is unknown is shown in Table II. Penalty weight α in the Variance-BIC was set to 5.0 after preliminary experiments. The threshold θ for the speaker clustering procedure (step 6) was determined so that the accuracy on the number of speakers (SNA) was maximum for each method. The same threshold value was used for all ten discussions, which means speaker number accuracy (SNA) on average may not be 100% even with the optimal case. The indexing performance for individual discussions is in Fig. 3. Here, "VQ," "GMM," and "SSMS" denote the result when the size of the mixtures or codebooks is 32.

The SSMS method we propose achieved a speaker indexing accuracy (SIA) of 97.0% when there were 32 mixture components. It outperformed the Variance-BIC, the VQ-based and the GMM-based methods including GMSS. It achieved the best performance over almost all the discussions. It is also verified that discrete VQ was chosen when utterances are short, and stochastic GMM was chosen for large clusters. We thus demonstrated the effectiveness of the proposed framework that selects an optimal speaker model (GMM or VQ) based on the BIC according to the training data. For reference, we also evaluated speaker indexing accuracy with a measure that counts the duration of utterances. It is defined by replacing n_i (number of utterances in the cluster) of (11) with l_i (number of frames belonging to the cluster). As a result, much the same tendency is observed: 86.7% for Variance-BIC $(\alpha = 5.0)$, 89.5% for VQ (32 cb.), 93.8% for GMM (32 mix.), 91.3% for GMSS, and 95.7% for SSMS (32 mix.). We confirmed that almost all of long utterances with a large number of frames are correctly labeled, and short utterances have dominant effect in the speaker indexing performance.

Next, the average indexing performance when the number of speakers is given in advance is shown in Table III. SSMS achieved a speaker indexing accuracy (SIA) of 97.0% with

TABLE III
SPEAKER INDEXING ACCURACY WHEN THE 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
GMSS	84.4
SSMS	
(4 mix)	66.8
(8 mix)	89.4
(16 mix)	91.6
(32 mix)	97.0

cb: codebook size, mix: mixture size

a 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 specifying the number of speakers has the same effect as using optimal threshold θ. This does not necessarily hold true for other methods, however, e.g., the VQ-based method with a codebook size of 32, because the best SIA was obtained when more than the actual numbers of speaker clusters were used.

The GMM-based method has more difficulty in estimating large mixtures with the data because there are so many short utterances for which the variances of some mixture components become too small, and this causes false matching. Therefore, clusters of the same speakers are not correctly merged. It is possible in the SSMS method to train models with 16 mixtures or larger because of the introduction of EVQ as an extension of the VQ-based model. SSMS realizes flexible data modeling and accurately merges clusters with identification based on the CLR. Actually, EVQ is more likely to be selected as the number of mixtures increases. GMM and SSMS are almost comparable in performance when the mixture size is small. GMM can be well trained and yields a better BIC value than EVQ in most cases with a limited number of mixture components.

GMSS, which adaptively controls the mixture size, did not perform better. The method often selected a single Gaussian distribution—especially for short utterances—which is a poor representation compared with VQ, and most of very short utterances were incorrectly clustered.

It is possible with the VQ-based method to train a stable model even with limited training data. However, it does not represent speaker information after a sufficient cluster size is obtained. Actually, the GMM-based method obtained better performance when the size of the mixtures and codebooks was the same.

Speaker indexing accuracy by the Variance-BIC is lower compared with other methods. Most of the short utterances were incorrectly clustered because a fixed penalty weight α was used despite large variations in the duration of utterances.

We then investigated the sensitivity of the speaker indexing accuracy to the threshold θ of the speaker clustering procedure (step 6). Fig. 4 shows the SIA plotted by changing the threshold θ. Also, Fig. 5 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. For the 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 that the accuracy is sensitive to this value, and that the peaks of SIA and SNA are obtained at totally different values of α. The SIA in the GMSS method is less sensitive to variations in threshold θ. However, the SNA changes with slight variations in threshold θ compared with the other methods. SSMS 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.

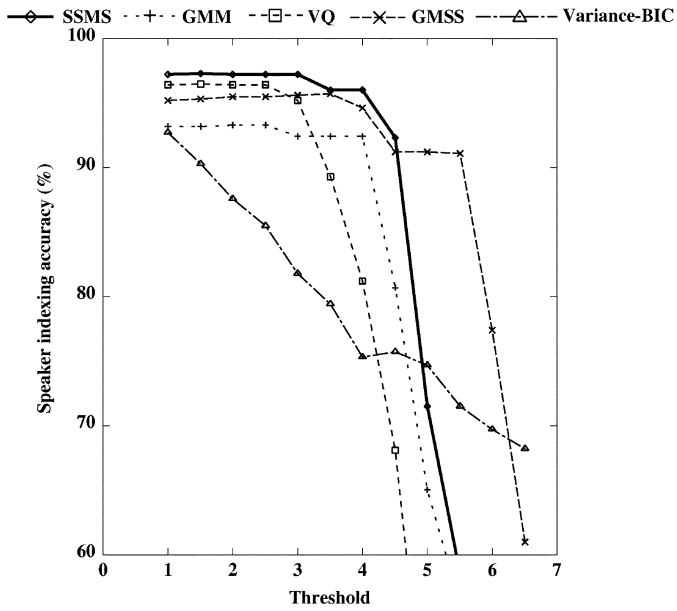

Fig. 4. Speaker indexing accuracy (SIA) versus threshold θ.

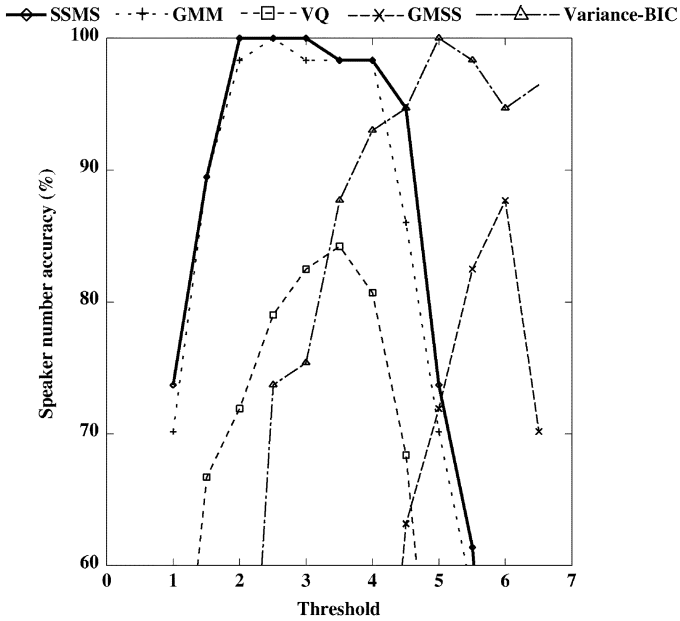

Fig. 5. Speaker number accuracy (SNA) versus threshold θ.

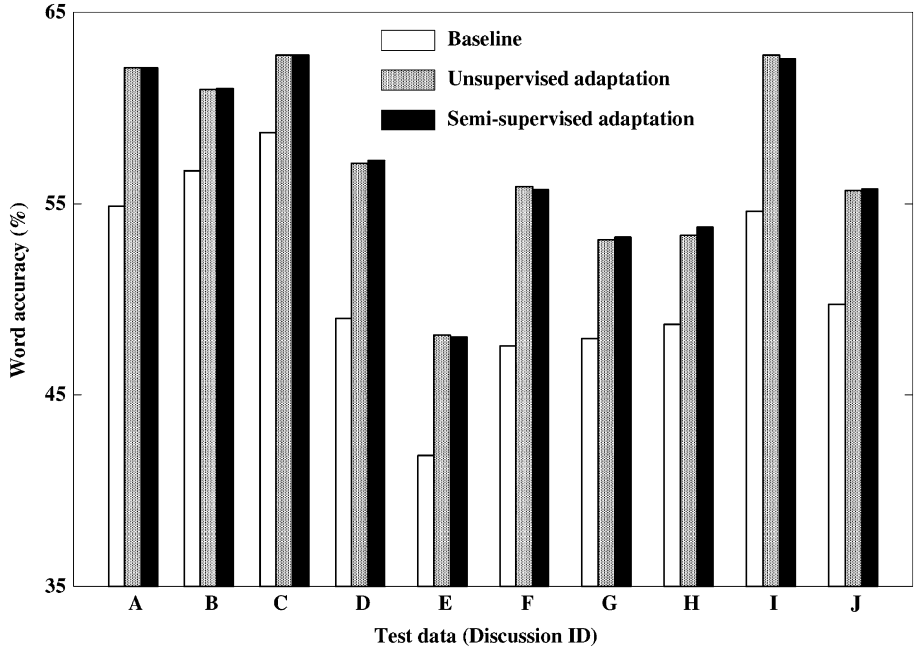

Fig. 6. Automatic speech recognition results for each discussion.

TABLE IV
AUTOMATIC SPEECH RECOGNITION RESULT

	Word accuracy (%)
Baseline	51.0
Unsupervised adaptation	57.2
Semi-supervised adaptation	57.2

V. ASR BASED ON SPEAKER ADAPTATION USING SPEAKER INDEXING RESULT

We then conducted automatic speech recognition (ASR) based on speaker adaptation using the indexing results. We used the speaker labels obtained by SSMS (32 mix.), which achieved the best indexing performance.

The baseline acoustic model is a phonetic tied-mixture triphone HMM (3000 states and 16 K Gaussians in total) trained with the Corpus of Spontaneous Japanese (CSJ) [42]. We use 43 phones, and all of them are modeled with left-to-right HMM of three states. We use 129 codebooks of 128 mixture components for tied-mixture modeling. The training data consisted of spontaneous oral presentations by 381 speakers that amounted to 60 h. 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 of the National Diet of Japan [43]. There are 36 053 vocabulary items. We used our Julius 3.3 decoder [44] for recognition with these models.

Unsupervised MLLR (Maximum Likelihood Linear Regression) speaker adaptation was performed for the baseline acoustic model using the results of speaker indexing. For each participant, utterances that were labeled as the speaker were used for adaptation. The initial ASR results with the baseline acoustic model were used for the phone transcriptions of utterances. We also perform semi-supervised adaptation of the baseline model using correct speaker labels and phone labels of the initial ASR results to investigate the upper limits in this scheme.

The average word accuracy obtained by the described methods is shown in Table IV. Here, "Baseline" denotes where the baseline acoustic model was used without adaptation. "Unsupervised adaptation" denotes the unsupervised adaptation using the speaker indexing and initial ASR results with the baseline model, and "Semi-supervised adaptation" denotes where correct speaker labels were used instead of the speaker indexing results. The word accuracy for each discussion is shown in Fig. 6.

Accuracy was 51.0% on average with the baseline model. Since the utterances in discussions are totally spontaneous and word perplexity is around 150, the ASR task is very difficult. Unsupervised adaptation improved it to 57.2%, and improvement was observed for all discussions. This demonstrates that unsupervised speaker adaptation based on the speaker indexing is very effective. The accuracy of semi-supervised adaptation was 57.2%, which is comparable to the totally unsupervised case. This means that the speaker indexing accuracy by the SSMS method is sufficient for adaptation of the acoustic model. For reference, we also conducted an experiment of the adaptation with the "Variance-BIC" speaker indexing method. The word accuracy was 56.9%, which shows some degradation from the result of the proposed method.

VI. CONCLUSION

We proposed a flexible framework in which an optimal speaker model (GMM or VQ) is selected based on the BIC. It automatically chooses the optimal model (GMM or VQ) according to the available training data and can be applied to a task in which the amount of training data is different for each speaker. The framework also makes it possible to use a discrete model (VQ) when the data is sparse, and to seamlessly switch to a continuous model (GMM) after a large amount of data is obtained.

As a typical real-world application, we implemented the proposed framework on unsupervised speaker indexing for a discussion audio where the duration of utterances is very short and its variation is large. For a discussion archive with a total duration of 10 hours, we demonstrated that the proposed method achieves higher indexing performance than conventional methods such as the Variance-BIC, VQ-based and GMM-based methods. SSMS achieved an accuracy of 97.0% both when the number of speakers was given beforehand and when the number was unknown. We also found that the method we propose is less sensitive to the threshold value for clustering. Thus, the method is shown to be not only accurate but also robust. Moreover, the proposed method achieved the sufficient performance without tuning the parameter α in the BIC.

We also applied and evaluated speaker indexing to automatic speech recognition of the discussions by adapting a speaker-independent acoustic model to each participant. It is shown that unsupervised speaker adaptation based on speaker indexing is very effective and the indexing accuracy with SSMS is sufficiently high for adaptation of the acoustic model.

As a future work, we will evaluate the proposed method on the NIST databases to demonstrate its generality.

REFERENCES

[1] J. He, L. Liu, and G. Palm, "A new codebook training algorithm for VQ-based speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 1997, pp. 1091–1094.

[2] R. D. Zilca and Y. Bistritz, "Text independent speaker identification using LSP codebook speaker models and linear discriminant functions," in *Proc. European Conf. Speech Commun. Tech. (EUROSPEECH)*, vol. 2, 1999, pp. 799–802.

[3] Q. Jin and A. Waibel, "A naive de-lambing method for speaker identification," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, vol. 2, 2000, pp. 466–469.

[4] M. Faundez-Zanuy, "A combination between VQ and covariance matrices for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, 2001, pp. 453–456.

[5] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[6] C. Tadj, P. Dumouchel, and P. Ouellet, "GMM based speaker identification using training-time-dependent number of mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 1998, pp. 761–764.

[7] R. Vergin and D. O'Shaughnessy, "On the use of some divergence measures in speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, 1999, pp. 309–312.

[8] L. Liu and J. He, "On the use of orthogonal GMM in speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 1999, pp. 845–848.

[9] W.-H. Tsai, C. Che, and W.-W. Chang, "Text-independent speaker identification using Gaussian mixture Bigram models," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, vol. 2, 2000, pp. 314–317.

[10] O. Thyes, R. Kuhn, P. Nguyen, and J.-C. Junqua, "Speaker identification and verification using eigenvoices," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, vol. 2, 2000, pp. 242–245.

[11] D. A. Reynolds, "Model compression for GMM based speaker recognition systems," in *Proc. Eur. Conf. Speech Commun., Tech. (EUROSPEECH)*, 2003, pp. 2005–2008.

[12] T. Matsui and S. Furui, "Comparison of text independent speaker recognition methods using VQ distortion and discrete/continuous HMMs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 1992, pp. 157–160.

[13] M. A. Przybocki and A. F. Martin, "The 1999 NIST speaker recognition evaluation, using summed two-channel telephone data for speaker detection and speaker tracking," in *Proc. Eur. Conf. Speech Commun. Tech. (EUROSPEECH)*, vol. 5, 1999, pp. 2215–2218.

[14] A. F. Martin and M. A. Przybocki, "Speaker recognition in a multi-speaker environment," in *Proc. Eur. Conf. Speech Commun. Tech. (EUROSPEECH)*, vol. 2, 2001, pp. 787–790.

[15] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech style," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, vol. 1, 2002, pp. 301–304.

[16] M. Nishida and T. Kawahara, "Unsupervised speaker indexing using speaker model selection based on Bayesian information criterion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, 2003, pp. 172–175.

[17] D. Roy and C. Malamud, "Speaker identification based text to audio alignment for an audio retrieval system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1997, pp. 1099–1102.

[18] A. E. Rosenberg, I. Magrin-Chagnolleau, S. Parthasarathy, and Q. Huang, "Speaker detection in broadcast speech databases," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 1998, pp. 1339–1342.

[19] K. Sonmez, L. Heck, and M. Weintraub, "Speaker tracking and detection with multiple speakers," in *Proc. Eur. Conf. Speech Commun. Tech. (EUROSPEECH)*, vol. 5, 1999, pp. 2219–2222.

[20] M. Sugiyama, J. Murakami, and H. Watanabe, "Speech segmentation and clustering based on speaker feature," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 1993, pp. 395–398.

[21] J. Murakami, M. Sugiyama, and H. Watanabe, "Unknown-multiple signal source clustering problem using Ergodic HMM and applied to speaker classification," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, vol. 4, 1996, pp. 2407–2410.

[22] H. Gish, M. H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 1991, pp. 873–876.

[23] Y. Moh, P. Nguyen, and J.-C. Junqua, "Towards domain independent speaker clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 2003, pp. 85–88.

[24] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification, and clustering of broadcast news audio," in *Proc. DARPA Speech Recognition Workshop*, 1997, pp. 97–99.

[25] S. Chen and P. Gopalakrishnan, "Speaker, environment, and channel change detection and clustering via the Bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 127–132.

[26] P. Delacourt, D. Kryze, and C. J. Wellekens, "Detection of speaker changes in an audio document," in *Proc. Eur. Conf. Speech Commun. Tech. (EUROSPEECH)*, vol. 3, 1999, pp. 1195–1198.

[27] M. H. Siu, G. Yu, and H. Gish, "An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 1992, pp. 189–192.

[28] L. Wilcox, F. Chen, D. Kimber, and V. Balasubramanian, "Segmentation of speech using speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, 1994, pp. 161–164.

[29] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1998, pp. 757–760.

[30] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification systems," in *Proc. Eur. Conf. Speech Commun. Tech. (EUROSPEECH)*, 1997, pp. 963–966.

[31] D. Charlet, "Speaker indexing for retrieval of voicemail messages," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, 2002, pp. 121–124.

[32] G. N. Ramaswamy, J. Navratil, U. V. Chaudhari, and R. D. Zilca, "The IBM system for the NIST-2002 cellular speaker verification evaluation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 2003, pp. 61–64.

[33] J. McLaughlin, D. Reynolds, E. Singer, and G. C. O'Leary, "Automatic speaker clustering from multi-speaker utterances," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 1999, pp. 817–820.

[34] 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," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 2002, pp. 577–580.

[35] S. Meignier, J.-F. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," in *Proc. Speaker Odyssey—The Speaker Recognition Workshop*, 2001, pp. 175–180.

[36] G. Kolano and P. Regel-Brietzmann, "Combination of vector quantization and Gaussian mixture models for speaker verification with sparse training data," in *Proc. Eur. Conf. Speech Commun. Tech. (EUROSPEECH)*, 1999, pp. 1203–1206.
[37] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantization design," in *IEEE Trans. Commun.*, vol. COM-28, 1980, pp. 84–95.
[38] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. Eur. Conf. Speech Commun. Tech. (EUROSPEECH)*, vol. 1, 1997, pp. 99–102.
[39] S. S. Chen and R. A. Gopinath, "Model selection in acoustic modeling," in *Proc. Eur. Conf. Speech Commun. Tech. (EUROSPEECH)*, vol. 3, 1999, pp. 1087–1090.
[40] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian information criterion," in *Proc. Eur. Conf. Speech Commun. Tech. (EUROSPEECH)*, vol. 2, 1999, pp. 679–682.
[41] 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," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 1998, pp. 3193–3196.
[42] H. Nanjo and T. Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2002, pp. 725–728.
[43] Y. Akita, M. Nishida, and T. Kawahara, "Automatic transcription of discussions using unsupervised speaker indexing," in *Proc. ISCA IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, 2003, pp. 79–82.
[44] A. Lee, T. Kawahara, and K. Shikano, "Julius—An open source real-time large vocabulary recognition engine," in *Proc. Eur. Conf. Speech Commun. Tech. (EUROSPEECH)*, 2001, pp. 1691–1694.

Masafumi Nishida received the B.E. degree in 1997, the M.E. degree in 1999, and the Ph.D. degree in 2002, all in electronics and informatics, from Ryukoku University, Shiga, Japan.

From 2002 to 2003, he was a Postdoctral Researcher at PRESTO, Japan Science and Technology Corporation (JST). Currently, he is a Research Associate at Graduate School of Science and Technology, Chiba University, Chiba, Japan. He has been working on speech recognition and speaker recognition.

Tatsuya Kawahara (M'91) received the B.E. degree in 1987, the M.E. degree in 1989, and the Ph.D. degree in 1995, all in information science, and all from Kyoto University, Kyoto, Japan.

In 1990, he became a Research Associate at Department of Information Science, Kyoto University. In 1995, he became an Associate Professor at Department of Information Science, Kyoto University. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ. In 1998, he became an Associate Professor at School of Informatics, Kyoto University. Currently, he is a Professor at Academic Center for Computing and Media Studies, Kyoto University. He is also an Invited Researcher at ATR Spoken Language Translation Research Laboratories. He has published more than 100 technical papers covering speech recognition, confidence measures, and spoken dialogue systems. He has been managing several speech-related projects in Japan including a free large vocabulary continuous speech recognition software project (http://julius.sourceforge.jp/).

Dr. Kawahara received the 1997 Awaya Memorial Award from the Acoustical Society of Japan and the 2000 Sakai Memorial Award from the Information Processing Society of Japan. Since 2003, he has been a member of the IEEE SPS Speech Technical Committee.